Behzad Mirzababaei, B.Sc M.Sc

Developing Full Tutorial Conversational Agents with Argumentation as a Learning Activity

DOCTORAL THESIS

to achieve the university degree of

Doktor der technischen Wissenschaften

submitted to

Graz University of Technology

Supervisor

Viktoria Pammer-Schindler, Assoc.-Prof. Dr.

Institute of Interactive Systems and Data Science

Graz, March 2024

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date

Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Dissertation identisch.

Datum

Unterschrift

Abstract

The significance of learning platforms has been underscored by both large-scale learning scenarios and the need for scalability and flexibility. Today's learning platforms have shifted from content-sharing to hosting diverse learning activities. Following the shift, this thesis presents educational conversational agents with which learners can experience a completely one-on-one tutorial session including receiving content (is "being taught") as well as quizzes, arguing or summarising content. An additional focus on argumentation as a learning activity distinguishes this research from existing literature, which mainly concentrates on providing argumentative feedback on written essays.

The overarching aim of this study is to enhance the learning experience with educational conversational agents which support various learning activities such as argumentation. The research methodology involves 1) conducting a within-subject experiment which compares a full-tutorial conversational agent, called DIGIBOT, with a web-based interactive system, called DIGIVIDget, in terms of user experience 2) focusing on argumentation as learning activity and developing machine learning classifiers for providing feedback on structural flaws in arguments (based on Toulmin's model of argument), 3) conducting a between-subject experiment in which we compare the impact of such argumentative feedback on writing argumentative essays, 4) defining a systematic workflow to identify argumentative components and provide feedback in various topics using pretrained transformer-based language models.

The findings show the potential of DIGIBOT especially in inclusive education due to the interactive nature of technology and the motivational effect of receiving immediate feedback. Regarding argumentation, the results show the feasibility of constructing accurate classifiers to identify Toulmin's components within a conversation. Finally, transformer-based language models demonstrate their potential to support learning engineers for creating an argumentation module in educational conversational agents, yielding classifiers with F1-macro scores ranging from 0.66% to 0.86%.

This research contributes significantly to the field by not only addressing the identified research gaps but also by offering insights into the future of educational conversational agents. By providing a glimpse into the design principles of full-tutorial conversational agents and demonstrating the effective integration of argumentation as a learning activity, the study lays the foundation for a new paradigm in educational technology. The findings contribute to the development of effective conversational agents that enhance learning experiences using various learning activities and bridge the content/learning dichotomy in computational environments.

Keywords: Educational conversational agents, argumentation, Toulmin's model, Transformer-based language models, LLMs

Kurzfassung

Die Bedeutung von Lernplattformen wird sowohl durch umfängliche große Lernszenarien als auch durch den Bedarf an Skalierbarkeit und Flexibilität unterstrichen. Moderne Lernplattformen haben sich vom Teilen und gemeinsamen Nutzen von Inhalten auf das Anbieten vielfältiger Lernaktivitäten verlagert. In dieser Arbeit werden pädagogische Konversationsagenten vorgestellt, die Lernenden ein komplettes Einzeltutorium anbieten, welches sowohl den Empfang von Inhalten ("Unterricht") als auch Quizfragen, Argumentation, und die Zusammenfassungen von Inhalten umfasst. Ein zusätzlicher Fokus auf Argumentation als Lernaktivität unterscheidet diese Forschung von bestehender Literatur, die sich hauptsächlich auf die Bereitstellung von argumentativem Feedback zu schriftlichen Aufsätzen konzentriert.

Das übergreifende Ziel dieser Studie ist es, die Lernerfahrung mit pädagogischen Konversationsagenten zu verbessern, welche verschiedene Lernaktivitäten wie z.B. Argumentation unterstützen. Die Forschungsmethodik umfasst 1) die Durchführung eines Within-Subjects-Experiments, in dem ein volltutorialer Conversational Agent, genannt DIGIBOT, mit einem webbasierten interaktiven System, genannt DIGIVIDget, in Bezug auf die Nutzererfahrung verglichen wird, 2) die Konzentration auf Argumentation als Lernaktivität und die Entwicklung von Klassifikatoren für maschinelles Lernen, um Feedback zu strukturellen Fehlern in Argumenten zu geben (basierend auf Toulmins Modell des Arguments), 3) die Durchführung eines Experiments, in dem die Auswirkungen eines solchen argumentativen Feedbacks auf das Schreiben von argumentativen Aufsätzen verglichen wird, und 4) die Definition eines systematischen Arbeitsablaufs zur Identifizierung argumentativer Komponenten und zur Bereitstellung von Feedback in verschiedenen Themenbereichen unter Verwendung von vortrainierten transformatorbasierten Sprachmodellen.

Die Ergebnisse zeigen das Potenzial von DIGIBOT insbesondere in der inklusiven Bildung, aufgrund der Interaktivität der Technologie und der motivierenden Wirkung des unmittelbaren Feedbacks. In Bezug auf die Argumentation zeigen die Ergebnisse, dass es möglich ist, genaue Klassifikatoren zu konstruieren, um Toulmins Komponenten innerhalb einer Konversation zu identifizieren. Schließlich wird das Potenzial transformatorbasierter Sprachmodelle offengelegt, Domänenexperten bei der Erstellung eines Argumentationsmoduls in konventionellen Bildungsagenten zu unterstützen, indem Klassifikatoren mit F1-Makro-Scores zwischen $0,66\%$ und $0,86\%$ geliefert werden.

Diese Forschungsarbeit leistet einen wichtigen Beitrag in diesem Forschungsfeld, indem sie nicht identifizierte Forschungslücken schließt, sondern auch Ausblicke in die Zukunft von konversationellen Agents bietet. Die durchgeführten Studien schaffen Verständnis überdie Designprinzipien von volltutoriellen Conversational Agents, demonstrierendie effektive Integration von Argumentation als Lernaktivität, und legen soden Grundstein für ein neues Paradigma in der Bildungstechnologie. Die Ergebnisse tragen zudem zur

Entwicklung effektiver Conversational Agents bei, welche Lernerfahrungen durch verschiedene Lernaktivitäten verbessern und unterstützen die Überbrückung der Dichotomie von Inhalten und Lernen in computergestützten Umgebungen.

Acknowledgements

My thesis would not have been possible without the endless help and support from my supervisor, colleagues, friends and my lovely wife. First and foremost I would like to thank my supervisor, Assoc. Prof. Viktoria Pammer-Schindler. I appreciate all her contributions of time, inspiring ideas and strong support to my PhD journey. She is the best supervisor that a PhD student could have. Whenever I meet difficulties, she is always there for help.

A huge thank you goes out to the best team ever, Digital Transformation Design. I was lucky to be part of this team. I learned a lot from every one of them, Alfred, Angela, Christof, Gert, Hermann, Irmtraud, Katharina, Leonie, Marie, Mia, Michael, Sebastiam, Simone and Viktoria. I would like also to extend my appreciation to the DDAI team and especially to Dr. Andreas Trügler who was like a mentor to me.

Special thanks to Angela not only for her motivational support but also for her delicious Tiramisu which made our team events memorable. Also special thanks to Alfred who is now in my friend zone for his technical support and all our discussion at lunch breaks.

To my friends, Ata, Bahar, Nazanin and Sara (Stefan's wife) who stood by me with encouragement. Studying PhD and living abroad has been quite an adventure and I do not think this adventure would have been the same without them.

Thanks to my parents for always being there notwithstanding the distance and the odd pandemic. Special thanks to my uncle, Reza, for inspiring me to focus on what matters the most in life despite my professional and academic career.

Last but not least, I extend my heartfelt gratitude to the most important person in my life, my lovely wife, Sara. Her unwavering support and encouragement sustained me throughout this academic journey. Her patience, understanding, and belief in my capabilities were invaluable motivators. She is the most and the biggest internal source of power, energy and motivation in my life.

This work was supported by the "DDAI" COMET Module within the COMET-Competence Centers for Excellent Technologies Program, funded by the Austrian Federal Ministry (BMK and BMDW), the Austrian Research Promotion Agency (FFG), the province of Styria (SFG) and partners from industry and academia. The COMET Program is managed by FFG.

Contents

List of Figures

List of Tables

List of Acronyms and Symbols

CA	Conversational Agent
AI	Artificial Intelligence
NLP	Natural Language Processing
GDPR	General Data Protection Regulation
PET	Privacy-Enhancing Technology
ML	Machine Learning
LMS	Learning Management Systems

1. Introduction

In recent years, conversational agents have emerged as powerful tools that significantly impact our daily lives. These agents, including chatbots and virtual assistants, have revolutionised the way we interact with technology, access information, and accomplish tasks. From simple customer support interactions to complex personal assistant functionalities, conversational agents have become increasingly prevalent in various domains, ranging from e-commerce and healthcare to entertainment and education.

Conversational agents have also emerged as a promising technology in the field of education, revolutionising the way students learn and interact with educational resources. These intelligent software applications utilise Natural Language Processing (NLP) and Artificial Intelligence (AI) to engage in human-like conversations with learners. By simulating dialogue-based interactions, conversational agents have the potential to enhance educational experiences, personalise instruction, and provide on-demand support.

The impact of conversational agents in education is significant and far-reaching. As technology continues to shape the educational landscape, conversational agents offer unique opportunities to improve learning outcomes and address individual student needs. One of the primary advantages of using conversational agents in education is their ability to provide personalised and adaptive learning experiences. This personalised approach enhances student engagement, motivation, and retention, as the agent can adapt its delivery of content and provide targeted feedback based on the student's responses.
However, the effective integration of conversational agents in educational settings comes with its own set of challenges. For instance, designing conversational agents that can understand and respond accurately to diverse student queries and contexts remains a significant task. Providing adaptive feedback which leads to better learning outcomes or performance is also important and challenging. Converting the learning content into a dialogic form which can be transferred by a conversational agent is also challenging.

The overarching goal of using conversational agents in educational scenarios is to enhance the learning experience by implementing various teaching techniques such as transferring learning materials step by step, having quizzes, giving personalised feedback, summarising the learning materials and arguing to learn.

1.1. The Importance of Argumentation

Argumentation is an essential skill that allows individuals to effectively communicate their ideas and beliefs to others. Whether it is in a professional setting or everyday conversations, being able to construct a strong argument can help to persuade others, resolve conflicts, and promote critical thinking.

> "*It is in argument that we are likely to find the most significant way in which higher order thinking and reasoning figure in the lives of most people. Thinking as argument is implicated in all of the beliefs people hold, the judgments they make, and the conclusions they come to; it arises every time a significant decision must be made. Hence, argumentative thinking lies at the heart of what we should be concerned about in examining how, and how well, people think.*" [Kuh92]

When individuals engage in argumentation, they are required to consider multiple perspectives, analyse evidence, and weigh different viewpoints. This process encourages individuals to evaluate their beliefs critically and consider alternative perspectives, which can lead to a deeper understanding of the issue. Moreover, by learning how to construct a strong argument, individuals can also better identify weaknesses in their own beliefs and seek out further information to strengthen their position.

Another important aspect of argumentation is its role in resolving conflicts. In both personal and professional settings, disagreements are inevitable. However, by engaging in argumentation, individuals can resolve conflicts in a constructive and productive manner. By presenting evidence and logical reasoning, individuals can reach a mutual understanding and find a solution that satisfies both parties.

In addition to its benefits in critical thinking and conflict resolution, argumentation is also an essential skill in many professional fields. Whether it is in law, business, or politics, being able to construct a persuasive argument can be a valuable asset. In these fields, individuals must be able to convince others of their ideas and negotiate effectively, which requires strong argumentation skills. Moreover, being able to present a clear and concise argument can also help individuals stand out in their careers, as it demonstrates their ability to think critically and communicate effectively.

By learning how to construct a strong argument, individuals can develop a deeper understanding of the world around them and engage in constructive discussions and debates. Moreover, argumentation is an essential skill in many professional fields and can help individuals stand out in their careers. As such, it is essential that individuals learn and practice the art of argumentation in order to succeed in both their personal and professional lives.

1.2. Argumentation in Education

One of the techniques for enhancing the learning experience is leveraging argumentation. Argumentation is an essential aspect of education that helps students develop critical thinking, effective communication, and problem-solving skills [JK10a]. In an educational setting, argumentation refers to the process of constructing and presenting a persuasive argument based on evidence and logical reasoning. Argumentation can take various forms, including debates, discussions, and presentations, and it plays a crucial role in shaping students' understanding of complex issues and promoting active engagement in the learning process.

One of the primary benefits of argumentation in education is that it promotes critical thinking. When students are required to construct an argument, they must gather evidence, evaluate different perspectives, and analyse the strengths and weaknesses of different viewpoints. This process encourages students to think critically about complex issues, engage with ideas that challenge their existing beliefs, and develop a deeper understanding of the material they are studying. Through argumentation, students learn how to assess the quality of evidence, identify fallacies in arguments, and make sound judgements based on logic and reasoning.

Moreover, argumentation in education is also an essential tool for effective communication. When students learn how to construct a persuasive argument, they are also learning how to communicate their ideas effectively. This involves being able to present their ideas in a clear and concise manner, anticipate counterarguments, and respond to criticism. By developing these communication skills, students can engage in meaningful discussions with their peers and teachers, work collaboratively on group projects, and succeed in future professional environments where effective communication is essential.

Another important aspect of argumentation in education is its role in promoting constructive problem-solving. When students engage in argumentation, they are often required to work collaboratively to develop a solution to a problem. This process encourages students to think creatively and consider multiple perspectives, while also identifying potential weaknesses in their proposed solutions. By learning how to construct a persuasive argument in this context, students can work towards finding a solution that is both effective and sustainable.

In educational settings (e.g. in classrooms), argumentation can be utilised by providing guidelines for teachers to encourage learners to argue and justify more or to provide situations in which the learners are asked to assert and analyse their opinions (e.g. [Geo+20]).

1.3. Argumentation in Educational Conversational Agents

Advances in NLP have revolutionized the field of argument mining, enabling the automated extraction and analysis of arguments from textual data. Argumentation can also be used in conversational agents or, in general, in computer-mediated learning environments (e.g. [Afr+21; Wam+21; WJL22]). The research on using argumentation in computer-mediated learning environments mainly focused on providing feedback on argumentative essays written by students and not on teaching argumentation per se. For instance, in [Afr+21], the authors developed a web-based learning environment in which learners can write essays and then receive feedback on the argumentative components based on a specific argumentative scheme.

In educational settings, the main focus was on argumentative essays written by students. Argumentative essays follow a specific template and contain structural features which are effective in detecting discourse elements. Such discourse elements such as "*in my opinion*", "*in conclusion*" were used by researchers to identify argumentative sentences in the essays (e.g. [PM09]). Due to focusing on argumentative essays, several corpora or datasets have been developed for argument mining in educational settings.

The developed corpora contain annotations which measure different aspects of an argument. For example, Persing and Ng [PN20] annotated 1000 essays based on the argument strength. In [PTT22], the English essays written by non-English speakers were annotated based on whether each sentence is argumentative or not and also the relation between argumentative sentences.

Learning to argue in educational scenarios is more than how to write an argumentative essay. Students need to be supported in how to provide, at least, a structurally complete argument which is useful in critical thinking, effective communication and constructive problem-solving.

In this thesis, we addressed the gap in the literature first by analysing how conversational agents can support various learning strategies, and second by investigating the impact of a conversational agent with argument-mining capability in learning environments.

1.4. Objectives and Main Contributions

The primary aim of the thesis is to augment the educational experience using conversational agents and argumentation. This objective is pursued along two distinct avenues:

1. Firstly, emphasis is placed on comprehending the characteristics of a fully tutorial conversational agent as an educational technology and analysing the user experience in using such an agent in comparison to other educational technologies.

2. Secondly, attention is directed towards using argumentation as a teaching strategy and providing adaptive feedback on learners' argumentation in fully tutorial conversational agents.

The contribution of the thesis can be summarised in five points as shown in Figure 1.1. The first two contributions are in line with the first objective and the rest contribute to the second objective.

1. The demonstration of a fully tutorial conversational agent that leverages various learning activities, including knowledge transfer, short quizzes, feedback, and summarising content, through systematic instructional dialogue design based on Bloom's revised taxonomy of learning objectives [AK01] (Chapter 4).

2. An analysis of user experience using two different methods of learning content delivery: freely navigable web-based content and a dialogic interaction. The study conducted with 31 master students of inclusive education (Chapter 5) provides insight into the effectiveness of each method.

3. The development of accurate classifiers to identify individual components of Toulmin's argumentative model [Tou03] in short argumentative statements. This research demonstrates the utility of such classifiers in the dialogue structures of conversational agents designed based on Bloom's revised taxonomy [AK01] for learning (Chapter 6).

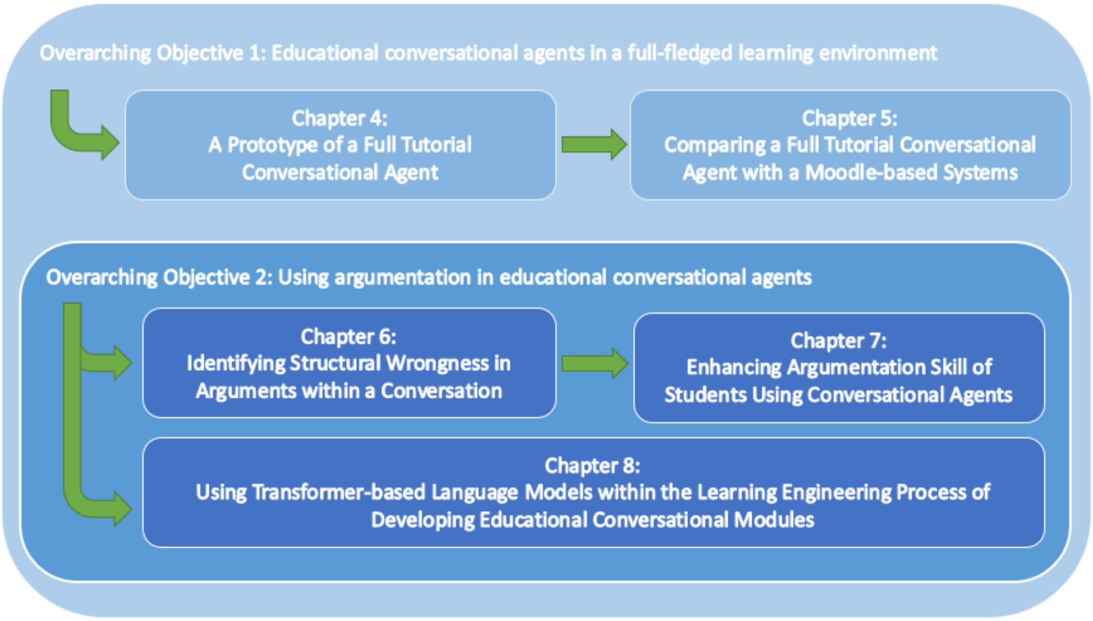


Figure 1.1.: The overarching objectives and the chapters related to each objective

4. The exploration of the impact of conversational agents on students' argumentation skills by identifying missing argumentative structural gaps. This study complements existing research on computational environments for teaching argumentation in longer texts (Chapter 7).

5. The provision of a step-by-step guideline to develop adaptive and argumentative conversational modules using pre-trained transformer large language models for various topics, provided that learning content and teaching strategies for engaging learners in argumentation are available (Chapter 8).

1.5. Outline

The present study is a comprehensive investigation of the usefulness of conversational agents in education and argumentation. The overall structure of the study takes the form of nine chapters, each of which contributes to the overall understanding of the whole study. The chapters are as follows:

- Chapter 2 provides an overview of the wider fields of educational conversational agents and argumentation by presenting the latest related work. The chapter offers a detailed literature review and analysis of the current state of research in these areas, highlighting the gaps in the literature and setting the foundation for the subsequent chapters.

- Chapter 3 offers the contributions and publications derived from the study and how they are connected.

- Chapter 4 illustrates a conversational agent that can cover all the learning steps such as transferring the content and having short quizzes based on Bloom's revised taxonomy [AK01]. This chapter largely corresponds to our paper which included full tutorial conversational agents [MPS22a].

- Chapter 5 explores and analyses how the user experience differs between learning via web-based interactive content and learning within a dialogue led by a computational tutorial agent. This chapter largely corresponds to our paper in which a fully tutorial conversation agent and a learned management system were compared [Mir+23].

- Chapter 6 covers the process of identifying the core components of Toulmin's model of argument in a conversation with an agent. The chapter addresses all the steps from collecting data to training machine learning models. The aim of this chapter is to provide a detailed description of the methodology used to develop the conversational agents and to identify the key components of Toulmin's model of argument [Tou03] in the conversation. This chapter largely corresponds to our paper in which we developed machine learning models to identify Toulmin's core components [MPS21a].

- Chapter 7 verifies the benefits of the proposed intelligent conversational agents, which are able to identify Toulmin's core components on students' argumentative skills. The chapter evaluates the students' argumentative skills in two completely different domains and examines the impact of the conversational agents on their learning outcomes regarding how to write a complete argument based on Toulmin's model of argument. This chapter largely corresponds to our paper in which we analysed the impact of providing argumentative feedback on writing argumentative answers or essays [MPS21b].

- Chapter 8 focuses on the generalisation of identifying Toulmin's core components in various domains. It provides a step-by-step guideline for identifying the core components in various educational domains, which could be used to enhance the effectiveness of conversational agents in different learning contexts. This chapter largely corresponds to our final paper in which we proposed a workflow for creating argumentative conversational modules using transformer-based language models. The paper was submitted to IEEE Transactions on Learning Technologies and it is currently accepted by minor revision.

- Finally, Chapter 9 discusses and concludes the work of the thesis, summarising the research findings, and outlining the potential avenues for future research in the field of conversational agents in education and argumentation.

2. Background

In this chapter, we cover the background and related works in conversational agents and argumentation. Firstly, we look at the classification of conversational agents (Section 2.1.1) and specifically educational conversational agents (Section 2.1.2). Secondly, we look at argumentation and argumentative schemes (Section 2.2.1). After that, argument mining is elaborated as one of the Natural Language Processing (NLP) tasks, Section 2.2.2. With a good understanding of the argumentative models and argument mining techniques, we move to argumentation in education (Section 2.2.3) and argumentative conversational agents (Section 2.2.4). Finally, we elaborate on the challenges in argumentative conversational agents, Section 2.4.

2.1. Conversational Agents

Conversational agents, commonly known as chatbots, have witnessed a remarkable evolution over the past several decades. From their beginnings as rule-based systems to the emergence of advanced artificial intelligence models, conversational agents have become increasingly sophisticated in understanding and engaging in human-like conversations.

The history of conversational agents can be traced back to the mid-1960s when the pioneering chatbot ELIZA, developed by Joseph Weizenbaum at MIT [Wei66], provided a glimpse into the potential of human-computer interaction. ELIZA employed pattern matching and scripted responses to simulate a Rogerian psychotherapist, demonstrating the ability to engage in rudimentary text-based conversations.
As technology advanced, more complex chatbots emerged. PARRY, developed by Kenneth Colby in 1972 [Col+72], simulated the behaviour of a person with paranoid schizophrenia. ALICE, created by Richard Wallace in 1995 [Wal95b], utilised pattern matching techniques and AIML to generate responses, winning multiple Loebner Prize Turing Tests and highlighting the progress in mimicking human-like conversation.

Due to advances in NLP, the new wave of conversational agents such as Siri, Google Assistant and ChatGPT demonstrated remarkable language understanding and the ability to generate coherent and contextually relevant responses, pushing the boundaries of the conversational agent's capabilities.

2.1.1. Conversational Agent's Classifications

From a technical standpoint, conversational agents employ two primary approaches to provide responses to users: generative and retrieval approaches. The generative approach is associated with chatbots that lack a predefined set of responses and they instead generate responses from scratch. This approach offers the advantage of the capability to create novel responses. The methods of generating agent's responses can be varied. For instance, Ritter et al. [RCD11] utilised statistical machine translation to generate responses by translating users' utterances into the agent's responses. Another approach, as demonstrated in Le et al. [LNN18], involved using an encoder-decoder architecture to generate agent responses based on previous users' interactions.

The second approach is retrieval-based agents, where, for each turn in a conversation, the agent selects the most relevant response from a predefined list of possible answers. Techniques such as employing different types of similarities (e.g. [LNN18; Rak+19]) are used to determine the best response. Retrieval-based agents rely on pre-existing conditional dialogue structures, allowing them to choose from a range of predefined responses. While the selection of responses from a predefined list has certain limitations, from a technological perspective, this approach is relatively simpler compared to the generative approach.

Besides response generation methods, conversational agents can be classified based on various categories. For instance, in [AM20], conversational agents were categorised based on the knowledge domain into open and closed domains and also based on their goals into informative, chat-based and task-based. Following the taxonomy of conversational agents in the literature, we focused on educational conversational agents which are retrieval, task-based and have a closed domain.

2.1.2. Educational Conversational Agents

The concept of computational systems resembling human tutors has a long-standing history, leading to extensive research on intelligent tutoring systems in various forms (e.g. [AS14; HSA19; MCd19]). With advancements in machine learning and NLP, conversational interaction with intelligent tutoring systems has gained increased attention [IL17]. Conversational agents hold high expectations as learning support tools, as they are anticipated to address students' sociocultural needs, and to promote engagement and motivation (e.g. [VR14]) by facilitating interaction in natural language, thereby creating a sense of conversing with a social entity such as a peer or a tutor in educational contexts.

In educational conversations, retrieval-based conversational agents with a predefined set of responses have commonly been employed (e.g. [Gra+00; Wol+22; Cai+19]). Although the agents' responses are limited and predefined, various methods can be utilised to select the best one based on learners' responses. For example, in [Gra+00], a list of possible responses for the agent was created and the agent used latent semantic analysis (LSA) to assess the similarity of users' responses to the list of expected answers. LSA is an NLP technique that analyzes the relationship between words and document topics. Graesser et al. ([Gra+00]) employed language modules with a comprehensive lexicon of approximately 10,000 words, where each entry included alternative syntactic classes and word usage frequency in the English language. Additionally, Graesser et al. ([Gra+00]) classified learners' responses into five categories: WH-questions, YES/NO questions, Assertion, Directive, and Short responses. Similarly, Wolfbauer et al. ([WPSR20; Wol+22]) utilised a reflective conversational agent for apprentices, where the apprentices' responses were matched to different predefined concepts, and the agent's response was selected from a pool of concept-related responses. In [WPSR20], a dictionary-based approach and regular expressions were used to classify apprentices' utterances. The chatbot ALICE [Wal95b] utilised AIML (Artificial Intelligence Markup Language) to match responses to different categories and determine the most suitable response for user inputs.

Topic-wise, educational conversational agents can focus on various dialogue subjects or topics, such as mathematics (e.g. [Gro+19; MS04; Sab+13; ATR14; ZJ17]), physics (e.g. [Van+02; PMB13]), medicine (e.g. [FF00; SH04; Mar+09]), or computer science (e.g. [Wal95a; WM11; Koe+13; Wan+15]). In these examples, conversational agents are utilised to support learning in specific domains.

Besides learning domains, the target groups of educational conversational agents are various, such as K-12 students (kindergarten to 12th grade) (e.g. [Wal95a; Dzi+10]), university students (e.g. [Van+02; SH04; Wam+21; WM11]), or apprentices (e.g. [WPSR20; Wol+22]).

In this thesis, we leveraged various techniques such as machine learning classifiers, dictionary-based approaches and predefined large language models to analyse the users' responses. As for the learning domains, we dealt with different learning domains such as the definitions of intelligence (Chapter 6), the basics of General Data Protection Regulation or GDPR (Chapter 4), privacy-enhancing technologies (Chapter 4) and how to write a concrete search query which is one of the main digital competences [RSY22] (Chapter 5). Regarding the target groups, we focused on students (mainly university students due to the topic of the conversations).

2.2. Argumentation

In general, an argument consists of a claim (conclusion) and relevant reasons (premises) [WRM08]. People have utilized argumentation - the skill of presenting and defending viewpoints through reasoned discourse - as an essential component of communication. From ancient philosophical debates to modern courtroom proceedings, argumentation has played a crucial role in shaping our understanding of truth, advancing knowledge, and resolving conflicts. The goal of argumentation can be for persuasion, justification, reaching agreement or deliberation [HW07]. From an academic point of view, argumentation is a multidisciplinary research field, which studies debate and reasoning processes, and spans across and ties together diverse areas such as logic and philosophy, language, rhetoric and law, psychology, and computer science. Argumentation has become increasingly central as a core study within AI [BCD07], due to its ability to conjugate representational needs with user-related cognitive models and computational models for automated reasoning.

2.2.1. Argumentation Models

While the term "argument mining" was initially introduced by Palau and Moens in 2009 [PM09], the exploration of argumentation and its impacts has a historical lineage dating back to the 4th century BC. Over the years, numerous methodologies for scrutinizing argumentation have been explored, resulting in the development of various theories, schemes, and diagrams aimed at elucidating the definition and structure of an argument.

The idea behind the development of new assessment and visualisation methods for representing arguments stems from the necessity for straightforward and impactful approaches to deconstruct, analyse, and ultimately comprehend complex arguments. Given that argumentation can attain a considerable level of complexity, there is a demand for more straightforward forms of representation to effectively convey it. The concept of illustrating and breaking down arguments is foundational in argument mining, involving the examination, evaluation, and eventual expression of arguments in a binary format that can be interpreted by various algorithms.

Argumentation, rooted in rhetoric and philosophy, has led to the development of argumentation diagrams, which serve as helpful tools for constructing and understanding arguments in formal or documents which have a clear structure. Originally used as practical aids for teaching logic, diagram techniques have evolved into a more sophisticated method, serving as a conceptual framework for argument modelling. These logical diagrams have significantly advanced the field of informal logic, offering a practical means to analyze and assess everyday arguments beyond the confines of formal logic. Reed et al. ([RG07]) provide a thorough exploration of argumentation diagrams and their intersections with other disciplines like formal and informal logic, law, and artificial intelligence. It is important to note that these diagrams were not explicitly designed for the nuances of modern argument construction in social media. The tasks of detecting, classifying, and evaluating argumentative content in noisy text call for more adaptable frameworks.

In 1857, Whately ([Wha97]) pioneered the use of diagrams in argumentation, aiming to simplify the teaching of argumentation during his time. Whately's diagram theory revolves around identifying the concluding assertion, then tracing the reasoning backwards to ground the original assertion, ultimately forming a tree with assertions and proofs. In Figure 2.1, the conclusion serves as the root of the tree, while the assertions are positioned underneath. Using the classical example of Socrates' syllogism ("Socrates is a man, all men are mortal, therefore Socrates is mortal"), the conclusion ("Socrates is mortal") would be the root, and the two premises (p1 - "Socrates is a man", p2 - "all men are mortal") would be the leaves. The complexity and depth of the tree are proportional to the reasoning process, potentially leading to an intricate procedure that necessitates well-structured arguments and experienced annotators.

Beardsley ([Bea50]) has contributed significantly to the field of Argumentation with a representation that features distinctive statements of the argument as interconnected nodes. These nodes, capable of diverse connections, give rise to three fundamental classes: (1) convergent, (2) divergent, and (3) serial arguments. In Figure 2.2, a logical flow is depicted, showcasing serial linking between premises that lead to the convergence for the final argument. In the case of a convergent argument, various premises ultimately contribute to establishing a reliable and robust argument, fortified with reasonable and enough support.

Beardsley's theory has a notable limitation, namely the absence of support between the statements in the nodes. Consequently, the statements comprising the argument are considered flawless, without the subjects for support, debate, or evaluation. This

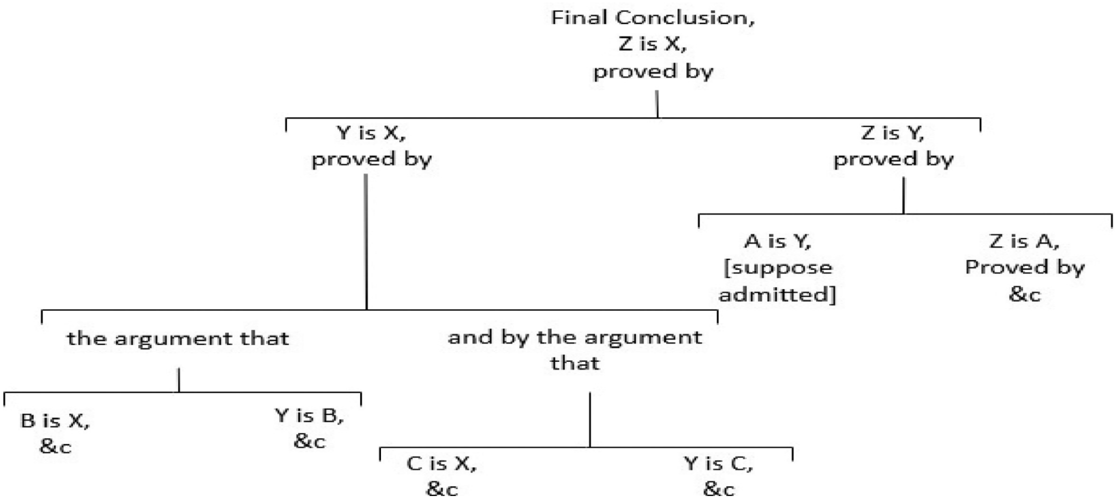


Figure 2.1.: Whately's diagram [Wha97] emphasises on analysing arguments based on backward reasoning. The final conclusion is represented as the root of the tree and its assertions are represented as leaves and the depth of the tree is proportionate to the complexity of the argument. The example is adapted from [Lyt+19].

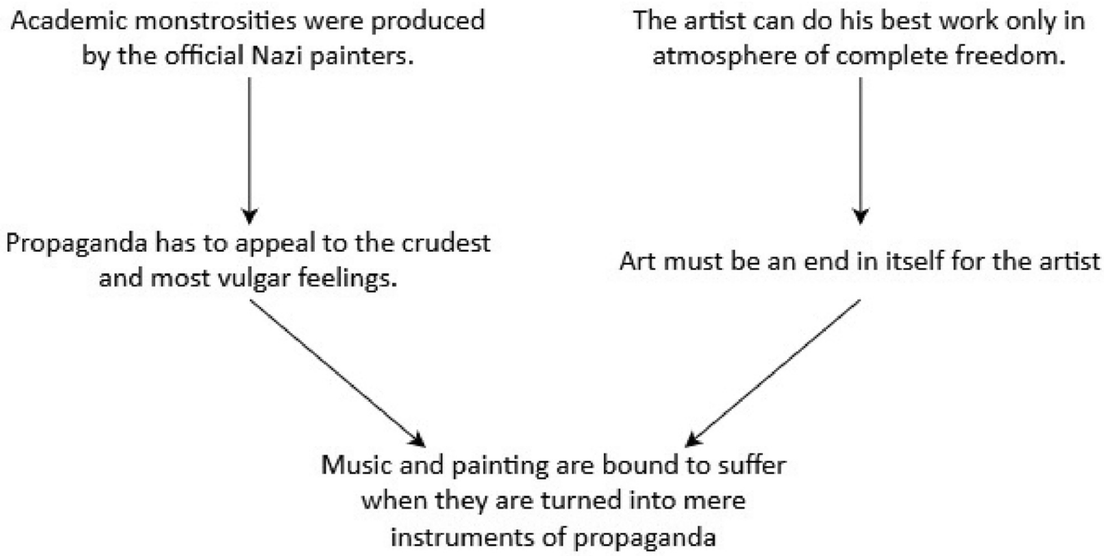


Figure 2.2.: Beardsley's convergent argument scheme [Bea50] provided with an example adapted from [Lyt+19]. The serial premises eventually lead (converge) to the final conclusion. It should be noted that the links between the premises are not evaluated.

constraint makes the theory not suitable for application in scenarios involving ambiguous, implicit, or imperfect arguments.

g tasks. Freeman's theory could be characterised as an upgrade of Beardley's theory [Bea50], as it uses the scheme of inductive/deductive reasoning and enhances it with the concept of modality, which indicates the strength of induced conclusion by the premises.

Freeman ([Fre11]) has developed an argumentation theory that embraces a data-driven approach for argument mining tasks. This theory can be seen as an advancement of Beardsley's theory ([Bea50]), incorporating the scheme of inductive/deductive reasoning and augmenting it with the modality concept. Modality, in this context, signifies the strength of the induced conclusion drawn from the premises.

In the 1980s, Mann ([Man84]) introduced a prominent theory on text organization, focusing on arranging the text into distinct sections. Each section consists of a crucial central part, known as the nucleus, vital for comprehending the text, and several satellites containing supplementary information about the nucleus. The relationship between the nucleus and the satellites involves various relations (such as circumstance, elaboration, and evidence). Based on the topic and the task, these relations can be modified and manipulated. The nucleus-satellite concept serves as a foundational element repeatedly employed to structure various sections of a text, ultimately forming a tree-structured hierarchy, as illustrated in Figure 2.3. In this hierarchy, the initial nucleus is substantiated by a series of premises. Each premise is sequentially articulated through the model of nucleus-satellite, manifesting a specific relation such as preparation, condition, or means. This argumentative schema has proven to enhance the effectiveness of sentiment analysis when complemented with argumentation [CT17].

Stephen Toulmin, a British philosopher, suggested one very influential scheme. It has been utilised in many studies, examining the role that different utterances might have in the persuasive perspective of the argument [Tou03]. This model which comes from a philosophical view, is essentially a structure for analysing arguments. Based on Toulmin's conceptual schema, Figure 2.4, an argument consists of six different components: a claim, a warrant, evidence, qualifiers, rebuttal, and backing. A claim is an assertion or a conclusion whose validity must be proven. Evidence is some information or a piece of knowledge that is used to make the claim. The warrant is used to establish the logical connection between the claim and the evidence. A qualifier is a word or phrase that shows the certainty of the claim. The rebuttal is an exception or an aspect for which the claim is not valid, in other words, another valid view to the claim. And finally, grounding materials or extra facts that support the warrant are called the backing component. Based on the model, only three of the components, the claim, warrant and evidence, which are also called the core components, are needed to have a complete argument.

This model or its modified versions have been used in numerous research (e.g. [Sim08; HG17; Wan+20; Geo+20; WF06]). As an example, in [Sim08], the author enriched teachers in the teaching and evaluation of argumentation in science contexts by using a program by which the teachers learn how to identify Toulmin's components in discussions and also teach students how to argue. Toulmin's model also has been used in computational argumentation. For instance, Habernal and Gurevych [HG17] used machine learning approaches to identify Toulmin's components of arguments in essays.

In this thesis, we dealt with the structure of an argument, therefore, we used Toulmin's

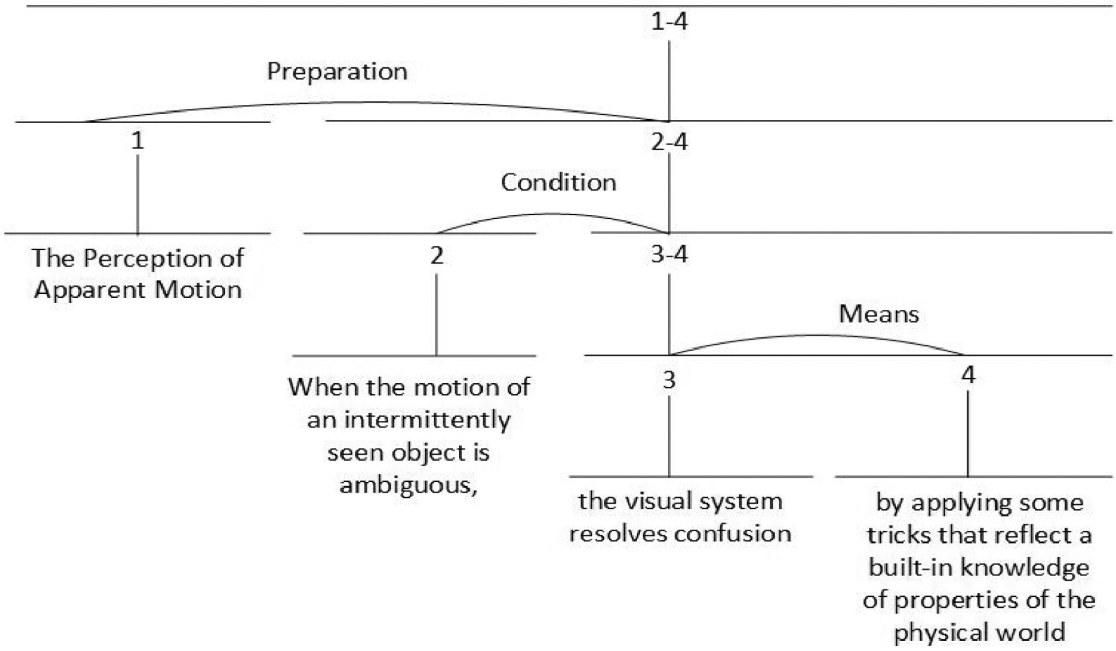


Figure 2.3.: Rhetorical Structure Theory scheme provided with an example [Man84] (adapted from [Lyt+19]). The example that is used is the theorem of perception of apparent motion (initial nucleus), which is justified by a set of premises, and each premise is analysed consequently based on the nucleus-satellite model. The example is adapted from [Lyt+19].

model of argument [Tou03] which breaks down an argument into different components. This model has been already used in similar studies such as [DV09; DNO00; Auf+08].

2.2.2. Argument Mining

Computational argumentation mining, also known as argument mining, is a field of study within NLP. Its primary objective is to enable machines to automatically recognize and comprehend arguments and their corresponding points in a given text. In other words, the main goal is to understand the points in an argument and get insights into how these points support or oppose each other. This is considered to be a challenging task within the realm of NLP. Having a deeper understanding of the structure of arguments is important for various applications such as debating technologies (e.g. [Slo+21]), legal decision-making (e.g. [Moe+07]), automated essay scoring (e.g. [OLB14]), computer-assisted writing (e.g. [SG17a]), providing tools for policy-making and socio-political sciences (e.g. [BŠ14]), software engineering (e.g. [Wan+20]) and also in business (e.g. [PKY15]).

Based on Lippi and Torroni ([LT16]), the identification of argument structure involves three sub-tasks as follows:

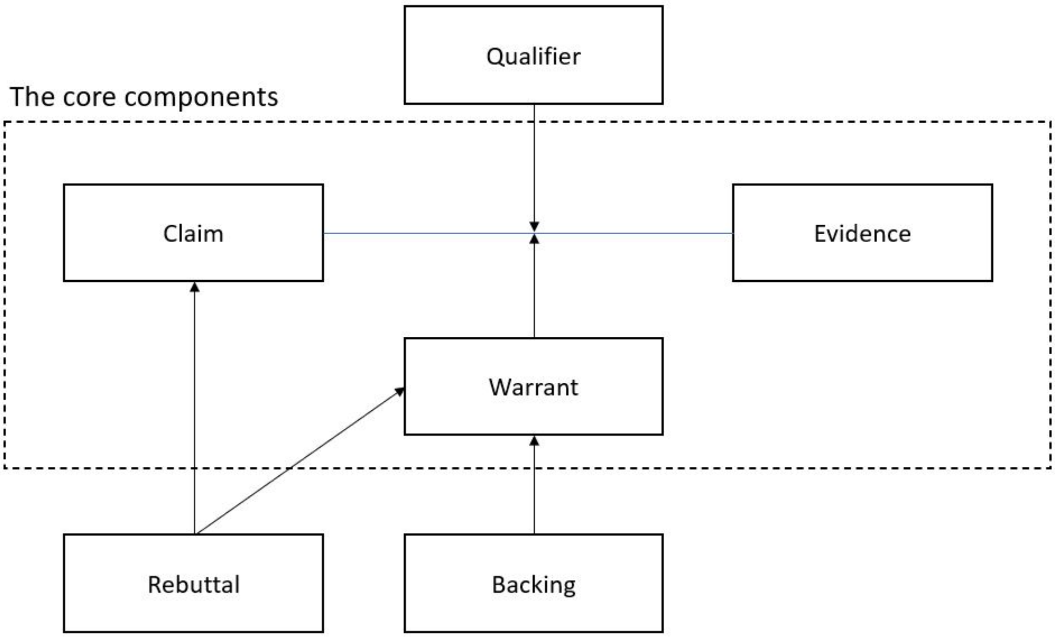


Figure 2.4.: Toulmin's model of argument [Tou03]

1. Determining the "argumentative" vs. "non-argumentative" parts of a text (e.g. [Moe+07; Wan+20]),
2. Classifying argumentative components into categories such as "Claim" or "Premise" (e.g. [SG17a; MM11; Wan+20])
3. Identifying relations between the components or arguments, "attacking" and "supporting" (e.g. [CV13; PM09; SG17a]).

In the first task, argument identification, the main goal is extracting or detecting the parts of documents that contain an argument; in other words, the parts are classified into argumentative and non-argumentative (e.g. [Moe+07; SG17a; PGQ16; Zha+16]). For instance, in Zhang et al. [Zha+16], the main research goal was to design a model that can detect argumentative sentences in online discussions. However, in Poudyal et al. [PGQ16], the focus was on law cases that in terms of formality are completely different from the online discussions. In another research, Dusmanu et al. [DCV17] tackled the first task of argument mining by identifying argumentative sentences in tweets.

In the second task, classifying argumentative components refers to classifying the components of arguments based on a specific scheme (e.g. [MM11; RWB12; SG14; HG17]). Based on the scheme, argumentative parts can be classified into different classes, such as claims and premises (e.g. [SG14]) or claim, backing, rebuttal, premise, and refutation based on Toulmin's model of argument [HG17]. For example, Habernal and Gurevych [SG14; SG17a] proposed a sequence labelling approach in which many different types of features such as lexical, structural, morphological, semantic, and embedding features were used to victories sentences. The authors used SVM^{hmm} [JFY09] which is an implementation of Support Vector Machines specifically used for sequence labelling. Using SVM^{hmm}, each word in documents was separately annotated based on the BIO encoding. This encoding is used to distinguish the boundary of argumentative components, and works as follows: The first word of an argumentative component is labelled with B which means the beginning of the component. The label I is used for the rest of the words in the component. All tokens in non-argumentative components are labelled with O.

The last task in the pipeline of argumentation is discourse analysis which refers to identifying the relations, as for support or an attack, among the argumentative components or specifically the claims and premises in documents [PM09; CV13; BŠ14]. Similar to Zhang et al. [Zha+16], in Boltužić and Šnajder [BŠ14], the authors dealt with online discussions. This work tried to match users' comments to a predefined set of topics and also predicted the relation between the comment and the topic which can be either supported or not supported (for or against).

There are also works that tackle all the levels. Persing and Ng [PN16] was among the first to present an end-to-end pipeline approach to determine argumentative components and their relationship using an Integer Linear Programming (ILP) framework. Similarly, Stab and Gurevych [SG17a] has proposed a joint model that globally determined argument component types and relations using ILP. Eger et al. [EDG17] has presented the first end-to-end neural argumentation mining model eliminating the need for designing hand-crafted features and constraints in all three argumentative tasks. In Wang et al. [Wan+20], the authors deal with online discussions about the usability on issue-tracking systems of open-source projects. Since a large number of issues and comments with different perspectives are posted daily, the contributors of projects face a major challenge in digesting the rich information embedded in the issue-tracking systems to determine the actual user needs and consolidate the diverse feedback. Thus, the authors' ultimate goal was to make usability issues more readable. To do this, they first identified argumentative comments, and then, classified them based on two independent dimensions, argumentative components (the second task of argument mining) and standpoints (the third task of argument mining).

Technologically, a range of machine learning methods has been applied to address different levels of argument mining. For instance, in [Moe+07], a multinomial naive Bayes classifier and maximum entropy model were used as classifiers for detecting arguments in legal texts. They converted the sentences to feature vectors which contained unigrams, bigrams, trigrams, verbs, and argumentative keywords such as "but," "consequently," and "because of," statistical features namely the average of word length and a number of punctuation marks. In Goudas et al. [Gou+14], the authors studied the applicability of some machine learning classifiers on social media text in two steps. First, they identified argumentative sentences by using different machine learning techniques such as Logistic Regression, Random Forest, and Support Vector Machine and second, through using Conditional Random Fields, the boundary of the premises in argumentative sentences was detected. Other machine learning methods such as support vector machine (e.g. [RWB12; Sar+15; HG17]), logistic regression (e.g. [Rin+15; DCV17]), random forest (e.g. [EKKG15; DCV17]), and conditional random field (e.g. [Gou+14; Sar+15]) are also used in argument mining.

In Wambsganss et al. [Wam+20c], an adaptive tool, named AL, by which students received feedback on the argumentative structure of their written text, was designed, built, and evaluated. They tried to answer two research questions that were about the acceptance of AL and how effective it was in writing more persuasive texts. For the latter research question, first, they created two different classifiers by which they identified argumentative sentences and also the relation among them, supported and non-supported. Second, they evaluated the texts by measuring readability, coherence, and persuasiveness. By illustrating these scores and their description, users understood how to improve their texts.

In Shnarch et al. [Shn+17], they presented an algorithm, named GrASP (Greedy Augmented Sequential Patterns), which was weak labelling of argumentative components using multilayer patterns. The algorithm produced highly indicative and expressive patterns by augmenting input n-grams with various layers of attributes, such as name entity recognition, domain knowledge, and hypernyms.

Besides the supervised machine learning approaches that rely on annotated training data, there are unsupervised approaches that eliminate the need for training data. For instance, in [PN16], a novel unsupervised approach was developed that focused on the task of end-to-end argument mining in persuasive student essays collected and annotated by Stab and Gurevych [SG17a]. They applied a bootstrapping method from a small dataset of arguments. They used reliable contextual cues and some simple heuristics, which relied on the number of paragraphs, the location of the sentence, and the context n-grams, for labelling the different components of arguments.

Another unsupervised approach has been presented by Ferrara et al. [FMP17]. Their approach was based on the topic modelling technique. In their research, they focused on detecting argument units that were at sentence-level granularity. Their method, named Attraction to Topics (A2T), had two main steps. The first step was identifying the argumentative sentences and the second step was classifying the argumentative sentences, which were discovered in the first step, to their role, as major claims or the main standpoint, claims, and premises.

Although most of the research in argumentation mining has focused on the English language, Peldszus [PS15] has collected a corpus of microtexts in German and used this corpus for argument component detection. Furthermore, Basile et al. [Bas+16] has studied relation prediction tasks in Italian news blogs. Similarly, there has been some recent work investigating argumentation mining beyond monologues, i.e., looking at the process of argumentation in dialogues. For example, Chakrabarty et al. [Cha+20] has proposed a method to identify the argument structure in persuasive dialogues that can model the micro-level (i.e., the structure of a single argument) and the macro-level (i.e., the interplay between the arguments) characteristics of arguments.

In this thesis, we focused on the first two tasks of argument mining. We identified the argumentative components based on Toulmin's model of argument in user's responses. We compared various machine learning algorithms and also used predefined large language models to identify the argumentative components. We applied the mentioned models in various educational domains within a conversation between a learner or student and an adaptive agent. The main goal of the agent was to analyse the student's argument in order to provide feedback on the missing argumentative components.

2.2.3. Argumentation in Education

Argumentation is the means by which we rationally resolve questions, issues, and disputes and solve problems. Embedding and fostering argumentative activities in learning environments promote productive ways of thinking, conceptual change, and problem-solving [JK10a]. Many science education scholars argue that argumentation is central to scientific thinking [DNO00; DO02; Kuh93; NDO99]. Three different strands of research have been carried out to use argumentation in educational scenarios. One strand of research focuses on including argumentation within teacher education [Erd06; EAYG06]. Such works investigate guidelines and teaching strategies for teachers in order to encourage students to justify better their opinions.

The second strand of research is interested in designing learning environments and course instructions and then assessing their impact on the argumentation skills of students (e.g. [Geo+20]). Georgiou et al. [Geo+20] analysed how the duration of students' engagement in an appropriate teaching environment affects the students' argumentation skills. For this purpose, they conducted an experiment in which 10th-grade Greek state school students were divided into two groups in which both groups received the same teaching materials about Biotechnology but at a different pace. In both groups, the students were asked to express their opinions which were assessed using Toulmin's model of argument. The only difference between them was the number of hours they needed to attend the classes per week.

The third and newer line of research focuses on computer-mediated environments for supporting learners in writing argumentative essays (e.g. [Afr+21; Wam+21; WJL22; Wam+22]). A fundamental motivation underlying such research is the promise of scalability in the face of large class sizes while still being able to give feedback specifically to each learner. The importance of such personal feedback, adapted to each learner's prior knowledge or task performance is in turn a foundational motivation for research in adaptive learning support [Ale17]. Such feedback is of course also important for learning how to argue, as has been found in research investigating teaching strategies for argumentation skills [DV10].

Researchers have used computer-based platforms to evaluate students' essays and offer automated feedback to improve their arguments. For instance, in [Afr+21], a web-based intelligent writing assistant was designed and tested. Four interfaces were developed to determine the most beneficial feedback for students. The feedback varied based on the unit span (sentence or sub-sentence) and the level of surface and content revisions. Surface revisions included feedback on grammar, fluency, and organization, while content revisions involved significant textual changes, such as claims, reasoning, and rebuttals based on Toulmin's argument model [Tou03]. Upon comparison, the interface that displayed details of surface and content revisions at the sentence level was found to be the most effective.

In [WJL22], the impact of automated feedback with social comparison on students' essays was investigated. They showed that triggering the basic psychological processes such as comparing to the social norm could lead to writing more convincing essays. To end this, they conducted an experiment in which three groups of undergraduate business students (71 students) were compared based on their average number of arguments based on Toulmin's model of argument. The average number of arguments for the groups who received automated feedback with a social comparison nudge was 4.78. However, the average number of the group who received only the feedback without the social comparison was 3.9 and for the last group who only had access to syntactic feedback was 3.64.
In [Wam+21], a tutoring system, called ArgueTutor, was developed to help students to write more convincing essays. The students' task was, first, to read the debate of two teachers on a specific topic and then write an argumentative essay. The system was turn-based. In each round, first, an argumentative essay should be written by a student and then ArgueTutor analysed the essays and gave feedback by deep learning methods. After each round, the students had the chance to improve the essay based on the received feedback. The feedback received by students constituted a short summary based on the number of argumentative components based on Toulmin's model of argument and a readability score which is calculated based on [Fle43].
In another study by Wambsganss et al. [Wam+22], they developed a system called ArgumentFeedback, based on nudging theory, to provide adaptive self-evaluation for students' argumentation skills based on Toulmin's model of argument. They conducted three studies involving 83 students to investigate the impact of individual argumentation self-evaluation on students' writing. The results showed that students who received the self-evaluation nudge produced more convincing texts with improved formal and perceived argumentation quality compared to the control group. The findings suggest that combining nudging-based learning applications and computational methods for argumentation self-evaluation can effectively enhance students' writing skills in traditional learning environments, promoting self-regulated learning. The topic of the studies was business pitches and the required data (a corpus of 200 student-written business model pitches in German) to identify the argumentative components was collected in a mandatory business model innovation lecture at a Western European university.

Besides the mentioned works which focused only on giving feedback on students' essays, in [Xia+22], the authors aimed to develop a tool, called Persua, to help users enhance the persuasiveness of their arguments in online discussions. They designed an interactive visual system that provides guidance on persuasive strategies by visualising the relationship between claims and different types of premises. They employed machine learning models to detect components of argumentative writing and their relationships and provided example-based guidance on persuasive techniques to improve the persuasiveness of arguments. To do that, they created a labelled dataset of persuasive strategies in arguments from ChangeMyView forum. A between-subjects study showed that Persua encouraged users to submit more times for feedback and helped them to improve the persuasiveness of their arguments compared to a baseline system.

In this thesis, our attention was directed towards the third area of research. Computer-mediated environments offer the advantage of remote accessibility, particularly beneficial in pandemic scenarios. Moreover, thanks to recent progress in NLP and AI, this technology not only supports a growing number of learners but also offers personalized feedback, enriching the overall learning process.

2.2.4. Argumentative Conversational Agents

Beyond educational context, conversational agents have also been studied as discussion partners for general argumentation. Such agents do not have an educational goal and they focus on having an argumentative dialogue. In general, the available argumentative conversational agents can be persuasive [TG17; CH20] or just convey the information by offering arguments that keep the dialogue comprehensive and meaningful [LNN18; Rak+19]. For instance, in [CH20], a conversational agent was developed that tried to persuade its audiences regarding a specific topic such as meat consumption. The agent selected an argument from its knowledge base which related to the audience's concerns to increase the chance of persuasion. The knowledge of the agent, which was collected by a crowdsourcing method, was a list of arguments and counterarguments about the topic.

In [Rak+19], a retrieval-based agent, named Debbie, has been presented. Their agent talked to its audiences about three topics: the death penalty, gun control, and gay marriage. The main goal of the agent was to keep the meaningful conversation by presenting various arguments until it ended by users.

Following the literature, we used conversational agents as a computer-mediated environment but to complete the literature, we focus on argumentation per se. Within the conversation between a learner and an agent, we focus on analyzing and giving feedback on the structure of the argument written by the learner which is novel w.r.t. the above-discussed literature.

2.3. Bloom's Revised Taxonomy and Argumentation

Bloom's revised taxonomy [AK01] attempts to classify learning stages as a consecutive process from remembering facts to creating new ideas based on the acquired knowledge. Based on the taxonomy, the learning stages are as follows: Remembering, Understanding, Applying, Analysing, Evaluating and Creating (as shown in Figure 2.5)

In this work, the developed conversational agents delivered the learning materials following Bloom's revised taxonomy [AK01]. The first two levels of the taxonomy are Remembering and Understanding. The developed agents cover these levels by providing required content and information and asking learners to recall them in the curriculum. The next four levels of taxonomy, considered as higher-order thinking skills, are Applying, Analysing, Evaluating and Creating. In this work, we handled these four levels by asking argumentative questions and then by providing feedback on learners' responses. For instance, an agent can first initiate discussions that prompt learners to recall and understand information (Remembering and Understanding). Secondly, learners are asked to answer an argumentative question through the application (Applying) and analysis (Analyzing) of the question and the provided information. Third, learners answer the

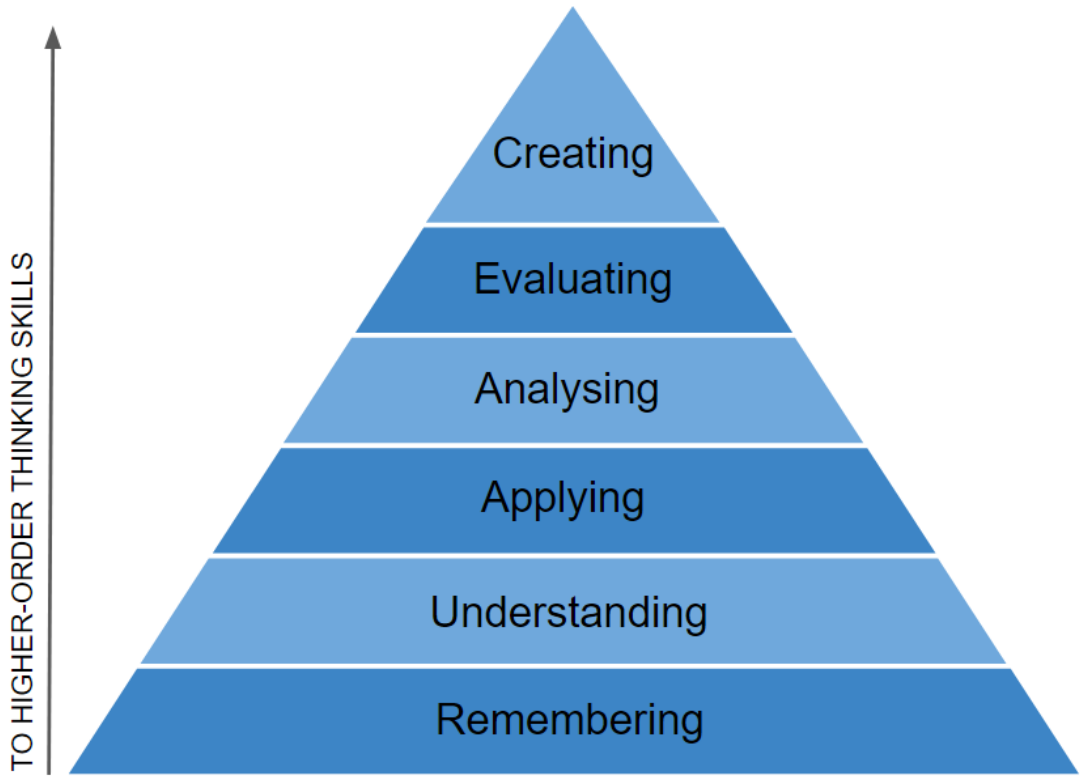


Figure 2.5.: Bloom's revised taxonomy [AK01].

question and justify it (Evaluating). Finally, the conventional agent analyses the learner's answer and provides feedback based on Toulmin's model of argument in order to support the learner in writing a complete argument based on Toulmin's model (Creating).
To summarise, the agent starts from lover-order to higher-order thinking skills of Bloom's revised taxonomy [AK01] first by asking learners to recall facts and concepts, second, by asking questions about the facts which push learners to apply, analyse and evaluate their answers, and finally, by providing feedback on their answers based on Toulmin's model of argument [Tou03].

The combination of Bloom's revised taxonomy [AK01] and Toulmin's model of argument [Tou03] in conversational agents represents a powerful synergy between instructional design and argumentation. This integration not only provides a diverse range of cognitive processes but also enhances the skills of effective reasoning and articulation. As we delve into the era of intelligent conversational agents, this combined approach promises to elevate the quality of educational interactions, fostering a holistic learning experience that is more than just content delivery.

2.4. Synthesis

The main goal of this thesis is to utilise argumentation within a conversation with an agent. The educational conversational agent augments the learners' educational experience by asking argumentative questions in various learning domains and also by providing adaptive feedback on structural flaws in the learners' arguments according to Toulmin's model of argument [Tou03]. This is achieved in two parts. In the first part, the focus is on identifying deficiencies in learners' structural argumentation and improving their argumentative knowledge with the help of an educational conversational agent. The previous studies utilised conversational agents to enhance the argumentation skill of learners by providing feedback mainly on essays written by learners (e.g. [WJL22; Wam+21; Afr+21]). Argumentative essays follow a specific structure and are written in paragraphs each of which has its own role. The structure of the essays can be used as a feature (the structural features) to identify the argumentative components. Structural features describe the position and frequency of text elements, including token position, punctuation, and argument components. In essay structure, the statistics such as the number of tokens in an argument component or an argumentative unit can be considered as an indicator. Stab and Gurevych [SG17b] reported the efficacy of these features in both identifying argumentative structures and classifying argumentative components. Besides identifying the argumentative components in structural essays, the feedback provided by an agent was based on the number of written argumentative components such as claims and premises in the students' essays (e.g. [WJL22; Wam+21]).
In this thesis, the focus is on (short) argumentative statements as would be expected in the context of a conversation with an intelligent agent in a given domain, and not on essays. The main challenge in providing argumentative feedback during a conversation is that the dialogue turns or the learners' statements do not have a specific structure to be used as an indicator in identifying the argumentative structure. In addition, in this work, the feedback is based on the missing argumentative components or on structural gaps in the learners' responses and not only the number of each component. Giving feedback on a missing argumentative component is more challenging than giving feedback on the number of argumentative components in argumentative essays. In this study, the agent's feedback requires to be more specific and point out the missing argumentative component to support learners to fulfil the gaps. These are the main challenges addressed in this thesis and the contributions complement existing research that has worked on longer argumentative essays (e.g. [Afr+21; WJL22; Wam+21]), that has differently conceptualized argumentation quality which is less suitable for direct feedback within a conversational agent.

This thesis complements the literature first by showing what fully tutorial conversational agents look like and then injecting argument into such agents by asking argumentative questions in various domains and providing feedback on structural gaps in learners' arguments. For this purpose, the thesis first discusses the characteristics of a fully tutorial conversational agent as an educational technology and assesses the user experience compared to other educational technologies. It was in two steps as follows:

1. How does a fully tutorial conversational agent that leverages various learning activities look like? The learning activities supported by the agent include transferring knowledge, having short quizzes with different types of questions, giving personalised and argumentative feedback, and summarising content. These learning activities are supported through a systematic instructional dialogue design based on Bloom's revised taxonomy [AK01] of learning objectives (see Chapter 5).

2. An analysis of user experience and user preferences in using two different methods of learning content delivery: freely navigable web-based content and a tutorial and dialogic interaction (see Chapter 5).

Second, the thesis discusses how to have argumentative dialogue turns in such agents as follows:

1. Toulmin's model of argument, as outlined in [Tou03], can be highly beneficial when incorporated into a conversational agent. This agent can assist users in constructing an argument that encompasses Toulmin's key components, including claims, warrants (underlying understanding), and evidence, all within a specific domain (see Chapter 6).

2. The impact of using such an agent and its feedback on students' argumentation skills. This study complements existing research on computational environments for teaching argumentation in longer texts (Chapter 7).

3. A systematic workflow consisting of a dialogue structure and an adaptation mechanism to be able to ask argumentative questions and give adaptive feedback in various learning domains based on Toulmin's model of argument using pre-trained large language models (see Chapter 8).

3. Thesis Structure and Contributions

The work in this thesis builds upon previous research and aims to enhance learners' educational experience by using argumentation as a teaching technique in fully tutorial conversational agents. The thesis is divided into two parts: The first part focuses on understanding the attributes of fully tutorial conversational agents as an educational technology which follows Blooms's revised taxonomy [AK01] and evaluating the user experience in comparison to other educational technologies. The second part concentrates on injecting argumentation in such agents defined in the first part. To do that, we first identified structural weaknesses in learners' arguments within a conversation with an agent and then the agent provided adaptive feedback on structural argumentative flaws in learners' responses following Toulmin's model of argument [Tou03].

The work in this thesis builds upon and synthesizes our previous research published in the following sources: The first two papers in this thesis have a common objective corresponding to the first objective of the thesis which is exploring the characteristics of fully tutorial conversational agents as educational technology and evaluating the user experience in comparison to other educational technologies. The goal is to gain insights into the effectiveness and usability of these conversational agents in enhancing learners' educational experiences. The first two papers are as follows:

3.1. Paper 1

> Mirzababaei, B., & Pammer-Schindler, V. (2022). An Educational Conversational Agent for GDPR. *In European Conference on Technology Enhanced Learning* (pp. 470-476). Cham: Springer International Publishing (Chapter 4)

This paper describes and shows what a fully tutorial conversational agent looks like. We designed and developed two prototype conversational agents that utilises various learning activities such as conveying the content, asking questions, giving adaptive feedback and summarising the learning materials at the end of the conversation following the revised version of Bloom's taxonomy [AK01]. This overall design of the dialogue structure allows for inserting more specific adaptive tutorial strategies. From a learner perspective, the learners experience a completely one-on-one tutorial session in which they receive relevant content (is "being taught") as well as experience active learning parts such as doing quizzes or summarising content. Our prototype, therefore, illustrates a move away from the dichotomy between content and the activity of teaching/learning in educational technology. Specifically, we developed two tutorial conversational agents, the first agent is called GDPRAgent which teaches a lesson on the European General

Data Protection Regulation (GDPR). This regulation governs how personal data must be treated in Europe. Instructionally, the agent's dialogue structure follows a basic GDPR curriculum and uses Bloom's revised taxonomy of learning objectives in order to teach GDPR topics. The second agent is called PETs Agent and it talks about privacy-enhancing technologies following Bloom's revised taxonomy.

We see the main contribution of this paper to existing research on conversational agents in education in the systematic instructional dialogue design, based on Bloom's revised taxonomy of learning objectives [AK01]. We aim to show that this structure can also be used in other domains than the GDPR and privacy-enhancing technologies. Further, we inject a few simple teaching (instructional perspective) and adaptation (technical perspective) strategies. A systematic guideline for educational conversational agent developers would be helpful that summarises which teaching strategies can be inserted in a single conversational agent lesson.

This paper addressed the first objective of the thesis which is showing the characteristics of a full tutorial conversational agents. Our prototype agents, therefore, illustrate a move away from the dichotomy between content and the activity of teaching/learning in educational technology.
The first author's contributions consist of writing the paper, literature research and developing the agent.

3.2. Paper 2

Mirzababaei, B., Maitz, K., Fessl, A., & Pammer-Schindler, V. (2023). Interactive Web-Based Learning Materials Vs. Tutorial Chatbot: Differences in User Experience. *In European Conference on Technology Enhanced Learning* (pp. 213-228). Cham: Springer Nature Switzerland (Chapter 5)

In the subsequent paper, we delved further into the first objective of the thesis and examined the distinctions in user experience when utilizing web-based interactive materials versus engaging in a conversation guided by a computational tutorial agent or a completely tutorial conversational agent similar to the one introduced in the preceding paper (Section 3.1). These two educational technologies materialise different interaction metaphors in the sense that users (=learners) can interact in different manners with what should be learned through this interaction. To analyse the differences in user experience and preferences, we conducted a study with 31 master students of inclusive education. One group interacted with web-based textual learning materials (DIGIVIDget condition, $n = 14$). The other group interacted with a tutorial agent (DIGIBOT condition, $n = 17$). Both groups received the same text-based content on formulating search queries for the Internet. Subsequently, two focus group discussions were carried out with participants who had also tested the respective other technology ($n = 12$).

The contribution of this paper towards literature lies in revealing differences in the user experience and preferences between these two educational technologies. Furthermore, the focus group discussions with participants who have experiences in inclusive education highlighted the potential of each educational technology in inclusive education for learners

with special needs or learning difficulties.
The first author's contributions consist of writing the paper, literature research, data analysis and the development of the conversational agent.

In the first two papers, we dealt with the first objective of the thesis which is comprehending the characteristics of a full tutorial conversational agent. Such agents can support various teaching techniques from conveying the learning materials step by step to having quizzes and giving adaptive feedback. In the next three papers, we addressed the second objective of the thesis which is using argumentation as a teaching strategy and injecting it as one sub-dialogue in the full tutorial conversation agents.

3.3. Paper 3

> Mirzababaei, B., & Pammer-Schindler, V. (2021). Developing a conversational agent's capability to identify structural wrongness in arguments based on Toulmin's model of arguments. *Frontiers in Artificial Intelligence*, 4, 645516 (Chapter 6)

This paper discusses the usefulness of Toulmin's model of arguments [Tou03] as structuring an assessment of different types of wrongness in an argument. To support learners in developing a structurally complete argument based on Toulmin's model of argument, we present a study and the development of classifiers that can identify the existence of Toulmin's core component, namely a claim, a warrant (underlying understanding), and evidence in users' responses. Based on a dataset (three sub-datasets with 100, 1026, and 211 responses in each) in which users argue about the intelligence or non-intelligence of entities, we have developed classifiers for these components: The existence and direction (positive/negative) of claims can be detected a weighted average F1 score over all classes (positive/negative/unknown) of 0.91. The existence of a warrant (with warrant/without warrant) can be detected with a weighted F1 score over all classes of 0.88. The existence of evidence (with evidence/without evidence) can be detected with a weighted average F1 score of 0.80.

The contributions that this article makes toward state-of-the-art are 1) to give evidence that reasonably accurate classifiers can be built for the existence of single components of Toulmin's model of arguments in (short) argumentative statements as would be expected in the context of a conversation with an intelligent agent in a given domain, 2) to show by an argument that such classifiers are useful within dialogue structures of conversational agents that are designed based on Bloom's revised taxonomy for learning [AK01], and 3) to show by argument how the same conceptual structure of Toulmin's model of argument can be used to further structure the identification of more complex types of faulty argumentation. These contributions complement existing research that has worked on longer argumentative essays (e.g. [Wam+21; Afr+21]), which has differently conceptualized argumentation quality that is however less suitable for direct feedback within a conversational agent, and broader work on argumentation mining on identifying groups of similar arguments [WSA17] or conversational agents for factual teaching [Rua+19; Gro+19].

The first author's contributions consist of writing the paper, data collection, data analysis, literature research, development and evaluation.

3.4. Paper 4

> Mirzababaei, B., & Pammer-Schindler, V. (2022). Learning to Give a Complete Argument with a Conversational Agent: An Experimental Study in Two Domains of Argumentation. *In European Conference on Technology Enhanced Learning* (pp. 215-228). Cham: Springer International Publishing (Chapter 7)

In this paper, we continued with the learning domain (the intelligence of entities) and the classifiers developed in our previous paper (see above). This paper reports a between-subjects experiment (treatment group $n = 42$, control group $n = 53$) evaluating the effect of a conversational agent that teaches users to give a complete argument and how the users can transfer the acquired knowledge to another learning domain. Utilising the classifiers, the agent analyses a given argument for whether it contains a claim, a warrant and evidence, which are understood to be essential elements in a good argument based on Toulmin's model of argument [Tou03]. The agent detects which of these elements is missing, and accordingly scaffolds the argument completion. The experiment includes three tasks. A treatment task (Task 1) in which participants of the treatment group converse with the agent and receive feedback on their argumentative answers, and two assessment tasks (Tasks 2 and 3) in which both the treatment and the control group answer an argumentative question. The results show that the participants who received feedback from the agent in Task 1, not only can use the acquired knowledge in the same domain (Task 2), but also transfer the knowledge and use it in another learning domain (Task 3).

The contribution of this paper towards literature lies in answering the question of whether a tutorial conversation leads to the learning of the taught argumentation structure (Toulmin's model) and applying the acquired knowledge in different learning domains. This complements existing work on computational environments for teaching argumentation in longer texts (e.g., [Afr+21; Wam+21]). Both the underlying computational methods needed to understand and feedback arguments are different, due to different lengths and styles of argumentation in essays and in conversations.

In alignment with the second objective of the thesis and taking a further stride forward, this paper (the fourth paper) delved into examining the influence of argumentative feedback, utilizing the models developed in the third paper (Section 3.3), on the writing of argumentative essays within two distinct domains.
The first author's contributions consist of writing the paper, data collection, data analysis, literature research, development and evaluation.

3.5. Paper 5

(*under review- Accepted with Minor revision*) Mirzababaei, B., & Pammer-Schindler, V. (2023-2024) Facilitating the Learning Engineering Process for Educational Conversational Modules using Transformer-based Language Models (Chapter 8)

In Paper 3, we showed how Toulmin's core components [Tou03] can be identified in learners' responses to that specific question and the topic (whether an entity is intelligent). Following that, in Paper 4, we used the models, developed in Paper 3, in a conversation to assess the impact of argumentative feedback on writing argumentative essays in two different domains. Both Papers 3 and 4 focused on one specific topic which is the intelligence of entities. To have such classifiers and conditional dialogues for various questions and learning domains, new classifiers with all reacquired machine learning engineering should be developed. In the last paper (Paper 5), we generalised our previous works and presented a systematic workflow to be able to develop a conversational module containing classifiers and dialogue structures in various learning domains. The primary purpose of this paper is to evaluate how good the conversational agent modules can be expected to be if a learning engineer follows our workflow. We, therefore, evaluated the workflow by separately evaluating three different conversational modules. For each module, we assessed 1) the classifier quality, as well as 2) how coherent the follow-up question that the agent would ask due to the classification results with regards to the user response.

Our work makes significant contributions to utilising argumentation in education by introducing a systematic workflow to develop a module that enables a conversational agent to ask argumentative questions on various topics and provide personalised feedback while minimising the technical requirements. This workflow aligns with the educational goals of teaching concepts, definitions, or terminologies through argumentation. The developed conversational module can be integrated into conversational interfaces or educational conversational agents, such as Quizbot [Rua+19], MathBot [Gro+19] or any educational conversational agents which convey knowledge and ask questions. By incorporating this module, these conversational agents gain the ability to pose argumentative questions and deliver adaptive feedback based on Toulmin's core components.

This paper completes the second thesis objective by presenting a systematic workflow which generalises the dialogues used in the previous papers. The workflow allows us to develop a conversational module in various learning domains. The module contains an argumentative question asked by an agent and adaptive feedback on the users' answers based on the missing argumentative components.
The first author's contributions consist of writing the paper, data collection, data analysis, literature research, development and evaluation.

4. Demonstrating a Full-Tutorial Conversational Agent with Multifaceted Educational Capabilities

4.1. Introduction

Following the first goal of the thesis, in this chapter, we demonstrate the multifaceted capabilities of a tutorial conversational agent. The agent covers a range of learning activities essential for functioning as a comprehensive tutorial conversational agent. The learning activities handled by the agent include transferring the learning materials in different forms, having short quizzes, providing personalised feedback based on learners' interactions, and summarising the content. From the learners' standpoint, this means not only receiving the necessary information but also gaining the ability to assess their own progress.

Large-scale learning scenarios as well as the pandemic situations underline the importance of educational technology in order to support scalability and spatial as well as temporal flexibility in all kinds of learning and teaching settings. Educational conversational agents build on a long research tradition in intelligent tutoring systems and other adaptive learning technologies but build for interaction on the more recent interaction paradigm of conversational interaction. Educational conversational agents, often referred to as virtual tutors or digital mentors, are sophisticated software programs designed to engage learners in dynamic, natural-language conversations. They possess a wide array of capabilities that go beyond conventional textbook-based instruction, offering personalised instant feedback, and adaptive learning experiences.

Wollny et al. [Wol+21] have categorised research on educational conversational agents into four main groups: Application, Design, Evaluation, and Educational Effect. In the design group, one of the subcategories focuses on the interaction between agents and learners (e.g. [PM21]). In this subcategory, researchers primarily concentrated on analysing how the agent's feedback or actions influence the learners' experience and learning outcomes. The educational scenarios defined in such research consist of dealing with one specific task which the agent initiates by posing a question or describing a problem. Subsequently, learners respond with their answers or proposed solutions, and finally, the agent assists them in completing the task by providing tailored feedback and helpful hints (e.g. [Gro+19; Gra+99]). In such scenarios, presenting the learning materials is not part of the conversation between the agent and learners. In other words, the content required to solve the problem or the instructional phase is not part of the scenario.

In contrast to the agents and scenarios mentioned earlier, our emphasis lies on illustrating the features of a comprehensive educational conversational agent with which learners also receive the learning materials. A full tutorial conversational agent seamlessly incorporates a diverse range of functionalities, smoothly transitioning from content delivery to interactive exercises, quizzes, providing instant feedback, and ultimately summarising the covered learning materials. In this chapter, we provide detailed descriptions of two such full tutorial conversational agents that cover two different topics:

1. GDPR Agent: This agent is designed to deliver a lesson on the European General Data Protection Regulation (GDPR), which dictates the proper handling of personal data in Europe. Instructionally, the agent's dialogue structure follows a fundamental GDPR curriculum and leverages Bloom's revised taxonomy [AK01] of learning objectives to effectively teach the core GDPR topics.

2. PETs Agent: This agent is dedicated to conveying learning materials about privacy-enhancing technologies (PETs). It comprehensively covers the four primary technologies and the situations in which each one is applicable. In an era characterised by digital interconnectivity, protecting personal information has become crucial, especially for companies entrusted with the responsibility of protecting sensitive and private data.

The chapter is organised as follows: The next section covers a brief pedagogical and technological background is covered. Following that, the two developed agents, GDPR Agent and PETs Agent, are explained in Section 4.3. Finally, the chapter is concluded in Section 4.4.

4.2. Pedagogical and Technological Background

Lifelong learning is necessary for individual, organisational and societal success and well-being. Simultaneously, the growing number of learners in educational settings or in workplaces, coupled with what often appears to be insufficient resources, presents a challenge in delivering a high standard of personalised and interactive instruction [Bri+14; Kim01]. Educational technology has long been investigated as a means to address this challenge.

We are particularly interested in the promise of conversational agents that act as tutors. Conversational agents constitute a human-computer interaction paradigm in which people can interact - so the ideal - in a relatively natural (for humans) way in natural language with technology. Ideally, with conversational agents, a learner can discuss concepts in a learning domain, move from talking about basics toward core complex definitions, do a self-assessment by answering questions, and receive feedback. This is what good educators do, given sufficient resources to interact bilaterally or with small groups of learners.

Much research in artificial intelligence for education has gone into developing computational systems that are able to, at least partially, fulfil some of these functions that (good) human tutors take on. Such systems are typically called intelligent tutoring systems [GV00; Koe+97]. More recently, researchers have investigated tutorial conversational agents, e.g., for question answering [Cla+18], helping students to efficiently use a large body of content [Akc+18], helping learners in assessing their own abilities [DK20], and providing administrative services such as answering students' questions on behalf of the academic faculty [Hie+18]. Many conversational agents that focus on teaching a topic are of course domain-specific, and by now research efforts span a plethora of subjects such as mathematics [ATR14], medicine [Jag+21], computer science [Mit03], physics and chemistry [PMB13]. Typical research foci have been on the agents' architecture, how to model learners' knowledge, different communication methods such as text or voice, or the impact of the appearance of agents on learners. Complementing such works, our research emphasis is on designing tutorial dialogues which cover learning materials by following Bloom's revised taxonomy of learning objectives [AK01] and how to insert different teaching strategies such as argumentation into this overall structure.

4.3. Description of the Prototype

In this section, we present two conversational agents, named GDPR Agent and PETs Agent, that carry out a complete tutorial conversation. The agents cover the complete content of a lesson step by step, ask questions after each step, give feedback on learners' answers, and summarise the provided content. The Agents thereby simulate a whole learning session in a one-to-one situation between a teacher and a learner.

The agents deliver the learning materials following Bloom's revised taxonomy [AK01]. The first two levels of the taxonomy are Remembering and Understanding. The agents cover these levels by asking learners to recall and explain a concept in the curriculum. The next three levels of taxonomy are Applying, Analysing and Evaluating. The agents cover these levels by asking questions in which a new situation is described. To answer the questions, the learners need to apply the already learned concepts to the new situation (Applying level) and the connection between the questions and their answers (Analysing level). To go further in Bloom's revised taxonomy, the agents ask the learners to justify their answers. Explaining or arguing covers the Evaluating level of the taxonomy. To summarise, the agents start from lover-order to higher-order thinking skills of Bloom's revised taxonomy first by asking learners to recall facts and concepts, second, by explaining the facts and concepts, and finally by asking questions about the facts which push learners to apply, analyse and evaluate their answers using Toulmin's model of argument.

The following subsections describe two implemented full-tutorial conversational agents, GDPR Agent which covers the basic GDPR content and PETs Agent which deals with four privacy-enhancing technologies.

4.3.1. GDPR Agent

GDPR Agent delivered the content in four steps, which topic-wise, we created to represent a typical introductory GDPR curriculum. First, it starts with a greeting and shows the agenda (Figure 4.1). Second, the agent talks about what is the GDPR and where and when it should be applied. Third, the definition of personal data and sensitive data,

Figure 4.1.: The responsive web page in which the greeting section of the conversation and the agenda are shown.

and their differences are covered. Finally, the seven data protection principles of the GDPR are presented. Figure 4.2 shows the order of the content in the whole dialogue. The GDPR content has been synthesised based on various online resources including the authoritative GDPR information[1]. Especially, we have benefited from FutureLearn[2], based on the open content license for non-commercial purposes, for the question parts of the dialogue, Step 3.4 and 4.2.

Different teaching strategies are embedded into the conversation. In this prototype agent, we followed the revised Bloom's revised taxonomy of learning objectives [AK01]. For instance, in the third step of the conversation which is about personal and sensitive data (See Figure 4.2), the agent first covers the remembering and understanding levels of Bloom's taxonomy by asking learners to recall and then write about the definitions of personal and sensitive data. Second, by focusing on the differences between these two types of data and asking learners some questions in which the learners need to apply the information in various scenarios, the agent covers the applying and analysing. Finally, the agent addresses the evaluating level by having an argumentative conversation in which the learners are asked to justify their answers. At this point, we therefore also followed the teaching strategy of learning through argumentation [JK10a]. Learning through argumentation guides learners to analyse a problem from various perspectives and also to distinguish what is correct and incorrect. Figure 4.3 shows how the agent asked follow-up questions in order to guide the learner to find out why the selected option was incorrect. Here the agent asked the user to justify his answer and then the agent explained a situation in which the user's answer was not valid. In general, the agent adapts to learners' responses. Based on each response, the agent asks the learner to think again about their own response and justify it and then, in case of selecting a wrong

[1] *https://eur-lex.europa.eu/homepage.html*
[2] *https://www.futurelearn.com/courses/general-data-protection-regulation*

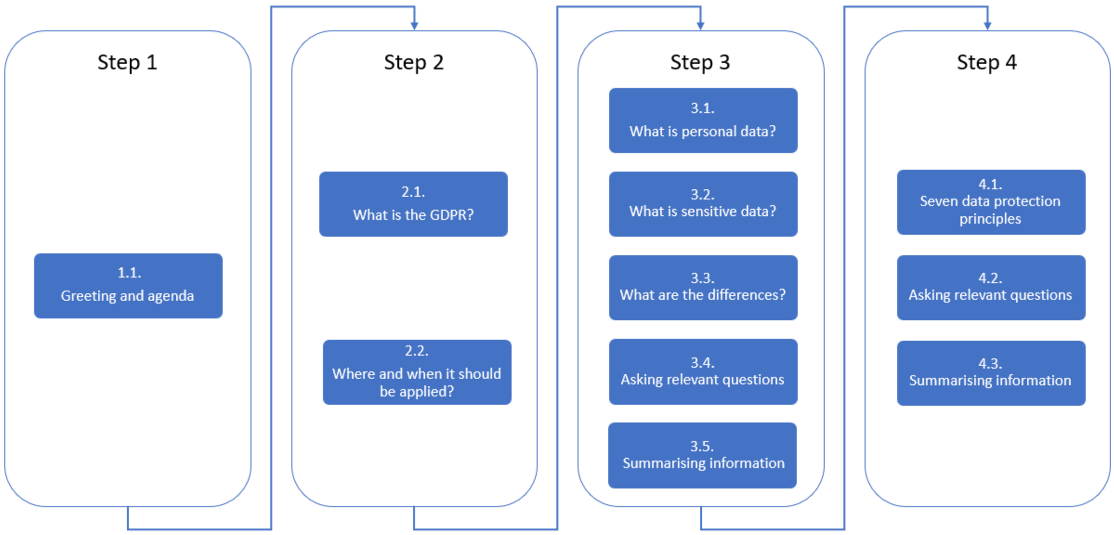


Figure 4.2.: The flow of the conversation. It includes four subsections: greeting, the GDPR and its scope, personal and sensitive data, and seven data protection principles

answer, the agent explains a situation in which the user's argument is not valid anymore. The mentioned example already shows, how an overall instructional design following Bloom's revised taxonomy of learning objectives allows and actually needs the insertion of more specific teaching strategies. Note that from an instructional perspective "teaching strategies" are inserted, whilst this means that from a technical perspective "adaptation mechanisms" need to be inserted. Here, the full spectrum of intelligent tutoring and adaptive teaching systems [Ale17] is available to conversational agent designers.

In the current prototype, we have inserted two more adaptation strategies. The first is to adapt to learner knowledge (cp. [Ale17], the taxonomy of adaptation strategies), and to exercise what a learner does not know. We did this in steps 3.4 and 4,2 (See Figure 4.2), such that when the agent asks questions about the different types of data (Step 3.4) and about data protection principles (Step 4.2), the number of questions for each learner depends on the number of his or her correct answers. We defined five different questions for Step 3.4 and six questions for Step 4.2, but the agent first asked three questions. In each step, if a learner answers at least two questions, the agent asks the learner to answer more questions. In case of agreement, the agent asks the rest of the questions. Otherwise, the conversation is continued. The second adaptation strategy could be understood as an adaptation that targets learners' affect. At the beginning of presenting a new topic, the agent asks what the learner already knows about the topic. In Figure 4.4 for instance, the agent asks the learner, about the GDPR's scope before giving the information. The agent is to some extent adaptive to the learner's responses. The agent uses keyword matching in order to understand the learners' responses. For each topic, a set of keywords is defined which helps the agent to have an adaptive reaction. For instance, in Figure 4.4, since the user did not know the answer, the agent gave encouraging feedback in order to motivate

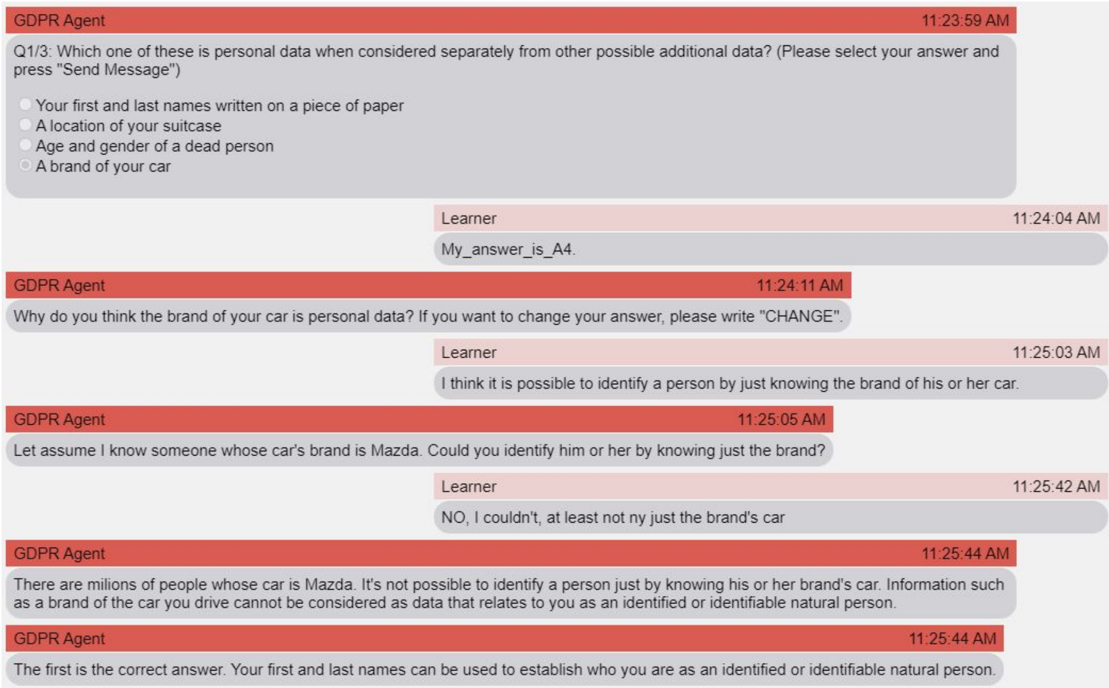


Figure 4.3.: The agent asks argumentative follow-up questions before showing the answer.

the learner. If the agent does not understand the user's responses, it will try to keep the conversation coherent and meaningful by giving a proper reply.

Technically, we have implemented GDPR Agent based on the open-source Bazaar framework [Ada+14] as back-end[3], and as an HTML/JS responsive web page for the front-end. This framework allows both rule-based and machine-learning-based classifiers to decide between dialogue branches. The complete dialogue including the content, questions, and the agent's feedback is presented in Appendix A.1.

4.3.2. PETs Agent

The second agent, called PETs Agent, delivers the content about privacy-enhancing technologies (PETs). The topic is important, especially in workplaces which deal with personal and sensitive data. Privacy is considered a fundamental human right. PETs protect individuals from unwarranted intrusion into their personal lives, allowing them to maintain autonomy and control over their personal information. In a broader perspective, GDPR or other data protection laws require businesses and organizations to implement privacy measures. PETs play a crucial role in achieving compliance with these regulations. Overall, PETs are a crucial component of a balanced and ethical approach to data collection, processing, and storage in the digital age. They help strike a balance between the benefits of technological progress and the protection of individual privacy

[3] *https://github.com/DANCEcollaborative/bazaar*

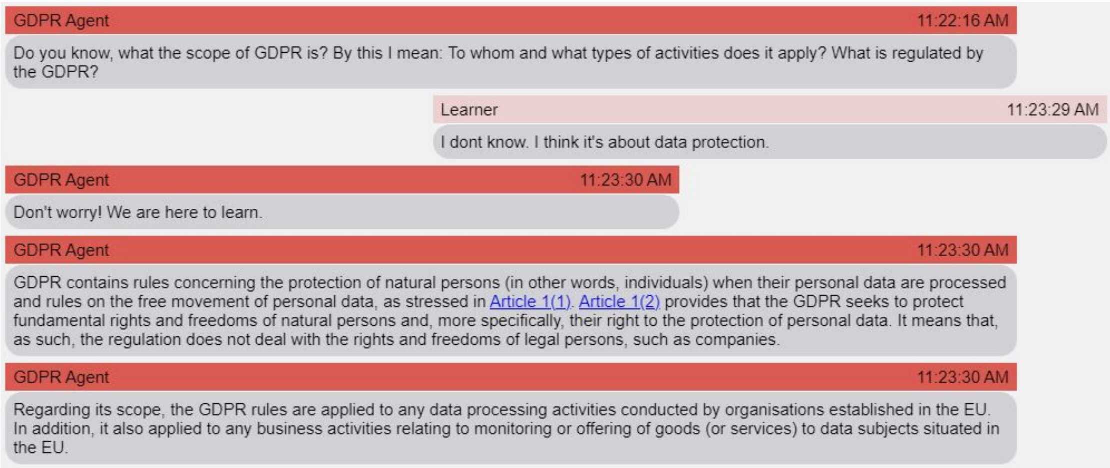


Figure 4.4.: The agent asks for the learner's idea about the GDPR's scope before teaching it. The agent is also adapted to the learner's responses

and rights.

Much like the GDPR Agent, the PETs Agent adheres to Bloom's revised taxonomy for content delivery [AK01]. The agent introduces topics systematically, beginning by inquiring about learners' existing knowledge on the subject and subsequently providing explanations. Following content delivery, the agent engages learners with questions and offers adaptive feedback.

The PETs content is covered in five steps. STEP 1: The first step is the greeting and the agenda. STEP 2: The second step deals with what these technologies are and why they are important. Since the topic is about protecting privacy, the agent also briefly presents the definitions of personal and sensitive data. STEP 3: The third step starts with the definitions of two main categories of PETs, the crypto-based and ML-based methods and finishes with the definitions of four main technologies, federated learning, differential privacy, homomorphic encryption and multiparty computation. STEP 4: In the fourth step, the learners are asked several questions about PETs. In each question, an imaginary situation is described in which a company or person needs to select one of the technologies as a suitable and feasible solution based on the situation. For example, one of the questions is as follows: "*Suppose there are two companies each having a private data set and it would be beneficial to them to find common entries in these sets without having to show the other party the whole data. What method might offer such a functionality?*" The user has four options to select and the agent gives different feedback based on the selected option. STEP 5: Finally in the last step, the user has the chance to talk about a specific use-case to find out which technology would be feasible. To do that, the agent starts asking questions about the characteristics of the learner's use-case to recommend the possible feasible technologies. For instance, the questions are about the number of involved parties and what should be protected (data or algorithms). The complete dialogue including the content, questions, and the agent's feedback is presented in Appendix A.2.

In addition to the two agents previously discussed, we also created a novel agent designed to instruct learners on writing a concrete search query, a crucial component of information and data literacy as outlined in the DigComp 2.2 [VRKP22] and DigCompEdu [Red+17] frameworks. This agent is thoroughly explained in the upcoming chapter (Chapter 5).

4.4. Conclusion

In this chapter, we showed educational conversational agents can do more than just supporting users to finish a specific task. Delivering the learning materials or teaching them can be part of the conversation handled by agents. Ideally, with conversational agents, a learner can discuss concepts in a learning domain, move from talking about basics toward core complex definitions, do a self-assessment by answering questions, and receive feedback. This is what good educators do and having such agents with the mentioned capabilities can enhance the learning experience. Educational conversational agents such as the GDPR and PETs Agents provide a learning journey in which the users experience a complete learning experience from receiving the content and having quizzes to evaluating themselves following Bloom's revised taxonomy [AK01].

We see the main contribution of our research to existing research on conversational agents in education in the instructional dialogue design, based on Bloom's revised taxonomy of learning objectives [AK01]. We showed that this structure can also be used in other domains. Further, above we have explained already a few teaching (instructional perspective) and adaptation (technical perspective) strategies. A systematic guideline for educational conversational agent developers would be helpful that summarises which teaching strategies can be inserted in a single conversational agent lesson. There is overall still room for improvement in research on being able to accommodate more complex question types and feedbacking them in intelligent tutoring systems. Each part of the conversation can be more efficient and more realistic. For example, working on NLP capabilities allows us to analyse and provide better feedback on open-ended questions.

In the next chapter, we focus on assessing the agent's usability, which is an important baseline that educational technology needs to meet. We also elaborate on what qualities of the learning process and learning outcomes change as the interaction paradigm is more conversational when compared to other interactive digital content formats. Specifically, we analyse the differences in how user experience differs between learning via web-based interactive content and learning within a dialogue led by a full tutorial conversational agent when both transfer the same learning materials differently.

5. Comparing an Interactive Web-Based Learning Materials Vs. Tutorial Conversational Agents: Differences in User Experience

5.1. Introduction

Today's learning platforms make content available and enable social interaction between humans. Tomorrow, such platforms could also host computational tutors that support learning through dialogues. In this chapter, we continue investigating full-tutorial conversational agents by which learners experience a full learning journey. Following the previous chapter, we explore how user experience differs between learning via web-based interactive content (similar to a learning management system) and learning within a dialogue led by a full-tutorial conversational agent (similar to what was presented in the previous chapter).

Learning management systems (LMS) are the standard educational technology used in education. LMS typically offer functionalities to share multimedia content, quizzes with different question formats, and social features like message boards and discussion forums. LMS are web-based and via responsive design work across all kinds of devices. Due to their widespread use, they now also have the advantage of being very common and well-established.
Other educational technologies, such as educational conversational agents, are much newer and much less widespread; they have become possible through advances in natural language processing. Agents' responses could be generative or retrieval. In generative models, the responses are generated based on the user's inputs, for example, OpenAI's ChatGPT which uses (large-) language models. However, in retrieval models, the best response has been selected from a pre-defined list of responses (e.g. [Gra+99; Cai+19; Wol+22]). While educational conversational agents can take multiple roles, in this work, we are particularly interested in tutorial conversational agents. These typically (are programmed to) convey content and to ask questions to learners or answer questions of learners.

Clearly, these two types of technologies are very different. However, they can complement each other, such that we see tutorial conversational agents as a type of technology that could be added to current LMS. Subsequently, LMS could continue to share multimedia content and interactive quizzes, and host social activities between human users, but in addition, would provide access to computational tutors. In other words, LMS would continue their move from sharing content to hosting learning activities (now: by social features; future: in addition by providing access to computational tutors). In this work, we delve more deeply into the obvious, namely that these two types of technologies are very different, and explore based on a user study in which we compare learning via features typically found in an LMS with learning via features typically found in a tutorial agent in terms of usability and overall user experience (e.g. [BDA22; Leo+21]).

The rest of the chapter is organised as follows. The second section delves into the description of LMS, detailed in Section 5.2. Subsequently, Section 5.3 presents an exploration of two learning technologies, namely DIGIVIDget and DIGIBOT. Moving on, Section 5.4 outlines the research questions that are posed and addressed in this study. In Section 5.5, the methodology, participants, and the approach employed to address the research questions are described. The subsequent section, Section 5.6, provides a description of the results obtained in alignment with the research questions. This chapter is concluded by a comprehensive discussion presented in Section 5.7, followed by a conclusive summary in Section 5.8.

5.2. Learning Management Systems

Learning management systems (LMS) are an educational technology offering functionalities for creating, managing, and delivering course material [TCL21]. Such functionalities include sharing of (multimedia) learning content, personalised user experience, flexible analytics, a variety of quiz formats and adaptive assessments, and social features including message boards or discussion forums [SM23]. Popular LMS for universities are for example Moodle[1], ATutor[2], Blackboard[3], and OLAT[4]. LMS are used in various ways to support teaching and to teach a wide range of learners. Additionally, LMS are applicable in different instructional settings supporting face-to-face, fully remote and blended learning.
Typically, LMS are implemented as online learning technology using responsive design and are therefore working on all kinds of devices (e.g. smartphones, tablets). This leads to their widespread use in different learning contexts, and thus, LMS are well-known and well-established.

Depending on the environment in which the LMS is embedded, its usage requires more or less self-regulated learning capabilities. In traditional learning settings (face-to-face), instructors design the duration, the pace or the learning sequence of the course, thus, there is no self-regulated learning for students [Shi20]. In contrast, in higher education, where the support and guidance of an instructor are limited, the set of skills subsumed under the term self-regulated learning is essential [Cic+18]. Self-regulated learning is defined as “the strategies that students use to regulate their cognition (i.e., use of various cognitive and meta-cognitive strategies) as well as the use of resource management

[1] *moodle.com*
[2] *atutor.github.io*
[3] *blackboard.com/teaching-learning*
[4] *olat.org*

strategies that students use to control their learning" [Pin99]. Students are seen as active participants in the learning process, and they need to construct their own goals, criteria, or standards [Pin04]. Monitoring their own learning progress towards these goals, criteria, or standards allows students to assess whether the learning process should continue as is or if some type of change is necessary [Pin04].

The utilisation of LMS provides the possibility to easily navigate through online content. Learners also need to self-pace their learning in line with their self-regulated learning strategies, which is often a challenge. At the same time, such learning settings allow learners to be very flexible in their learning; thus they appreciate well-designed LMS and learning materials in this regard.

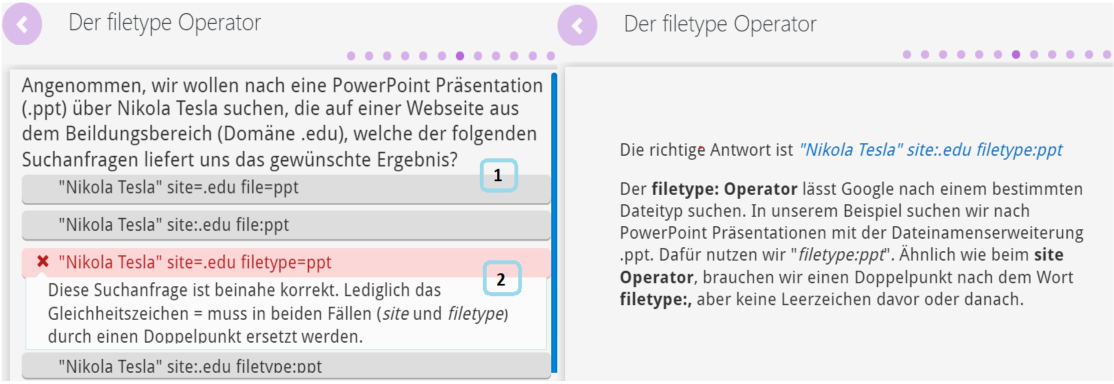
In relationship to LMS, educational conversational agents are typically more prescriptive in terms of how to navigate through the content via their dialogue structure (e.g., [Gro+19]). The typical characteristics of a tutorial agent include conveying content, giving individualised responses and feedback to learners' responses, and proactive guidance through a learning/teaching path. Interaction, as one of the main characteristics of this technology, increases learners' motivation compared to a more traditional learning system [Rua+19]. These two technologies were compared by Grossman et al. [Gro+19]. They created an educational conversational agent - called MathBot - which acted as a tutor with scripted dialogue and compared it with an LMS. Their goal was to compare MathBot, as an automated text-based tutor, with Khan Academy, as an LMS which used videos and written tutorials, in terms of user preferences and learning outcomes. As for user preferences, 53% of participants preferred Khan Academy with videos to MathBot (42%). However, compared to text-based Khan Academy, 47% preferred MathBot, which was more than 44% of Khan Academy which used only text. The difference was not significant in learning outcomes, although the MathBot performed slightly better.

5.3. DIGIVIDget and DIGIBOT

This study aims to evaluate and contrast the user experience related factors such as usability and user preferences between two educational technologies - DIGIVIDget as an educational technology similar to a typical LMS and DIGIBOT as an educational conversational agent which conveys the same text-based materials. Subsequently, we present the learning domain and the design of both educational technologies.

5.3.1. Learning domain

21st-century skills include a broad spectrum of digital competences and skills, including information and data literacy as suggested by the DigComp 2.2 [VRKP22] and the DigCompEdu [Red+17] frameworks. Being able to search for information on the Internet [Fes+19] or formulating appropriate search queries is one crucial digital competence. In this regard, our goal was to convey the skills of formulating concrete and meaningful search queries that can be applied in common search engines such as Google or Bing. Search engines usually offer a set of search operators in the form of words or symbols that allow users to find more refined and targeted results but are hardly used in practice.



(a) Slide 1 (b) Slide 2

Figure 5.1.: Covering the filetype operator by the DIGIVIDget

The search operators covered by both tools are quote (*“”*), or (*OR*), asterisk (***), range (*..*), site (*site:*), filetype (*filetype:*), intitle (*intitle:*) and exclusion (-).
In this work, we formulated the intended learning outcomes in the form of competence-oriented learning goals [Fes+21]. Both tools convey content related to the same learning goal, which is “You are able to search for information on the Internet in a targeted way.” but each tool conveyed it differently.

5.3.2. DIGIVIDget

The DIGIVIDget is a learning platform implemented in an adapted version of Moodle 3.11.3. The adaptations refer especially to the look and feel of the platform: i) the user interface does not resemble Moodle, ii) the content is structured along teaching units, each consisting of one or more competence-oriented learning goals [Fes+21], and iii) teaching units convey the learning content in the form of micro-learning cards implemented using H5P (*h5p.org*). Generally, the DIGIVIDget offers a set of teaching units on digital competences. One is the formulation of targeted search queries using search operators.

To convey the learning content about how to formulate targeted search queries, we have prepared the content as a set of 13 micro-learning cards. After a general introduction presented on the first micro-learning card, we introduce each search operator in the form of a presentation with integrated quizzes on two slides (see Figure 5.1): On the first slide (Figure 5.1a, point 1), the user receives a question related to the current introduced search operator, *filetype*. After answering the question, the user receives feedback on the correctness of the given answer (Figure 5.1a, point 2). In the second slide (see Figure 5.1b), the user is presented with the correct answer and an additional explanation about the operator.
The DIGIVIDget allows learners to freely navigate through the micro-learning cards and by going through all the micro-learning cards and learning the respective content, users can achieve the learning goal.

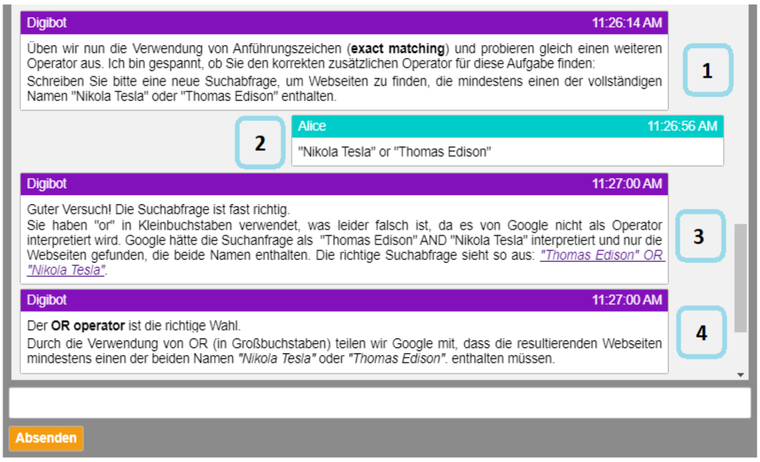


Figure 5.2.: Covering the OR operator by the DIGIBOT

5.3.3. DIGIBOT

The DIGIBOT is an educational conversational agent that acts as a tutor. It is designed to convey the same learning materials with the same wording and order as DIGIVIDget but in a dialogic form. The dialogue starts with a general introduction and greeting. Then, the agent talks about each operator one by one in the form of a dialogue, including asking questions (multiple choice or free-text questions) about operators, giving adaptive feedback, and an explanation about the operators. Figure 5.2, a screenshot of the DIGIBOT, shows how the agent covers the content related to the OR operator. In Figure 5.2, point 1, the user, called Alice, receives a free-text question related to the OR operator. The user can answer it by writing the answer in the textbox at the button. After having answered the question by the user (point 2), the agent gives adaptive feedback on the correctness of the given answer (point 3). The agent's responses differ based on the user's responses. Finally, the agent gives a general explanation about the search operator (point 4). The DIGIBOT covers the learning materials by asking questions and giving adaptive feedback in a dialogic manner. The whole dialogue's structure and the questions related to each operator are listed in Appendix A.7. The DIGIBOT was implemented upon the Bazaar framework [Ada+14].

The main difference between the DIGIVIDget and the DIGIBOT is how they convey the learning materials. The DIGIVIDget presents the content with micro-learning cards enabling the learners to freely navigate through them. However, the DIGIBOT conveys the same content through a conversation with learners.

5.4. Research Questions

Above, we have described both tools. They materialise different interaction metaphors in the sense that users (=learners) can interact in different manners with what should be learned through this interaction. The DIGIVIDget's metaphor is that of technology showing content to the learner, and the learner actively reading it and navigating through the content. The DIGIBOT's metaphor is that of an active entity (=the tutor) that engages the learner in a conversation around the learning content. As DIGIVIDget and DIGIBOT materialise different strategies for providing learners with the knowledge they need to reach (the same) learning goals, user experience is expected to differ.

In this work, we structure our exploration of how user experience differs into research questions that target different user-experience related constructs such as usability and user preference (RQs 1-3). Further, we aimed to explore the potential of each approach in inclusive educational contexts (RQ4), in line with our selection of study participants (see next Section 5.5 on methodology.)

- RQ1: *How was the DIGIVIDget perceived by the participants regarding its usability and usefulness?*
- RQ2: *How was the DIGIBOT perceived by the participants regarding its usability and usefulness?*
- RQ3: *What are users' preferences towards the DIGIVIDget and the DIGIBOT?*
- RQ4: *What is the potential of the different approaches for application in inclusive educational contexts from experts' point of view?*

5.5. Methodology

5.5.1. Participants

Study participants were recruited in a master's program on inclusive education at an Austrian university and they were asked to participate in the study as part of a mandatory course in the winter term 2022/2023. Study participation was not mandatory and in no way required to pass the course. 31 students out of 36 students in the course participated in the study. The majority of the participants (87.1%) were female. All of them had at least a Bachelor's degree in educational science or a related field (psychology, social work, pedagogy). At least 19 out of the 31 have practical experience as teaching assistants or personal assistants in working with people with disabilities and/or young people with special learning needs (for the others, this is not known).

5.5.2. Procedure and Instruments

We answered the above research questions with an explorative study carried out in January 2023. The study was divided into two phases: In the first phase, the 31 study participants were randomly assigned to two groups. One group used the DIGIVIDget ($n = 14$) and the other group used the DIGIBOT ($n = 17$). Immediately after their learning experience, the participants filled in a questionnaire comprising 10 items regarding the system (System Usability Scale (SUS), ten items) [Bro+96] and six items regarding their evaluation of the learning content.
The SUS consists of five items representing negative aspects (even-numbered items, for

example, “*I feel the system is unnecessarily complex.*”) and five items representing positive aspects (odd-numbered items, for example, “*I can very well imagine using the system on a regular basis.*”) of a system’s usability. The additional six questions were developed by the authors and concerned the participants’ evaluation of the delivery of the content, learning activities and learnability, for example, *“Even without or with little previous knowledge, the content can be well understood.”*. All questions were rated on a scale ranging from 1 (“I completely disagree”) to 5 (“I completely agree”). The internal consistency of the SUS in our sample (re-coded negative items) was good (Cronbach’s alpha of .89) and the internal consistency of the self-developed scale was sufficient (Cronbach’s alpha at .84). The quantitative data collected through the questionnaires forms the basis for answering RQs 1 and 2.
At the end of the questionnaire, we added one free-text question asking *“What did you like most about the DIGIVIDget/DIGIBOT and the learning activity? What did you not like so much?”*, where only either DIGIVIDget or DIGIBOT was presented. This question aimed to contribute to answering RQs 1, 2 and 3.

In the second phase, a smaller sample of six participants was drawn from each group. These twelve participants were also asked to try the technology they had not used in the first phase and participate in a focus group discussion one week later. All six focus group participants drawn from the DIGIBOT group also tried the DIGIVIDget and participated in the focus group discussion. From the DIGIVIDget group, one participant did not join the focus group discussion. The following questions guided the focus group discussion: (1) How did participants like each learning experience as a whole (trying both approaches one after the other), and what did they like in particular, what not so much? (2) What (if any) effect do participants see in the order of the two approaches when working through both? (3) What advantages and disadvantages do participants see (when comparing both approaches)? (4) Which approach did participants prefer and why? (5) What is the potential of both tools in the field of inclusive education in terms of competence and knowledge gain as well as learning engagement and motivation? (6) What are possible barriers to use?
Both focus group discussions took place online and lasted around 40 minutes; they were recorded, transcribed verbatim, and analysed applying qualitative content analysis [Kuc18]. Qualitative analysis of the discussions was used to give additional insight into RQs 1 and 2, and to answer RQs 3 and 4.

5.6. Results

5.6.1. Perceived Usability and usefulness (RQs 1 and 2)

To answer RQ1 and RQ2, we use the results of the questionnaires regarding the usability of the DIGIVIDget and the DIGIBOT. Table 5.1 shows the median and the interquartile range for each question in both groups.

We can summarise that the median scores of odd questions (positive aspects, higher median means higher perceived better usability) for the DIGIVIDget group are between $MD = 3$ ($IQR = 1$) and $MD = 4$ ($IQR = 1$), and for the DIGIBOT group, all median

Table 5.1.: The median (MD) and the interquartile range (IQR) for each SUS question and the overall mean (M) and standard deviation (SD) are shown.

#	Question	DIGIVIDget ($n = 14$) MD (IQR)	DIGIBOT ($n = 17$) MD (IQR)
1	I can very well imagine using the system on a regular basis.	3 (1)	4 (1)
2	I feel the system is unnecessarily complex.	2 (1)	1 (1)
3	I feel the system is easy to use.	4 (2)	4 (2)
4	I think I would need technical support to use the system.	2 (2)	2 (1.5)
5	I find that the various features of the system are well integrated.	4 (1)	4 (1)
6	I find that there are too many inconsistencies in the system.	2.5 (1)	2 (1)
7	I imagine most people will learn to master the system quickly.	4 (1)	4 (2)
8	I find the operation to be very cumbersome.	2 (2)	1 (1)
9	I felt very confident using the system.	4 (2)	4 (1.5)
10	I had to learn a lot of things before I could work with the system.	1.5 (2)	1 (0)
Overall		$M = 3.58$ ($SD = 0.77$)	$M = 4.12$ ($SD = 0.72$)

scores were at 4 (IQR between 1 and 2). For the even questions (negative aspects, higher median means lower perceived poorer usability), the median scores in the DIGIVIDget group are between $MD = 1.5$ ($IQR = 2$) and $M = 2$ ($IQR = 2$), and in the DIGIBOT group between $MD = 1$ ($IQR = 0$) and $MD = 2$ ($IQR = 1.5$).
To compute an overall mean score for the SUS, we re-coded the scores of even-numbered questions or negative aspect items (by subtracting the given scores from 5 and adding 1 to them). The results showed that the overall SUS score for the DIGIVIDget group was $M = 3.58$ ($SD = 0.77$) while the score for the DIGIBOT group was $M = 4.12$ ($SD = 0.72$), indicating slightly higher perceived better usability of the DIGIBOT. However, the difference between the two groups is statistically not significant based on the Mann-Whitney-U-Test ($U = 166.5, p = 0.06$).

Besides the questions regarding usability and usefulness, the participants were asked six questions related to the content and the learning activities. Table 5.2 shows the median and the interquartile range for each of the questions related to the content and the learning activities. Similar to the SUS, we calculated a mean score for the overall scale. The overall scores for the DIGIVIDget groups and the DIGIBOT groups are $M = 3.31$ ($SD = 0.89$) and $M = 3.66$ ($SD = 0.83$) respectively, indicating that both systems were perceived as appealing and supporting learning (as shown by the rather high medians). There is only a small, statistically not significant difference in this regard between the two approaches ($U = 146.5, p = 0.28$).

Finally, we report results on a free-text question to participants that asked what they liked most and what least of the tool they had used in their condition (which could be the DIGIVIDget or the DIGIBOT). 21 out of 31 participants answered this question (DIGIVIDget group: $n = 8$; DIGIBOT group: $n = 13$).

Table 5.2.: The median (MD) and the interquartile range (IQR) for each question related to the content and learning activities and the overall mean (M) and standard deviation (SD) are shown.

#	Question	DIGIVIDget ($n = 14$) MD (IQR)	DIGIBOT ($n = 17$) MD (IQR)
1	The contents are presented in an appealing way.	3.5 (2)	4 (1.5)
2	Even without or with very little previous knowledge, the content can be well understood.	3 (3)	4 (3)
3	The content is presented in a way that what has been learned can easily be applied to similar situations in practical use.	3.5 (1)	4 (2)
4	The learning activity is designed in varied ways.	3 (2)	3 (2)
5	The learning activity is designed in such a way that it motivates to stick with it and complete it in one piece.	3.5 (2)	3 (1)
6	I find that learning activities that are prepared like this one supports self-directed learning.	4 (2)	4 (2)
Overall		$M = 3.31$ ($SD = 0.89$)	$M = 3.66$ ($SD = 0.83$)

As for DIGIVIDget, the participants liked the navigability and being able to try different answers before receiving the correct answer. In addition, the most negative aspect, mentioned by three participants, was the lack of enough introductory explanation to the topic and how to proceed through the content. As for DIGIBOT, study participants liked the agent's responses, which were found to be short, concise and helpful. The most negatively mentioned aspects of the DIGIBOT were related to the user interface design. For instance, it was mentioned that *"Maybe the design could be made a little more lively and appealing"*.

5.6.2. Users' (Comparative) Preferences (RQ3)

Two focus group discussions were held as a follow-up after the participants used the tools. In focus group 1, the participants used the DIGIVIDget first and then the DIGIBOT, while in the second focus group, they used first the DIGIBOT and then the DIGIVIDget. Participants of the first focus group are referred to as P1 - P5, and participants of the second focus group as P6 - P11.

In the first question, the participants were asked to briefly summarise their impressions of both tools. Overall, the participants liked both tools, the DIGIVIDget and the DIGIBOT: Regarding the DIGIVIDget, P5 stated: *"I thought both contents were actually cool. I found that even with "this classic" [DIGIVIDget], there were already many different methods inside. So it was already quite varied designed."*. referring to the DIGIVIDget and the different types of presenting the content (e.g. quizes, slides) Besides the mentioned diversity of activities in the DIGIVIDget, users also liked that with the DIGIVIDget they can take as much time as they need to deepen and strengthen the knowledge *"If you have the classic learning platform [DIGIVIDget], you can take more*

time and adjust to it and repeat it" (P3). In contrast, the dialogic structure of the DIGIBOT potentially put some time pressure on them, as was mentioned in both focus groups. Another positive aspect of the DIGIVIDget mentioned in one of the focus group discussions was the clear structure of the content, while the DIGIBOT might become confusing *"if it is a large quantity [of information] that should be conveyed"* (P6). In both focus group discussions, the interactivity of the DIGIBOT was highlighted as a very positive aspect and, for example, P10 stated that learning with the DIGIBOT was a lot of fun *"So it was kind of very funny and at the same time you really learned something because it was so constructive and you had to write something into it yourself. And this was somehow a lot of fun."*. Some participants even indicated a clear preference for the DIGIBOT on an affective level, such as P11, who said *"I personally liked that one [DIGIBOT] better. I found this interactive element totally enjoyable"*or on the level of learning effect, such as P1 who stated *"the DIGIBOT alone would have been sufficient"*. In particular, direct feedback was mentioned by all participants and they agreed that this was a particularly motivating and encouraging aspect of the DIGIBOT; for example, P1 found the DIGIBOT very exciting as his learning situation changed completely and turned into *"a more motivated learning through the very positive and very appreciative feedback, so I felt like I was more motivated, but also somehow more eager to interact with the tool [DIGIBOT]"*.
Note that interacting with an educational conversational agent was a new experience for all focus group participants. This is illustrated by the following statement from P8 *"I found it [DIGIBOT] nice that it was something different and that there was simply some variety in it"*.

In the second question, we asked focus group participants about the order in which they had used both tools, and whether they had any preference regarding this. In focus group 1, the participants had used the DIGIVIDget first and the DIGIBOT afterwards. Three of the participants in this group explicitly stated that they liked the order in which they used the tools. P4, for example, said *"I actually think it's great. I noticed that at first with the classical system [DIGIVIDget] it was so that I could try it out and grasp, also in terms of content, what is it actually about and how does it actually work? And then through the bot [DIGIBOT] actually the idea to repeat once again for deepening"*.
In focus group 2, the participants had used the DIGIBOT first and the DIGIVIDget afterwards. Opinions regarding the order differed, such as P7: *"I find the classic [DIGIVIDget] good so that you learn with it at the beginning to get prior knowledge and then can start with the other [DIGIBOT]"*. Another participant from this group, P8, contradicts this: *"So I think that basically both will work, but I would also say the DIGIBOT first, [...] this interactive is, I think, good to get into. And later, when you have the slides, you can actually learn very well with it"*. P11 supports this but qualifies that there may have been a bias towards this order as it was the actual order in which tools were used for P11.

Regarding concentration, in both focus groups the predominant experience reported was that the participants were more focused and concentrated when learning with the DIGIBOT. For example, P5 stated *"I was more concentrated with the DIGIBOT because it was more interactive in my eyes and required more concentration"* and P6 said *"simply* *that one has the feeling [that] it reacts situational to me at the moment is conducive to attention"*. Only P10 disagreed with the prevailing opinion and said that *"I learn more concentrated when learning as typical with a text or slides [...] there I am more concentrated than with a chatbot, because I think this is somehow more relaxed"* - implicitly pointing to a potential (time) pressure perceived when working with the DIGIBOT.

5.6.3. DIGIVIDget and DIGIBOT in Inclusive Education (RQ4)

Finally, in both focus groups, we discussed the potential of both tools for inclusive education (RQ4). As study participants are doing their masters in inclusive education this was a reasonable discussion to lead. Almost all participants saw potential for tutorial agents in inclusive education, based on their comparative experience with both types of technology. For example, P5 stated that *"[...] the DIGIBOT has great potential for inclusion because in the context of learning difficulties, I would say, it is easier to use than the classic system [DIGIVIDget]"*. Other participants from both focus groups pointed out that children and young people (e.g. from disadvantaged backgrounds), who are hard to get excited about learning, are typical learner groups in inclusive education. They could benefit from DIGIBOT, or conversational agents in general, as it is a technology that is seen as *"very close to life worlds of the younger generation"*. P4 who works with children stated that *"Especially where there is a learning disability or especially reading difficulties, or comprehension difficulties, I thought that it would appeal well to exactly these girls, would motivate them"*. Also P11 pointed to the potential of tutorial agents to positively impact motivation: *"With children and young people, I can totally imagine that the motivation to learn increases insanely because they simply have fun with it [DIGIBOT]"*.

5.7. Discussion

We have developed and investigated two educational tools, DIGIVIDget and DIGIBOT, regarding usability and preferences by conducting a study with 31 master students in inclusive education. In this investigation, DIGIVIDget stands in for how traditional learning platforms that have developed from learning management systems (LMS), support learning by providing (interactive) learning materials in a web-based and freely navigable manner. The DIGIBOT stands in for tutorial agents as a new interaction paradigm in which content and interaction are integrated into a dialogue.

Regarding usability (RQs 1 and 2), the DIGIBOT achieved better System Usability Scale (SUS) results than the DIGIVIDget. However, the difference was not statistically significant. To this knowledge, we can contribute via the discussions led within the focus groups, which further compared the DIGIVIDget with the DIGIBOT. Our guiding research questions for the focus group discussions were to understand user preferences towards the two technologies (RQ3) and thoughts about applicability and ways of using them (potentially in different ways) in inclusive education.

Regarding the degree of concentration, the focus group discussions indicate that participants were more focused and concentrated when using the DIGIBOT. This is in line with

the findings of Cai et al. [Cai+19] in which the conversational agent kept the learners focused as they talked to and received feedback from the agent. In general, an interactive learning process needs and facilitates more concentration. Concretely, the DIGIBOT asks questions and immediately gives personalised feedback to users' responses. Contrasting this, the DIGIVIDget does not require as much interaction and makes skimming through content easier.
Opinions varied regarding the preferred order of usage. Discussions point to individual preferences with underlying agreement about main user experience characteristics of the two technology types: some participants referred to the DIGIVIDget as their preferred first tool to cover the whole content because they are allowed to repeat each part and can freely navigate through the whole content. These participants would use the DIGIBOT subsequently to practice the content through open questions in the DIGIBOT's dialogue structure. Other participants considered the DIGIBOT a better first tool to get into the topic as it clearly structures the content and facilitates, and would then use the DIGIVIDget to practice and focus on more challenging parts of the content.
Note that a suitable or preferred order of usage might also depend on the learners' domain knowledge or cognitive capability (relevant in inclusive education). In this regard, a tutorial agent (in our case: DIGIBOT) is more suitable for those having less domain knowledge and hence needing and benefitting more from guidance (cp. [KSC06]). On the other hand, the LMS-style way of supporting learning towards the set learning goals (in our case: the DIGIVIDget) allows free navigation. While this requires more self-regulated learning competence as it less strongly requires concentration and interaction, this allows learners to cover content at their own pace, and also to select content. This all points to more suitability for learners with an overall higher self-regulated learning competence or motivation, more domain knowledge, or more punctual interests (as is typical for instance for professionals as learners, cp. [LPS22]).

Finally, the focus group discussion also addressed the potential of the tools in inclusive education, especially the DIGIBOT mentioned by both focus groups. In line with the above discussions regarding user experience and a priori suitability of tutorial agents vs. interactive and freely navigable web-based content, discussions highlighted especially two aspects: Learners with learning difficulties could benefit from tutorial agents due to their characteristic of being more interactive hence both requiring and facilitating more concentration, and due to their characteristic of more strongly requiring to follow the predetermined structure than freely navigable web-based content. Further, also children and the younger generation in more general, as well as learners who have difficulty getting excited about learning are typical target groups in inclusive education. These target groups could benefit from tutorial agents because according to the experts participating in the focus group discussion, they are potentially more motivating. Some of this may be reduced as conversational agents become more usual (now, their novelty helps make them more motivating), but the dialogic nature of interaction with tutorial agents may be intrinsically more motivating for these target groups.

5.8. Conclusion

In this chapter, we investigated that the two types of explored technologies constitute different interaction metaphors, in the sense that they allow different ways of learning through interacting with technology and content made available through it. Today's learning platforms have already moved forward from platforms that primarily share content to platforms that host learning activities (now: by social features that connect humans). This move could be continued to provide access to computational tutors and thereby support a different type of learning activity. Educational technology research now needs to think of how to combine and use the different interaction metaphors, and for what kinds of user groups which is more suitable. This work is an (early) investigation of these questions and points to follow-up work that explores the same overarching question in different domains, over time as the novelty of conversational agents is reduced, and specifically with user groups that have been identified as being different in terms of preferable interaction metaphor.

6. Identifying Structural Wrongness in Arguments within a Conversation

The thesis follows two main goals: 1) presenting a full-tutorial conversational agent which leverages various teaching strategies, and 2) using argumentation as a technique to enhance the learning experience of learners. The second goal is addressed in the next three chapters. We investigate argumentation as a teaching technique in educational conversational agents. In Chapter 6, we train Machine Learning (ML) classifiers to identify arguments in a conversation about the topic of intelligence. We use ML classifiers in an argumentative dialogue in which the agent asks argumentative questions and provides adaptive feedback based on the learners' answers to the questions. The agent analyses the arguments based on Toulmin's model of argument [Tou03] and then provides feedback. following the current chapter, in Chapter 7, we investigate the impact of such classifiers and feedback on the structure of arguments written by learners on writing argumentative essays. Finally, in Chapter 8, we describe a systematic workflow to be able to create such argumentative dialogues and classifiers in various learning domains using pre-trained transformer-based large language models.

6.1. Introduction

In this Chapter, we focus on how to enable a tutorial conversational agent to identify different types of structural wrongness in an argument during a conversation with learners where they are asked a question about the intelligence of a specific entity. Specifically, the learners are asked to apply agreed-upon definitions of intelligence to a concrete (type of) entity, such as "*a cat*", "*a chair*", or "*a self-driving car*". The choice of topic – discussing in what sense an entity is intelligent – has been made on the background of understanding the development of AI literacy as important in a society pervaded by increasingly powerful technology that is based on data analytics and other methods from AI [LM20]. One puzzle piece in this is to understand what AI is (ibid); as a precursor and surrounding discussion, we see the question of what is meant by intelligence, and more specifically in what sense different entities can be understood as intelligent. In the ensuing discussion, the conversational agent has the role of a tutor who develops the student's argumentation into a reasonable and clear argument. Such an agent needs to assess the structure of the learners' arguments. In this work, for argument classification or identifying the components of arguments, we used Toulmin's model of argument [Tou03] (Section 2.2.1) and focused on the core components, namely claim, warrant and evidence.

In comparison to the previous works (see Section 2.2.4), in this study, we worked on an argumentative-educational conversational agent in which the agent gave feedback on missing core components. The conversational agent tried to support learners to write an answer which has all the core components instead of persuading users or giving (counter-) arguments based on similarity to continue the conversations (e.g. [TG17; CH20; LNN18; Rak+19]). Regarding the identification of the core components, we followed previous studies in which traditional ML algorithms such as Random Forests [Bre01] and SVM [Joa98] were utilised. In line with these studies, we used ML methods such as K-Nearest Neighbors, SVM, Decision Trees, Random Forest and Ada Boost. We explicitly did not use deep learning methods in this work, as we have too little data; and do not use transfer learning as no suitable models from sufficiently similar problems are available.

The rest of the chapter is organised as follows. In Section 6.2, we concretise the research questions that we ask and answer in this work. In Section 6.3, we describe the method used to answer the research questions, including data collection, annotation, inter-rater agreement, data pre-processing, feature selection, overarching model development, and evaluation process. In Section 6.4 we describe results in line with the research questions, and we conclude the chapter with a discussion in Section 6.5 and a conclusion in Section 6.6.

6.2. Research Questions

In pursuing our overall goal, to study how to enable a conversational agent to identify different types of structural wrongness in an argument, we here investigate the suitability of using Toulmin's model of argument within conversational agents to operationalise what a good structure of an argument is and subsequently to identify different structural wrongness. In the present chapter, we study a conversational agent with whom one can discuss a single question: *Is <an entity> intelligent, and in what sense?* In this domain of discussion, we ask and answer the following three research questions:

- RQ1 (overarching): Can Toulmin's model of argument be used to model different types of structural wrongness within conversational agents in the given domain?
- RQ2: How well can components of Toulmin's model of argument be identified in the given domain?
- RQ3: Can a conditional dialogue structure with conditions based on the existence of components from Toulmin's model of argument lead to coherent conversations in the given domain?

Our methodology is as follows:

- Develop classifiers that operationalise Toulmin's model of argument in order to provide evidence for RQ2 (how well can different elements of Toulmin's model of argument be identified) in this case (preparatory work: Sections 6.3.2 – 6.3.6; classifier development and evaluation in Section 6.4)
- To set up a conditional dialogue structure with conditions based on the existence of arguments following Toulmin's model of argument and show, by example, that

it can lead to a coherent conversation (existential proof by example; in answer to RQ3).

- To discuss the overall suitability of Toulmin's model of argument as a suitable basis for modelling different types of wrongness in conversational agents (RQ1) based on results on the collected dataset.

As discussed in the related work and background section (Chapter 2), by answering these research questions, we contribute to the existing scientific discourse around conversational agents in education and argumentation mining knowledge about how Toulmin's model of argument can be operationalised and how this operationalisation can be used within a conversational agent to allow a coherent conversation that helps users develop a – structurally – good argument. This is useful, and novel in complement to existing research and knowledge on conversational agents that use domain background knowledge to facilitate the acquisition of factual knowledge, to develop argumentation along the dimension of content or educational conversational agents that moderate discussions by injecting discussion prompts.

6.3. Methodology

Below we describe the data collection study (Section 6.3.1), the educational materials used in the data collection (Section 6.3.2), the data annotation process and labels used (Section 6.3.3), the achieved inter-rater agreement as a measure of the quality of annotations and subsequently datasets (Section 6.3.4), the data processing (Section 6.3.5) and finally the feature selection for the three classifiers that aim to identify the existence of a claim, a warrant, and evidence in a given user statement (Section 6.3.6).

6.3.1. Data Collection

To collect data, Amazon Mechanical Turk[1] (MTurk) was used. It is a system for crowdsourcing work and it has been used in many academic fields to support research. By using crowdsourcing methods, a large number of diverse arguments can be collected and the data are free from researchers' bias [Cha+19]).

The data were collected in three rounds. In each round, essentially the question "*Is <an entity> intelligent or not? Why?*" was asked to study participants. The materials prepared for all three rounds are described in Section 6.3.2 below.

To increase the chance of having data without spelling or grammatical errors and also meaningful errors, we defined some qualification requirements for participants. The participants, who wanted to answer the questions, were required to be master workers. It means they needed to have a minimum acceptance rate of 95% in order to qualify to answer the questions. This qualification requirement ensures the high quality of the results. Furthermore, an additional qualification requirement was considered which was having an academic degree equal to or higher than a US bachelor's degree. The reason

[1] https://www.mturk.com/

Datasets	Number of collected responses	Qualification requirement
Dataset 1	100	HIT Approval Rate (%) >= 95
Dataset 2	1026	HIT Approval Rate (%) >= 95 At least US Bachelor's Degree
Dataset 3	211	HIT Approval Rate (%) >= 95 At least US Bachelor's Degree

Table 6.1.: MTurk experiments for collecting data

behind that was to have better responses in terms of formality and without spelling or grammatical errors. The data have been collected in three rounds (see Table 6.1).

As it is shown in Table 6.1, in the first pilot study, 100 responses regarding the question, "*Is <an entity> intelligent or not? Why?*" have been collected and the only qualification requirement was having an approval rate of more than or equal to 95. In the second round, 1026 responses were collected. However, the second qualification was also added. In the last round, the same qualification requirements similar to the second round were used and 211 new responses have been collected to use as a test set. In the end, overall, 1337 records have been collected. The data that have been collected from the first two rounds, Datasets 1 and 2, are considered as validation data and training data, and the records of the last round, Dataset 3, are considered as test data.

6.3.2. Apparatus – Educational Scenario "*Is <an entity> intelligent or not? Why?*"

We prepared the following materials for data collection: five different definitions of intelligence with brief explanations for each definition, a list of eight entities, and defining some properties for a good response, and a few samples of good/bad responses.

The following five definitions were given and explained as follows to study participants: We will call an object intelligent if it *thinks humanly, acts humanly, thinks rationally, acts rationally; or if it is able to learn from experience in order to better reach its goals.* These definitions were chosen on the background of understanding intelligence as a foundational concept for arguing about capabilities as well as non-capabilities of AI. The first four are discussed as having an impact on the discussion around intelligence in relation to the development of AI and inspired different directions of AI research (cp. [RN02]). The fifth definition more closely mirrors the understanding of learning in psychology and learning sciences.

Every study participant was asked to decide and argue about the intelligence of one (type of) entity, which was chosen such that in each dataset, the following categories are similarly represented: Inanimate objects, plants, animals, AI-enabled technologies. These categories are ontologically different, general judgements about their intelligence are possible, and we can expect different types of argumentations per category. As a general judgement, inanimate objects can be considered to not be intelligent according to any definition, plants could be with some difficulty argued to be intelligent as a species

Category	Dataset 1 & 2		Dataset 3 (test data)	
	# responses	Average tokens	# responses	Average tokens
Animals	296	36.73	53	29.77
Plants	277	34.06	55	34.85
Inanimate objects	277	31.54	52	30.23
AI-enabled technologies	276	39.13	51	32.31

Table 6.2.: The descriptive statistics of different categories of entities in the datasets.

if evolutionary aspects are put to the forefront, and animals and AI-enabled technologies could, in general, be argued to be intelligent even though in a differentiated manner.

In Dataset 1, these categories were instantiated by: tables (inanimate objects), trees (plants), cats, fish (animals), and Google search engine (AI-enabled technologies). We collected 100 records for Dataset 1 which means 20 records for each entity.

For Datasets 2 and 3, we used two examples per category, and these were office chairs and the New York Statue of Liberty (inanimate objects), sunflowers, Venus flytraps (plants), snakes, monkeys (animals), self-driving cars, Google search engine (AI-enabled technologies). We collected 1000 records for Dataset 2 which were 125 records for each entity and 200 records for Dataset 3 which were 25 records for each entity. We collected more records for Datasets 2 and 3. The extra records were due to a few short answers because we asked others to answer them again.

While collecting the data, it was also explained that a good response should be argumentative, contain a claim, reasoning, and an explanation, have at least 10 words, and be checked again for correcting typos. Furthermore, examples of good and bad responses were also illustrated in the explanation. In Table 6.2, some statistics related to the collected data are shown. For Datasets 1, 2, and 3, we collected 20, 125, and 25 responses respectively for each entity. Since some of the responses were too short or irrelevant, we did not approve of them and then asked new participants to answer them again. That is the reason behind the small deviations in the number of responses for each category. However, we used all the responses, rejected and approved responses, in our models. Overall, 349, 332, 329, and 327 responses were collected related to animals, plants, inanimate objects, and AI-enabled technologies respectively.

6.3.3. Data Annotation

The whole annotating process was done by two annotators (the authors). The whole process had three steps. First, in a group session, we reached a conclusion about the definitions of each component and how to annotate them. Second, we randomly selected 100 records from Dataset 2 and annotated them separately in order to measure the agreement (The detail of measuring inter-rater agreement is mentioned in the next section). In the last step, the first author annotated the rest of the unannotated data. The data were annotated based on three core components of Toulmin's model of arguments: Claim, warrant, and evidence.

Three different annotation values were considered for the claim: “positive” which means the user claimed that the entity is intelligent; “negative” which refers to the opposite direction which means the user’s claim is that the entity is not intelligent, “unknown” refers to responses in which there is no specific claim or stance regarding the question.

For the warrant, two different values were considered, “with warrant” or “without warrant” which refers to the existence of a warrant in the response: “with warrant” is assigned to responses in which at least one of the definitions of intelligence is mentioned.

For evidence, a binary value was considered. The responses are annotated with “with evidence” if there are some parts in the responses in which users use their background knowledge or observation to justify their claims. Table 6.3 represents the collected data in terms of these labels. The collected and annotated data are freely accessible for other researchers[2]. We explicitly discarded at this stage an evaluation of how reasonable the evidence is; this is discussed further in Section 6.4.3.

Component	Claim			Warrant		Evidence	
Annotation	Positive	Negative	Unknown	With warrant	Without warrant	With evidence	Without evidence
Training data	477	594	55	691	435	835	291
Test data	102	99	10	111	100	159	52

Table 6.3.: The number of different labels for each component in training and test data.

6.3.4. Inter-rater Agreement

One of the reasons that make argumentation mining and its sub-tasks such a challenging task is having disagreements in annotating datasets. Most datasets that are available do not report inter-rater agreements [LT16]. In principle, the overall quality of the argument is something that humans cannot agree about sometimes because, based on [Wac+17], some parts of the argumentation quality are subjective, and overall quality is hard to measure. In [Wac+17], it was also shown that some dimensions of argument quality in practice were not correlated to any argument quality in theory or some practical dimensions could not be separated and matched to theoretical dimensions of argument quality.

In general, analysing arguments and annotating texts is controversial most of the time and it leads to having more challenges in tasks such as detecting claims, warrants, or evidence. To train and evaluate the performance of detecting the core components, high-quality annotated datasets are required. In this study, Cohen’s κ value is used for evaluating inter-rater agreements. In this method, the inter-rater agreements among the

[2] $https://www.frontiersin.org/articles/10.3389/frai.2021.645516/full$

labels and agreements occurring by chance are taken into account. The equation for κ is:

$$\kappa = \frac{Pr(a)-Pr(e)}{1-Pr(e)}$$

In this equation, $Pr(a)$ is the relative observed agreement among raters, and $Pr(e)$ is the hypothetical probability of chance agreement. Different thresholds are defined for the value of κ. In general, the range of κ is from zero to one and the higher amount means the higher agreement between the raters. If the raters are in complete agreement then $\kappa = 1$, if there is no agreement among the raters other than what would be expected by chance, $\kappa = 0$. Based on [LK77], the values below 0 as poor, between 0 and 0.20 as slight, between 0.21 and 0.4 as fair, between 0.41 and 0.6 as moderate, between 0.61 and 0.80 as substantial, and above 0.81 as almost perfect inter-rater reliability. In [ST08], the threshold of 0.5 was recommended for exploratory research. In NLP tasks, the agreement is considered significant when κ is greater than 0.6 [CV18]. The values of κ for the claim, warrant, and evidence components were 0.94, 0.92, and 0.65 respectively. The κ value for the claim and warrant is more than 0.9 which means there is almost perfect inter-rater reliability. The definitions of claim and warrant components are straightforward and the coders exactly know what they are looking for. In contrast, the evidence could be anything based on users' background knowledge or observations that are related to the users' claims. So, there is a chance that in some responses the coders have different opinions. Even though there are unlimited ways of providing evidence that supports the claim of whether and in what sense an entity is intelligent, there is substantial agreement between the two raters on the existence of evidence ($\kappa = 0.65$). In an analysis of disagreements, the disagreements mostly stemmed from different quality thresholds of raters on what would be acceptable to count as evidence or not. For instance, there were disagreements for these samples, "*No. The Statue of Liberty cannot think and has no mind or brain.*" or *"An office chair is not intelligent because neither it can do work on its own nor it can think and act.*".

6.3.5. Data Pre-processing

The pre-processing steps are the same for all the models we created for detecting claims, warrants, and evidence. The steps are as follows: i) converting all responses to lowercase form, ii) removing additional spaces in the beginning, ending, and middle of the responses, iii) replacing the various form of the entities' names with a specific token, "ENT", iv) tokenising and lemmatising the responses.

Replacing the entities' names is crucial for two reasons. First, by replacing the entities' names with "ENT", it will be possible to create only one claim detection model to cover all types of entities. Second, we wanted to ignore the impact of the entities on the model predictions because the names of entities will affect the models' outputs. For example, in 86% of responses in which we asked about the intelligence of monkeys the users' claims were positive. It means the model of detecting claims tends to assign a positive claim to the responses related to the monkey entity. This justification is also valid for other entities such as "*an office chair*". In 91% of responses related to office chairs, the claim was negative which means the users claimed that the entity is not intelligent. In the next subsection, the features used to create the models are presented.

6.3.6. Feature Selection

To create classifiers, user responses need to be converted to vectors in order to be used by machine learning classifiers. In this subsection, we report on features in the sense of how the vectors are created. Overall, we developed three classifiers, one for each core component of Toulmin's model of argument: claims, warrants, and evidence. For each classifier, different features were used; and we report for each classifier separately which features were used below. Some features were nonetheless shared for all classifiers (general features – namely TFIDF representation of the user response), and some features were component-specific, i.e. specific to the core component of Toulmin's model of argument.

Claim: We report on features that were used as input to the classifiers that aimed to detect the existence of a claim in user response. We aimed to differentiate between three classes, positive claims, negative claims, and unknown claims. We identified these classes with two groups of features, general features and component-specific features.

Term-frequency-inverse-document-frequency (TFIDF) was used in this work throughout as a general representation of user responses: As a full document set, Datasets 1 and 2 are used; and the dictionary vector contains bigrams and trigrams. The unigrams were ignored because they could not be informative and indicative. Words such as "is", "intelligent", "not" did not lead us to a correct prediction about claims. However, bigrams and trigrams, for instance, "is intelligent" and "is not intelligent" are required to predict users' claims. After pre-processing steps (see Section 6.3.5), only the 500 most frequent bigrams and trigrams for the whole Datasets 1 and 2 were used as TFIDF vectors. The underlying rationale was, to avoid high sparsity vectors.

In addition, we used general background knowledge as well as information from pilot studies to add features that needed to be considered both specific to the "claim" as one component in an argument that shall be classified; and that was specific to the particular question that has being asked (is an entity intelligent or not). Two regular expressions were used to indicate whether a response was started or ended by phrases or words such as "yes,", "no,", "it is intelligent", "it is not intelligent" or not. If one of these patterns can be found in a response, based on being positive or negative, a ternary value, -1, 0, 1, was added to the general feature vector of the response.

Warrant: In this study, we asked participants to use one of five definitions of intelligence as a link between their claim and their concrete evidence: "acting humanly", "acting rationally", "thinking rationally", "thinking humanly", "learning from experience to better reach its goals". We aimed to differentiate between two classes only: With a warrant in the sense of a reference to one of these definitions, and without a warrant. Note that in Toulmin's description, it is said that a warrant can also be implicit as an underlying understanding based on which a human is making an argument. Therefore, we looked for explicit warrants in labelling, and subsequently in classification.

Part of the feature vector for the warrant classifier was the same TFIDF representation of the user response (the length of vector: 500; terms, bigrams and trigrams are represented). Our goal was to identify the existence of warrants (or their absence) in the sense of detecting the usage of one of five pre-defined definitions of intelligence. Hence, regular expressions were used, as a component-specific feature, to detect the presence of the different forms of definitions in responses. The phrases that we looked for by regular expressions were indicative phrases such as "act humanly", "think rationally", "can learn", "learn from experience", or "reach better". Based on the existence of these patterns, a binary value, 0 and 1, was added to the general features.

Evidence: We aimed to differentiate between two classes, that some evidence is given, or that it is not. Since no pre-defined list of facts or observations was given in the present studies, the evidence part in our study was the most free and hence the most variable part of user responses. As general features, similar to other components, we used TFIDF vectors but with different parameters. As a full document set, Datasets 1 and 2 were used; and the dictionary vector contained unigrams and bigrams. In contrast to the claim and warrant that trigrams phrases can be indicative for identifying the existence of claims and warrants, for detecting evidence components, unigrams and bigrams, such as, "no brain", "inanimate object", "prey", or "making tools", can be discriminative. Furthermore, in contrast to the feature vectors for detecting claims and warrants, the 3000 (instead of 500) most frequent unigrams and bigrams for the whole dataset were used in the dictionary vector, and TFIDF was computed based on this reduced dictionary vector. The underlying rationale for length reduction was again to avoid high sparsity vectors; the length was still larger because we expected more reasonable variance in the evidence part of the given arguments. Besides the length of vectors and n-grams, to remove phrases related to the claims and the warrants, we ignored all phrases that occurred in more than 30% of all responses to have more relevant and meaningful phrases.

As component-specific features, we used two different evidence-specific feature sets, a list of evidence-specific keywords, and the length. The evidence-specific keywords where we assume that when one of them appears in the user response, there is a high likelihood that this keyword is part of the evidence for the argument, and hence that the statement should be in the "with evidence" class. The evidence component is the only argument component in which users need to talk about aspects specific to the entities and based on their experience and background knowledge. It means that users use their own keywords to justify their claims. Table 6.4 shows the 30 keywords that we identified in Dataset 2. In order to identify these keywords, we used Dataset 2 and did the following pre-processing: First, all phrases related to the claim and warrant (component-specific features) were eliminated. Second, we extracted unigrams and removed stop words. For each remaining unigram, a vector with the length of the number of responses was created that showed the existence of the unigram in each response. Then, Matthew's correlation coefficient between each vector of unigrams and the class values of the evidence class (with/without evidence) was calculated. The 500 most correlated unigrams were chosen; the cut-off was empirical because subsequent unigrams seemed too random. This yielded the bold entries in Table 6.4. The non-bold entries are the keywords that have been added to the list because they are synonyms, similar, or relevant to other high-correlated unigrams such as "by human", "handmade" and "made by" as synonyms or relevant to "man-made".

instinct	**plant**	**prey**	steel	**inanimate**	**sunlight**
hunt	**brain**	**object**	**trap**	handmade	**living**
survive	**lifeless**	**aware**	**insect**	**man-made**	**grow**
tool	alive	**cognition**	feed	made by	**food**
group	**metal**	**program**	by human	**stone**	**sun**

Table 6.4.: The features that are used for identifying the existence of evidence in responses.

These words and phrases are related to entities and study participants often used them in the evidence part of responses to argue why a particular definition of intelligence (warrant) applied to an entity or not. This feature set conceptually captures evidence that is made at an abstraction level that is higher than the single entity types in the sense of referring to an entity's decisive characteristic as being an inanimate object, or as having a brain; which in turn would be true about many more entities than the Statue of Liberty, or snakes. To use this feature, a binary value was added to the general vector of evidence to indicate the existence of this evidence-specific feature.

The second evidence-specific feature set was the length in terms of the number of words in the responses. Conceptually, if one wants to make a claim, refer to a predefined definition and in addition describe evidence that links the claim and warrant, one needs more words than if one does not add evidence. When cross-checking this intuition in Datasets 1 and 2, there is a significant difference between responses that contain evidence ($M = 39.16, SD = 23.27$, in Datasets 1 and 2) and without evidence ($M = 25.10, SD = 16.73$, in Datasets 1 and 2). To show this, Welch's t-test experiment was done and it showed that the difference was significant, $t = -11.07, p - value < 0.0001$ and the degree of freedom$=$ 701.63. To show that the significant difference was due to the length of the evidence component and not the other components, first, we reduced the length value by 4 and 5 words if responses had the claim and the warrant component respectively. Then, we did Welch's t-test again on the new values for the length feature for responses with evidence ($M = 32.27, SD = 22.9$) and without evidence ($M = 18.39, SD = 16.86$). Based on the new length value, there was a significant difference in terms of length, $t = -10.95, p - value < 0.0001$ and the degree of freedom$=$ 684.48. This feature intuitively makes sense, and yet of course is very coarse, in the sense that it can fail in single instances if no claim or warrant exists (shorter overall response which still includes evidence), can fail in single instances if claim and warrant are expressed very verbosely; and of course, absolutely fails to capture the correctness of the evidence or soundness of the overall argument in any way.

In Table 6.5, we summarized the features that were used for identifying the existence of the core components in responses.

Component	General feature	Component-specific feature
Claim	TFIDF of bigrams and trigrams (The length of vector= 500)	Regular expressions to indicate phrases such as "it is (not) intelligent"
Warrant	TFIDF of bigrams and trigrams (The length of vector= 500)	Regular expressions to indicate the proposed definitions of intelligence
Evidence	TFIDF of bigrams and trigrams (The length of vector= 3000)	The entity-specific keywords (Table 6.4), The length of responses based on the number of words

Table 6.5.: The features used for training classifiers of claim, warrant, and evidence components.

6.4. Results

6.4.1. How well can components of Toulmin's model of argument be identified in the given domain? (RQ2)

In this section, we answer RQ2 - How well can different elements of Toulmin's model of argument be identified? by developing classifiers for the three core components of Toulmin's model of argument, namely claims, warrants, and evidence based on Datasets 1 and 2 as training data and Dataset 3 as unseen test data for evaluating these classifiers (see especially Section 6.3.1 on Data Collection and the three different datasets). The classifiers were developed using vector representations of user statements using features as described above (Section 6.3.6). We use traditional ML methods such as K-Nearest Neighbors, SVM, Decision Trees, Random Forest and Ada Boost. We explicitly do not use deep learning methods in this work, as we have too little data; and do not use transfer learning as no suitable models from sufficiently similar problems are available.

For selecting the best classifier for each core component, we measured F1-score in 10-fold cross-validation on Dataset 2 for mentioned traditional ML methods. Furthermore, we used Dataset 1 as a held-out dataset to compare the ML models based on F1-score. After we had selected the best classifier, a final model was trained based on both Datasets 1 and 2. In order to avoid overfitting, the dataset for tuning hyperparameters is a little bit larger than that for initial model training and comparison and more diverse (Datasets 1 and 2 have been collected with slightly different materials – see Section 6.3.2; and the dataset for evaluation needed to be previously unseen, as is standard practice in ML literature. We note that Datasets 1 and 2 were lexically relatively similar, first because we had removed concrete entity names in pre-processing (replacement with ENT), and second because user arguments differ mainly across categories of entities (inanimate object, plant, animal, AI-enabled technology) and not so much between different entities (e.g., cat vs. snake).

Claim component

In this subsection, we describe the development and evaluation of a classifier for deciding the existence and the direction of a claim. In Table 6.6, you can see real responses, before applying pre-processing steps, related to the different values of a claim.

Table 6.6.: Real samples regarding the different values of the claim component. There are users' responses without any modification.

User's response	Claim
"Monkeys and humans are evolutionary speaking very close. Whilst it can't be said to think or act "humanly" (by definition only humans can do that), it can certainly think and act both intelligently and rationally, and most certainly learns from experiences. Therefore it is intelligent."	Positive
"I think that a self-driving car is intelligent. It learns from experiences and adapts and makes decisions based on what it has learned."	Positive
"I think a venus flytrap just wants to feed itself. That would be the goal it wants to reach."	Unknown
"the New York Statue of Liberty is made of copper and it exhibits positivity to the people around it and also the toes of this statue denotes the stableness to the world."	Unknown
"no I don't believe a self-driving car is intelligent I believe the people who wrote the code that make the car self-driving are intelligent. The car can only do what is it is programed to do."	Negative
"no"	Negative

We compared standard machine learning classifiers (K-Nearest Neighbors, SVM, decision tree, Random Forest, Ada Boost) using 10-fold cross-validation over Dataset 2 and evaluation of performance on Dataset 1 as a held-out dataset in order to identify the best classification model. To compare classifiers, we report the mean and standard deviation of macro-F1-scores over all training-and-test iterations. The result is shown in Table 6.7.

Classifiers	**The result of 10-fold cross-validation on Dataset 2**		**The result of using Dataset 1 as a held-out dataset**	
	Average of macro F1-scores	Standard deviation of macro F1-scores	Macro F1-score	Accuracy
K-Nearest Neighbors	0.61	0.02	0.56	0.70
SVM	0.76	0.07	0.63	0.93
Decision Tree	0.75	0.03	0.71	0.88
Random Forest	0.79	0.07	0.80	0.94
Ada Boost	0.68	0.07	0.62	0.92

Table 6.7.: The result of 10-fold cross-validation on Dataset 2 in detecting claims and evaluation of performance on the held-out dataset.

As it is shown in Table 6.7, the Random Forest classifier achieved the best results, and hence we proceeded to finetune this classifier. We observed that average F1-scores are reasonable for multiple classifiers (SVM, decision tree, Random Forest); this highlights the fundamental feasibility of separating statements with claim from those without a claim. The data that was used to train the final Random Forest classifier was the whole datasets 1 and 2 (1126 records). Since Dataset 2 was imbalanced and also it contained the majority of records, the whole training data became imbalanced. There were 594 records with "Negative" labels, 477 records with "Positive" labels, and only 55 records with "Unknown" labels. The data was extremely imbalanced since only 4.8 per cent of training data are annotated with "Unknown" labels. To tackle this, we generated synthetic examples via the Synthetic Minority Oversampling Technique (SMOTE), which generates new synthetic examples based on their selected nearest neighbours [Cha+02].

In order to select the tunning parameters of the Random Forest classifier, Grid-

SearchCV function of Scikit-learn [Ped+11] was used. The tunning function was parameterized for training the Random Forest classifier with 5-fold cross-validation. The parameters that we tried to optimize were the numbers of estimators (n_estimators) and maximum depth (max_depth). The rest of the parameters used default values[3]. For the number of estimators, the range of [100, 150, 200, 250, 300] and for the maximum depth, the range of [10, 20, 30, 40, 50] were considered to find the best combination. The optimum Random Forest in terms of macro F1-measure corresponds to 200 estimators and a maximum depth of 40. In the final step, the Random Forest classifier was trained on the oversampled Datasets 1 and 2. Table 6.4.1 illustrates the performance of the model assessed on the unseen test data, Dataset 3.

	Precision	Recall	F1-score	# of instances
Positive	**0.97**	0.91	**0.94**	102
Negative	0.95	**0.93**	**0.94**	99
Unknown	0.33	0.60	0.43	10

Table 6.8.: The performance of detecting claims on the test data (Dataset 3) based on each class.

As you can see, the results are shown based on each class based on precision, recall, and F1-score. All the scores for positive and negative classes are more than 90%. For precision, the positive class has the highest score; for recall, the negative class. In terms of F1-score, both categories achieve the same score. The unknown category was extremely imbalanced. Regarding the "Unknown" category which was extremely imbalanced, there were only 55 responses in training data, Datasets 1 and 2, and only 10 in Dataset 3. Overall, the performance on the positive and the negative class was very satisfactory. For the unknown class, it was not. The imbalanced precision and recall values mean that given an "unknown" label for a user statement, there is a reasonable likelihood that it will be wrongly classified as unknown (low precision). On the other hand, there is a very small likelihood of a user statement labelled as positive or negative to be anything else. In a separate experiment, a new model was created without using SMOTE, for the second time but as an up-sampling method. Without using SMOTE, all the scores of the minority class were zero.

Furthermore, since the claim detection model was a multi-class classifier, macro and weighted metrics are reported. Besides these scores, overall accuracy and Cohen's κ are reported. In Table 6.4.1, the macro and the weighted score of precision, recall, and F1-score and also accuracy and Cohen's κ are illustrated.

Macro average precision, which is the average of precision of all classes, is 0.75. However, the weighted average precision, which is the average precision based on the number of records for each class, is 0.93. We also measured macro and weighted averages for recall and F1-score metrics. The claim model had 0.91 accuracy which is satisfiable.

[3]The version of 0.23.2 of Scikit-learn was used in this study.

Random Forest classifier	Precision	Recall	F1-score	# of instances
Macro average	0.75	0.81	0.77	111
Weighted average	0.93	0.91	0.91	100
Accuracy	0.91			
Cohen's κ	0.83			

Table 6.9.: The overall performance of detecting claims on the test data (Dataset 3).

Warrant Component

In this subsection, we describe the development and evaluation of a classifier for deciding the existence of an explicit warrant in the sense of an explicit reference to one of five pre-defined different views on intelligence. In Table 6.10 several real responses from the study are shown.

Table 6.10.: Real samples regarding the different values of the warrant component.

User's response	Warrant
"Yes, I think that any action that involves the act of thinking and acting, involves a certain level of intelligence, in my opinion they are very intelligent, because they are born doing things that we humans are not born doing, they learn new things, things which is outside the animal world, things that only we humans learn, but of course there is a limitation in that."	With
"I think a monkey is very intelligent because it can learn just like a human."	With
"Snakes have the ability to adjust their behaviour as determined by their surroundings and, as such, are able to learn from their experiences, so, yes, they are intelligent."	With
"A self-driving car is intelligent as long as it has the correct information for it to function. It needs to have "brains" in order to work properly."	Without
"No, I think that the actins of reptiles which include apparent stealth and self direction, do not correspond to selecting from a set of alternative actions. The action is the only option and it is conjured by the needs of instinct"	Without
"It was intelligent it shows the friendship between two countries namely france and united States and mostly it representing liberty the enlightening the world. The torch really shows the path to freedom."	Without

Again, we compared give standard machine learning classifiers (K-Nearest Neighbors, SVM, decision tree, Random Forest, Ada Boost) similar to what we did for the claim component. The result is shown in Table 6.4.1.

Based on the results in Table 6.4.1, Random Forest classifiers were selected for detecting the existence of warrant in user's responses. Similar to the claim classifier, GridSearchCV function of Scikit-learn [Ped+11] was used to tune the parameters of the Random Forest classifier. The hyperparameters that we tried to find their optimum values were the numbers of estimators (n_estimators) and maximum depth (max_depth). For the first hyperparameter, the number of estimators, the range of [50, 100, 150, 200, 250] and for the maximum depth, the range of [10, 20, 30, 40, 50] were considered to find the best combination. The optimum Random Forest in terms of F1-measure corresponds to 100 estimators and a maximum depth of 30. Table 6.4.1 reports the performance of the model assessed on the unseen test data, Dataset 3.

Based on Table 6.4.1, the category of "with warrant" had the highest precision, however,

Classifiers	The result of 10-fold cross-validation on Dataset 2		The result of using Dataset 1 as a held-out dataset	
	Average of macro F1-scores	Standard deviation of macro F1-scores	Macro F1-score	Accuracy
K-Nearest Neighbors	0.76	0.03	0.55	0.58
SVM	0.85	0.03	0.61	0.61
Decision Tree	0.81	0.04	0.64	0.64
Random Forest	**0.87**	0.02	**0.68**	**0.68**
Ada Boost	0.85	0.03	0.65	0.65

Table 6.11.: The result of 10-fold cross-validation on Dataset 2 in detecting warrants and evaluation of performance on the held-out dataset.

Random Forest classifier	Precision	Recall	F1-score	# of instances
With warrant	0.95	0.83	0.88	111
Without warrant	0.83	0.95	0.88	100
Accuracy	0.89			
Cohen's κ	0.77			

Table 6.12.: The overall performance of detecting warrants on the test data (Dataset 3).

the best recall was related to "without warrant" category. The overall accuracy and Cohen's κ were 0.89 and 0.77 respectively. Besides the metrics, which are reported in Table 6.4.1, the average F1-score for the model was 0.88. These values are overall very reasonable. Especially, however, we note that for our use case, the lower precision and higher recall for "without warrant" means, additional fine-tuning might need to penalize further a wrong "without warrant" classification; in this case the conversational agent would mistakenly ask for an explicit warrant (a reference to one of the five definitions of intelligence in our study) even though the user had already given one. This should only be done given substantial evidence that such a question does more harm (=annoys users) than good (=helps users develop clear argumentative structures).

Evidence Component

In this subsection, we describe the development and evaluation of a classifier for deciding the existence of concrete evidence that illustrates the (non-)intelligence of an entity lasts. In Table 6.4.1, several responses are shown.

Similar to the warrant and claim section, we compared give standard machine learning classifiers (K-Nearest Neighbors, SVM, decision tree, Random Forest, Ada Boost). The result is shown in Table 6.4.1.

Similar to the claim and warrant classifiers, GridSearchCV was used for fine-tuning the model's parameters by training on Datasets 1 and 2 together. The tunning function was parameterized for training a Random Forest classifier in which 5-fold was selected for cross-validation with different numbers of estimators and maximum depth. The values that we considered for the number of estimators was [100, 200, 300, 400, 500] and for

User's response	Evidence
"In my opinion, a monkey is an intelligent being, as he presents aspects similar to those in humans, such as concern for the group, being able to perceive what is best for his community with its due limitations, motor intelligence, intelligence to solve situations that demand creativity."	With
"Actually, yes, I do. It doesn't "think humanely, or act humanely." I'm not sure if it thinks rationally or not, but it acts rationally: seeking out light in order to maximize its nutritional opportunities. It also, as all plants, learns from experience, in that it grows to match environmental conditions."	With
"I don't believe Google search engine meets the definition of intelligent because humans are behind the code of Google so Google itself is not doing the thinking. It is also only acting on what humans tell it to do. The only learning it might do is remembering what you've searched for previously and remembering cookies."	With
"Based on the definition provided the venus fly trap is not intelligent. I believe it meets some of the criteria (Thinks and acts rationally, learns from experience) but not all. It does not think or act humanly"	Without
"yes because it behaves humanly and can be able to adapt to changes to its environment"	Without
"A Table is unintelligent, because it cannot think like a human, move on its own or adapt behavior to a changing environment."	Without

Table 6.13.: Real samples regarding the different values of the evidence component.

the maximum depth was [40, 50, 60, 70]. The optimum Random Forest in terms of the average of F1-measure corresponds to 300 estimators and a maximum depth of 60. After finding the best parameters, a Random Forest classifier was trained on Dataset 1 and 2. The performance of the model was assessed on the unseen test Dataset 3 (Table 6.4.1).

Based on Table 6.4.1, the highest precision and recall were related to the category of "with evidence" The overall accuracy and Cohen's κ were 0.83 and 0.45 respectively. In addition to the metrics mentioned in Table 6.4.1, the average F-score for this model was 0.80. In our case, identifying the evidence component is the most challenging part in comparison to the other components, since the evidence part of an argument is based on users' experiences or observations. These precision and recall values are overall very reasonable. In comparison to the warrant classifier, in which users needed to mention warrants explicitly, there was no explicit answer for the evidence part. Thus, even if the evidence classifier wrongly predicts the category of "without_evidence", the whole conversation remains coherent because, in this case, the agent just asks the user to elaborate more on the response.

6.4.2. Can a conditional dialogue structure with conditions based on the existence of components from Toulmin's model of argument lead to coherent conversations in the given domain? (RQ3)

Above, we have ascertained that the existence of core components from Toulmin's model of argument can be detected reasonably well for the given dataset. In this section, we ask whether the availability of such classifiers enables us to create a conditional dialogue structure that can lead to coherent conversations (RQ3). We answer this research ques-

Classifiers	The result of 10-fold cross-validation on Dataset 2		The result of using Dataset 1 as a held-out dataset	
	Average of macro F1-scores	Standard deviation of macro F1-scores	Macro F1-score	Accuracy
K-Nearest Neighbors	0.87	0.02	0.70	0.77
SVM	0.86	0.01	0.44	0.70
Decision Tree	0.86	0.01	**72**	0.77
Random Forest	**0.90**	0.01	**0.72**	**81**
Ada Boost	0.88	0.02	0.63	0.74

Table 6.14.: The result of 10-fold cross-validation on Dataset 2 in detecting evidence and evaluation of performance on the held-out dataset.

Random Forest classifier	Precision	Recall	F1-score	# of instances
With evidence	0.83	0.96	0.89	159
Without evidence	0.79	0.42	0.54	52
Accuracy	0.83			
Cohen's κ	0.45			

Table 6.15.: The overall performance of detecting evidence on the test data (Dataset 3).

tion by example, in the following senses: First, we are just looking for a single reasonable dialogue structure. There surely are many reasonable dialogue structures, but we just need one. Also, in the current study, we are just interested in showing that the dialogue structure can lead to coherent conversations; not in showing how often this is the case in a given setting. In showing the quality of the conditional dialogue structure we, therefore, use the concept of conversation coherence as a quality indicator. We understand the linguistic concept of coherence, as denoting the extent to which a text (in this case: the conversation between the agent and the user) is meaningful and thoughts in it are well and logically connected. We, therefore, use conversational coherence as a fundamental quality that a tutorial conversation needs to have.

Coherence in the case of a retrieval-based conversational agent relies on the quality of i) the conditional dialogue structure and ii) the developed classifiers as well as the alignment of the two. The conditional dialogue structure needs to be well designed in that it is overall a reasonable path through a conversation, with an introduction, and a reasonable sequence of questions that suit the overall goal. The developed classifiers need to be able to decide between the conditional branches. The alignment between the two is necessary because depending on the quality of the classifiers, the responses of the agent need to show a different level of confidence towards the human user, in order to better perform in cases of the wrong classification.

In the below example conditional dialogue structure, the question about an entity's intelligence is shown as embedded in a longer tutorial interaction. The interaction follows the revised version [AK01] of Bloom's proposed taxonomy of educational goals [Blo+56]. In the revised version, the first four steps of the taxonomy are introducing knowledge, remembering, understanding and applying the knowledge. The focus of this study was on the applying step. Here we elaborated on each step.

- Introduction: The introduction is adapted to a use case setting in which the definitions of intelligence have already previously been discussed, e.g., in an introductory lecture on artificial intelligence.

- Introduction: The introduction is adapted to a use case setting in which the definitions of intelligence have already previously been discussed, e.g., in an introductory lecture on artificial intelligence.

- Remember: This part asks the user to repeat the learned definitions. Conditional branching with feedback can be designed for, but was outside our scope in this study.

- Understand: This part asks the user to explain in own words. Conditional branching with feedback can be designed for, but was outside our scope in this study. Additionally, it could be advisable even if problematic reasoning were detected here, to proceed immediately to the application stage, in order to switch between concrete and abstract reasoning; and only to come back to this level of understanding after a successful argumentation on a concrete example was carried out.

- Apply: This part is in focus of the present study, and the goal of the below dialogue structure is to show how the classifiers that decide upon the existence of core components of Toulmin's model of argument can be used to decide between branches in the dialogue structure. The dialogue flowchart is illustrated in Figure 6.1; in the subsequent explanation, the identifiers in brackets denote the decision points from Figure 6.1. The classifiers are executed sequentially. We first check for the existence and direction of a claim (C2), and act on identifying a missing claim; then we check for the existence of a warrant (W2), and act on identifying a missing warrant; finally, we check for the existence of evidence (E2), and act on identifying a missing evidence. Whenever a component (claim, warrant, evidence), is detected as missing, the agent uses increasing levels of scaffolding. The first scaffold is to point concretely to the missing core component (C3, W3, E3); the second scaffold is, to give the learner the start of an argumentative sentence or paragraph that just needs to be completed (C4, W4, E4). When the last scaffold fails, in the current dialogue structure, the conversation is (gracefully) ended, currently by apologizing for its own capability.

In Figures 6.2 and 4, two example conversations are given that showcase coherent conversations. In the conversation shown in Figure 6.2, the agent asks the user to argue whether and why a snake is intelligent. The user's response, "A snake is intelligent because it is able to survive, which indicates the ability to adapt to changing circumstances" (a response from Dataset 2), passes through all three classifiers (claim, warrant, evidence, see Section 6.4.3). Subsequently, the tutorial conversation is over. In the conversation shown in Figure 6.3, the agent asks the user to argue whether and why a sunflower is intelligent. the user's response, "No, the sunflower is not intelligent", only contains the

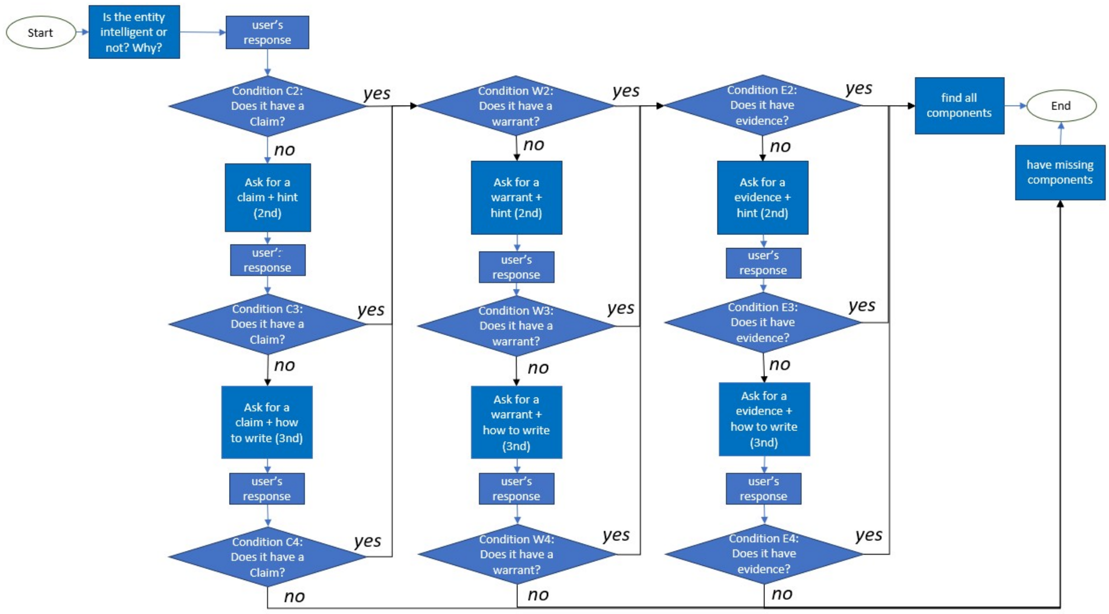


Figure 6.1.: The different states that the agent reaches based on the user's responses regarding the main question of the conversation, "Is <an entity> intelligent or not? Why?"

claim, but no warrant or evidence. In this case, the agent first shows its agreement and then asked for warrants to complete the argument. From this step onwards, it was us as authors who completed the remainder of the tutorial dialogue, just to show how a reasonable dialogue could ensue: The user's utterance, "acting humanly/rationally", fulfils the lack of the warrant component. Now, the only missing component is the evidence. In this step, the agent requests the user to add evidence or background knowledge to justify the claim. If the agent cannot identify the missing components, it will give the second chance to the user and ask again. As the agent again cannot find the evidence component, it asks the user to elaborate again. This is the last chance and if the agent cannot find the missing components again, the conversation will be ended. In Figure 6.3, the agent needed to find a connection between having no brain and thinking or acting to consider it as evidence part.

The agent's responses and follow-up questions are selected based on the predictions of classifiers. So, the conversations are coherent if the classifiers perform correctly. If they do not, the conversations can still be coherent as shown in the second example. This is created by the agent showing uncertainty when asking the user (=learner) to elaborate (C4, W4, E4). All agent's responses and branches are listed in Appendix A.3.

The above example conversations show that coherence can be achieved with the example conditional dialogue structure that makes use of classifiers that identify the existence of claims, warrants and evidence as core components of Toulmin's model of argument.

Note, that the example conditional dialogue structure does not show how incoherent

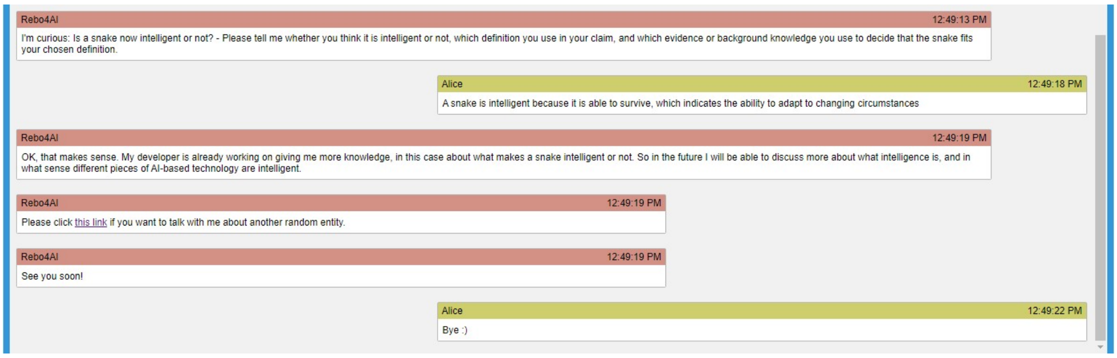


Figure 6.2.: A coherent conversation when all the core components were mentioned by the user.

Figure 6.3.: A coherent conversation when some of the core components were not mentioned by the user.

user responses can be caught and reacted to; and appropriate responses to wrong answers for the stages of remembering and understanding are not discussed either in this study.

6.4.3. Can Toulmin's model of argument be used to model different types of structural wrongness within conversational agents in the given domain? (RQ1)

In this subsection, we respond to the overarching research question, of whether and how Toulmin's model of argument is a suitable basis for modelling different types of the wrongness of arguments for use within conversational agents (RQ1). Answering this research question, will also immediately lead over to a broader discussion of our work in Section 6.5.

Firstly, we point out that using Toulmin's model of argument allows us to assess the structural characteristics of responses and in this sense structural quality. This means, that we can detect the existence of components of a reasonable argument; but we cannot – not by using Toulmin's model of arguments – say anything more about the content-wise plausibility.

For this purpose, we find that Toulmin's model of arguments works very well: With a comparatively small dataset, we were able to develop reasonably accurate classifiers (see Section 6.4.3) that are useful within a conditional dialogue structure to decide between branches (RQ2). We note that even though the developed classifiers model structural quality in the sense of the existence of necessary components, this assessment is related to content: The "warrant" classifier uses as features substantially content of the predefined definitions. The "evidence" classifiers use as features substantially content-related keywords that relate to how people argue about the intelligence or non-intelligence of entities. This highlights that in quality, structure and content are inter-related.

Despite this dependence of assessing structural quality on content-related features (quality indicators), we secondly observe that identifying the existence of Toulmin's core argumentative components does not per se allow us to assess content-wise plausibility of the made argument. For instance, in this response "yes because it acts rationally by providing humans comfort", in which it refers to an office chair, all the core components of Toulmin's model of argument were mentioned, but the content is arguable. However, we could use Toulmin's argument components to model different types of wrongness: For instance, it could be, that the evidence per se is not correct (a fictional example could be to say that "snakes are regularly observed to talk with each other in sign language"); it could also be, however, that the given evidence does not usefully relate to the used definition ("sunflowers move their heads with the direction of the sun, which shows that they learn from experience"). More generally speaking, each of Toulmin's model of arguments can have an independent value of "correctness" (whereby the value, in general, cannot be assumed to be binary), as well as interconnected values of content-wise quality in terms of how well, content-wise, the different parts of the argument align with each other.

Following this observation, we ask, how such content-wise quality assessment can be implemented? The answer to this can be found both in existing literature, and in future work: In the existing literature on argument mining, both the identification of similar arguments to one made by the user, and the identification of groups of arguments has been treated [Ada+14]. Argument similarity can be used when expert statements are available in the sense of a gold standard, and grouping arguments can be used when agreement with a majority opinion is a good marker of argument quality. On the other hand, fine-granular argument component detection and reasonable links between components that on their own might be correct or at least sufficiently reasonable to identify further problems is a topic for future research.

6.5. Discussion

In our endeavour to achieve our primary objective, which is to explore the capability of a conversational agent to recognize various forms of structural flaws, we are currently examining the applicability of Toulmin's argument model within conversational agents (RQ1). To address the primary goal, firstly, we developed and analysed the classifiers that identified the missing core components of Toulmin's model, namely claims, warrants and evidence, in a particular conversation around in what sense a given entity is regarded as intelligent or not (RQ2). Secondly, We have demonstrated by example that the conditional dialogue structures created based on the existence of Toulmin's core components led to coherent conversations (RQ3).

To answer RQ1, we first answered RQ2 and then RQ3. In RQ2, we developed and compared several machine learning classifiers to identify the missing core components in the user's answers to the agent's question about whether a specific entity is intelligent or not. Random Forest classifier achieved the best performance with the F1-scores of 0.77, 0.88 and 0.89 in identifying the existence of claims, warrants and evidence respectively. The results of classifying the core components were promising, although, the reported F1-scores were just valid and interpretable for the specific question asked by the agent. RQ2 contributes to the literature by focusing on the identification of Toulmin's components in the dialogue turns. Previous studies mainly addressed the identification of argumentative components in essays written by students which are lengthy and have a specific structure which can be utilised as a feature in the identification of the components (e.g. [Wam+21; Afr+21; SG17b; SG14]. In RQ3, we showed by examples that a conditional dialogue structure based on the existence of the core components led to coherent conversations which is a crucial quality of a tutorial conversational agent.
Achieving promising performance in the classification of Toulmin's core components in the given domain and utilising the classifiers to achieve coherent conversations demonstrate the suitability of Toulmin's model in modelling the different types of structural flaws within a conversation with a tutorial agent.

the current study also has several limitations: Firstly, in our data collection task, study participants received a specific explanation of what constitutes a good argument. Our concern was to have sufficient number of arguments that contain all components of Toulmin's model. Further, on the background of our research being on educational technology, it is reasonable to expect that users would receive some explanation for this. However, in settings, where no a priori explanation is given, it is to be expected that the distribution of classes (which components of Toulmin's model exists in a given user statement) is different than the distribution in our data set; and subsequently performance of the developed classifiers will vary.

Secondly, we have shown these results for a particular conversation. The classifiers use domain-specific (i.e., dataset-specific) features, like the limited-length TFIDF vectors, or the thirty terms most highly correlated with the "evidence" label (see Section 6.3.6). This means, for different conversation topics, still some feature re-engineering would need to occur. While our approach to feature engineering can be assumed to generalize, this is a) an assumption and b) still will result in different concrete features. Examples of a different conversation that is structurally similar to the one discussed in this study are an ethical dilemma. By definition, ethical dilemmas are situations in which, depending on underlying prioritization of different obligations, different courses of action would be reasonable. Such conversations could be conceptualized in Toulmin's model of argument as laying out as a claim which course of action one would choose (claim), laying out which obligation was most highly prioritized in choosing this course of action (warrant), and giving additional reasoning as to why the chosen priority is reasonable (evidence).

Thirdly, as discussed above in Section 6.4.3, while we do argue that Toulmin's model of arguments can also be used to structure identifying content-wise types of wrongness in arguments by means of argument mining, we have not shown this in the present work. Finally, we have not shown the effect of conversing with the agent on actual learning in an experimental study with human subjects.

These limitations also point out the direction of interesting future work, and stand in for research challenges that are being widely understood to be ambitious and are being addressed in educational technology and conversational agent research at large: Transferability of domain-specific classifiers; identifying more complex types of wrongness in arguments (i.e. argumentations where single components may make sense but do not fit together, as discussed towards the end of Section 6.4.3) and effectiveness of conversational agents as intelligent tutors in comparison to other teaching methods.

6.6. Conclusion

In this study, we put light on the suitability of Toulmin's model of argument in a conversational agent which asks an argumentative question about a specific topic (in this case, the intelligence of specific entities) and gives feedback on the structural flaws in the users' answers to the question. To ascertain the usability of Toulmin's model in a conversation (RQ1), we developed and analysed the classifiers which responsible to identify the missing Toulmin's core components in the users' responses (RQ2) and demonstrated by example the utilising the classifiers led to the coherent conversation (RQ3).

The contributions that this study makes towards state-of-the-art are a) to give evidence that reasonably accurate classifiers can be built for the existence of single components of Toulmin's model of arguments in (short) argumentative statements as would be expected in the context of a conversation with an intelligent agent in a given domain, b) to show by an argument that such classifiers are useful within dialogue structures of conversational agents that are designed based on Bloom's taxonomy for learning, and c) to show by argument how the same conceptual structure of Toulmin's model of argument can be used to further structure the identification of more complex types of faulty argumentation. These contributions complement existing research that has worked on longer argumentative essays [Wam+20c], that has differently conceptualized argumentation quality that is however less suitable for direct feedback within a conversational agent, and broader work on argumentation mining on identifying groups of similar arguments [WSA17] or conversational agents for factual teaching [Rua+19].

7. The Impact Argumentative Feedback by Agents on Writing Argumentative Essays

The previous chapter demonstrated the usability of Toulmin's model in identifying the core components, claims, warrants and evidence, within a conversation between a conversation agent and a learner. We also verified that providing feedback on structural gaps in learners' arguments by an agent leads to coherent and meaningful conversations. In this chapter, we show the impact of receiving such feedback on writing argumentative essays in two different domains.

7.1. Introduction

Being able to argue is an essential skill for participation in everyday and professional life [KW04], and for participation in society [DO02; Kuh93]. Based on this understanding, research has been carried out on different challenges regarding how to inject argumentation into teaching and also how to teach it. One strand of research focuses on including argumentation within teacher education [Erd06; EAYG06]. Such works investigate guidelines and teaching strategies for teachers in order to encourage students to justify better their opinions. A related line of investigation is interested in designing learning environments and course instructions and assessing their impact on the argumentation skills of students [Geo+20].

In this chapter, we take up a related and newer line of research that focuses on computer-mediated environments for utilising argumentation [Afr+21; WJL22]. A fundamental motivation underlying such research is the promise of scalability in the face of large class sizes while still being able to give feedback specifically to each learner. The importance of such personal feedback, adapted to each learner's prior knowledge or task performance is in turn a foundational motivation for research in adaptive learning support [Ale17]. Such feedback is of course also important for learning how to argue, as has been found in research investigating teaching strategies for argumentation skills [DV10]. This is the research strand that we continue and complement. Our particular approach to this challenge is to study a conversational agent for teaching argumentation directly within conversations. The underlying rationale is that good argumentation is expected and helpful in many private, professional and public conversations.

Adjacent to such research, our goal has been to develop an educational conversational agent that teaches a good argumentation structure within a conversation, rather than give feedback on an artefact developed outside the conversation (e.g., feedback on an argumentative essay written prior to the tutorial conversation). Naturally, argumentation, as given within a conversation, will be shorter and needs to be more compact than argumentation developed carefully over one or more pages in writing such as essays. Furthermore, we evaluate our intervention, the educational conversational agent, on two different argumentative domains in order to study whether the argument structure that the agent teaches can be transferred to a different domain of argumentation.

The rest of the chapter is organised as follows. In Section 7.2, we elaborate on the conversational agent developed for this study. Section 7.3 describes the research questions that we ask and answer in this work. In Section 7.4, we describe the method used to answer the research questions, including the procedure, the participants recruited for this study, data collection, annotation, inter-rater agreement and data analysis. In Section 7.5 we describe results in line with the research questions, and we conclude the chapter with a discussion in Section 7.6 and a conclusion in Section 7.7.

7.2. Conversational agent

The conversational agent that we have developed acts as a tutor towards the student with the goal to convey knowledge about good argumentation by developing a good argument in a concrete example together with the student. The knowledge about what constitutes a good argument is based on Toulmin's model of arguments [Tou03] (see Section 2.2.1). This model has already been widely used in research in educational settings, e.g., to assess students' opinions [Geo+20], to evaluate students' essays [HG17], and in online discussions to support the consolidation of opinions [Wan+20].

At the beginning of the conversation, the agent asks a question that demands an argumentative answer. If the user's argument is missing one of the three core components (claim, warrant, evidence), the agent explains which elements it understands the argument to already have and which to be missing, and asks the user to complement the answer by adding the missing elements. In the experiment described in this work, the agent asks questions of the type "*Is X intelligent or not? Why?*" X was replaced either by an animal (snakes, monkeys and sharks) or with an AI-enabled technology (Google search engine and self-driving cars). An example question asked to study participants therefore was "*Is a self-driving car intelligent or not? Why?*"
To identify which of the three core components are given in a user response, we used classifiers that were created and evaluated specifically for this type of question and described in the previous chapter (Chapter 6). The resulting classifiers achieve an accuracy of 0.91 for detecting whether a claim exists in a given argument, 0.89 for detecting whether the warrant exists in a given argument, and 0.83 for detecting whether evidence is given in the argument. By utilising the classifiers, the agent tried to teach the learners what a structurally complete argument should contain based on Toulmin's model of argument.

Figure 7.1 shows one of the dialogues within our experiment. The introduction, greeting, and ending parts of the dialogue were removed. Figure 7.2 shows the complete flowchart of the agent. Note that the agent asks three times for each missing core component and gives repeatedly stronger and more direct hints to help users to fill the

structural gap in their answers. If more than one component is missing, the agent prioritises the missing elements along the following sequence: Claim, warrant, and evidence. This means, that if all three components are missing, the agent will step-by-step scaffold the creation of a structurally complete argument.
Technically, we have implemented the agent using Bazaar framework [Ada+14][1].

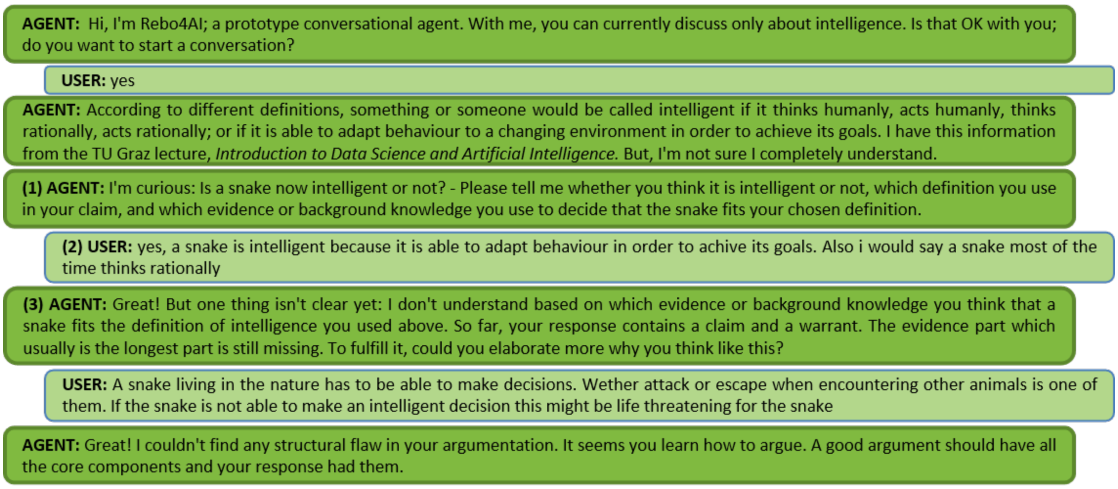


Figure 7.1.: A coherent conversation is taken from our study. (1): The agent asks its initial question. (2): the user answer is missing an element in the argumentation. (3): the agent identified the missing component and asked for it. Finally, the user completed the argumentation by adding evidence.

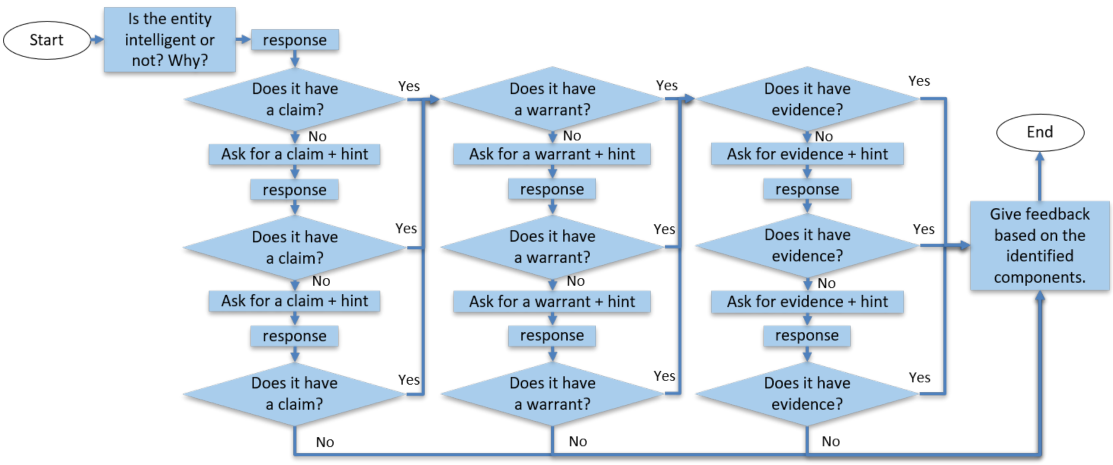


Figure 7.2.: The different states that the agent reaches based on user responses.

[1]https://github.com/DANCEcollaborative/

7.3. Research Questions

In this chapter, we ask and answer the following research questions, all with respect to the given experiment:

RQ1 - Classifier accuracy: How accurately do the used classifiers detect the existence of a claim, warrant and evidence respectively?

RQ2 - Conversation coherence: How coherent are the tutorial conversations?

RQ3 - Learning: Do users learn to structure arguments using the three core components of Toulmin's model of argument (claim, warrant, evidence) in the same argumentation domain as in the conversation with the agent?

RQ4 - Learning transfer: Can users apply the learned argument structure to a different domain of argumentation than is discussed in the conversation with the agent?

The questions are hierarchical in the sense that with each research question, we are aiming to ascertain a specific quality of the agent and the user-agent interaction, and each quality is the foundation for the quality that we are assessing in the next question. For instance, if the classifiers that we used in the investigated agent are not reasonably accurate in the here presented experiment (RQ1), then it is highly unlikely that tutorial conversations are coherent (RQ2). If conversations overall were incoherent, then we would expect that this has a negative impact on whether users actually learn to argue well with the given agent (RQs 3 and 4). On the other hand, the different qualities do not constitute strictly necessary preconditions. For instance, we have designed the agent's dialogue structure in a way, that some classifier inaccuracy is covered by the way the agent phrases its responses.
The contribution of this study towards literature lies in answering RQs 3 and 4 on whether a tutorial conversation leads to the learning of the taught argumentation structure (Toulmin's model). This complements existing work on computational environments for teaching argumentation in longer texts (e.g., [Afr+21; Wam+21]). Both the underlying computational methods needed to understand and feedback arguments are different, due to different lengths and styles of argumentation in essays and in conversations.

7.4. Methodology

7.4.1. Procedure

In order to answer the above research questions, we conducted a between-subjects experiment with two groups (treatment and control group). The experiment was a voluntary assignment set in a university lecture, named "Introduction to data science and Artificial Intelligence", at the Technical University of Graz. Before conducting the experiment, all materials were piloted within our research team.

The experiment contains three tasks. The overall procedure is shown in Figure 7.3. At the beginning of the experiment, all participants receive information about the tasks

and all necessary learning materials that explain Toulmin's model of argument and that give foundational information from the two different domains of argumentation used in the experiment tasks. These two domains of argumentation are: What is intelligence? Related learning materials describe different definitions of intelligence. We call this domain of argumentation the "intelligence domain"; it is used in Tasks 1 and 2. The second domain of argumentation is ethics, and related learning materials are about utilitarian and deontological ethics. We call this domain of argumentation the "ethics domain"; it is used in Task 3. The information about the tasks as well as learning materials remained accessible to study participants throughout the experiment.
The treatment group first exercises using Toulmin's argument structure in the intelligence domain (Task 1) with the agent. Then, the treatment group is given Task 2, which is also in the intelligence domain. Here the users simply answer the argumentative question without receiving feedback. Finally, the treatment group answers an argumentative question in the ethical domain (Task 3).
The control group starts with Tasks 2 and then Task 3. Note that the control group finishes with Task 1, which contains the intervention. This task was included for the control group as the experiment was an optional assignment in a university course, so that all students who decided to participate in the experiment would have the opportunity to talk to the conversational agent, and such that all students would have three argumentative tasks. The control group's performance on Task 1 was not used for the purpose of this study.

The data collected from Task 1 done by the treatment group answers RQs 1 and 2. The comparison of the performance on Task 2 between the treatment and control group answers RQ3 on whether the conversation with our tutorial agent helps learners to learn Toulmin's argument structure in the same domain of argumentation. The comparison of the performance on Task 3 between the treatment and the control group answers RQ4 on whether learners can transfer the learned argument structure to a different domain of argumentation.

7.4.2. Materials: Argumentation topics and tasks

Tasks 1 and 2 are about intelligence. Similar to the study mentioned in Chapter 6, in these tasks, the participants were asked to answer the question: "*Is X intelligent? Why?*". X was substituted either with a type of animal or by a (type of) AI-enabled technology. For animal categories, we used: sharks, eagles, monkeys, and snakes. For the AI-enabled technologies, we used: the Google search engine and self-driving cars. The two categories "animal" and "AI-enabled technology" are ontologically different, such that we can expect different lines of argumentation regarding their intelligence. In both categories, one can argue for both intelligence and non-intelligence of the entities, depending on the underlying definition of intelligence.
In Task 3, the participants were confronted with the trolley problem. By this we mean, the participants need to answer this: *There is a trolley coming down the tracks and ahead, there are five people tied to the tracks and are unable to move. The trolley will continue coming and will kill the five people. There is nothing you can do to rescue the five people*

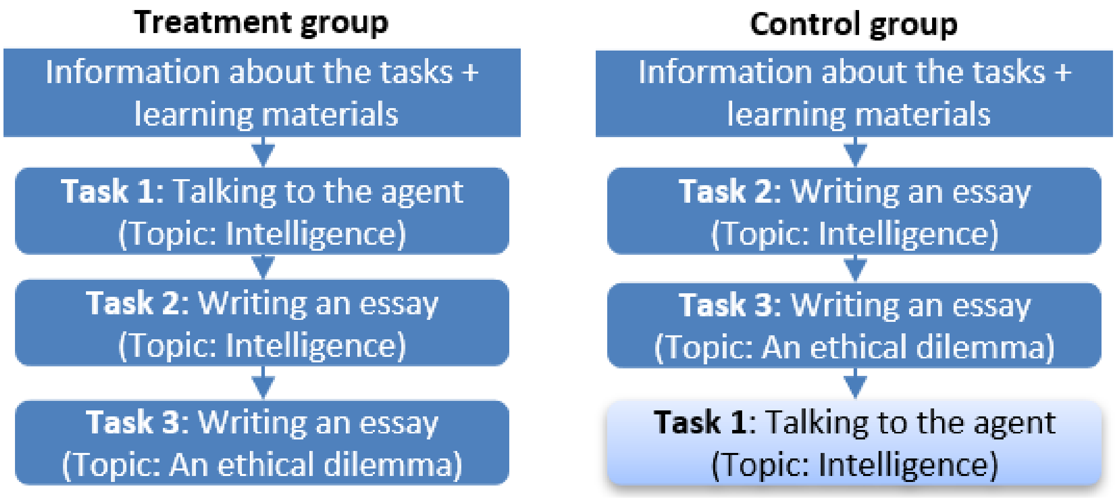


Figure 7.3.: The between-subjects study design.

EXCEPT that there is a lever. If you pull the lever, the train will be directed to another track, which has ONE person tied to it. What do you do? Please, justify your decision.

The choice of argumentation topics has been made such that the questions are structurally similar: By relying on different definitions, in one case of intelligence, in the other about what constitutes ethical behaviour, one can argue always for both types of answers (intelligent yes/no; pull the lever yes/no).

7.4.3. Participants

Study participants were recruited in an introductory university lecture, named "Introduction to Data Science and Artificial Intelligence", at the Technical University of Graz. The experiment was contextualised in the lecture as an optional bonus task for which points were received based on the most complete argument given in any of the three tasks. Overall 95 students participated in the experiment fully, i.e. such that they completed all three tasks in the treatment group or at least Tasks 2 and 3 in the control group. We randomly and equally split all the students into two groups. However, because of technical problems, we had to exclude some participants who could not finish all the tasks. Finally, 42 participants were in the treatment group and 53 in the control group.

7.4.4. Data Annotation

In all tasks, we annotated the user statements sentence by sentence. By statements, we mean: Every user response of the treatment group to an argumentative question of the conversational agent in Task 1 (depending on into which branches of the adaptive dialogue structure the user enters this could be more than once in a single conversation) and responses to the argumentative question given in Tasks 2 and 3. All the sentences have been annotated by four coders (including the first author), such that for each of them, we express which of the three core components of Toulmin's model (claim, warrant, evidence) is contained. Therefore, for each component, a binary value was defined to indicate the existence of the component. Then we called a sentence argumentative if it contains at least one of the core components. Furthermore, for Task 1, we annotated the conversations of the treatment group based on coherence. Similar to the core components, a binary value was defined to show whether a conversation is coherent or not.

Before starting the annotating process, a codebook in which the details about identifying the core components of Toulmin's model were explained. The annotation process was done based on the codebook in three rounds. In the first round, we randomly selected 25% of data that was independently annotated by all raters based on the codebook. The goal of this round was to reach a shared agreement on codes, and clarify and update the codebook where necessary.
In the second round, we randomly selected another 25% and assigned them to all four annotators. We computed inter-rater reliability and discussed the remaining differences in order to further improve the codebook and shared agreement.
For the 50% of data, annotated in the first two rounds, we had four ratings for each statement. Where we could not reach an agreement, we selected the final annotation based on the majority.
In the third round, we divided the rest of the data among all annotators for coding based on the codebook.

Overall, the data as analysed for the purpose of this study therefore contains 51 single responses written by the treatment group in Task 1, 904 sentences written by both groups in Task 2, and 600 sentences written by both groups in Task 3. The dataset is published on Zenodo[2]. Doccano [Nak+18] was used for annotating the data.

7.4.5. Inter-rater Agreement

We used Fleiss kappa [Fle71] to measure agreement among the coders. Based on [LK77], the values below 0 are understood as a poor agreement, between 0 and 0.20 as slight, between 0.21 and 0.4 as fair, between 0.41 and 0.6 as moderate, between 0.61 and 0.80 as substantial, and above 0.80 as almost perfect inter-rater reliability. However, we also note that not all the natural language processing tasks are not the same and we cannot define a specific threshold for all computational linguistic tasks [AP08].

In Tasks 1 and 2 which were about intelligence, the κ value for argumentative sentences was 0.69. We call a sentence argumentative if it contains at least one of the core components. The κ value for the claim, warrant and evidence were 0.90, 0.80 and 0.77 respectively. In Task 3, which was about an ethical dilemma, we achieved moderate inter-rater reliability for argumentative sentences, $\kappa = 0.58$. There was a substantial agreement for the claim (κ=0.77). However, the warrant (κ=0.36) and evidence (κ=0.42) components turned out to be not easily distinguishable [Erd07; VGK19]. We, therefore, did not

[2] *Zenodo*

use warrant and evidence annotations for comparing the groups on Task 3. Additionally, in Task 1, with respect to the coherence coding, we reached a significant agreement (κ=0.83).

7.4.6. Data Analysis

RQ1 (classifier accuracy) and RQ2 (coherence of tutorial conversations) in principle assess the qualities of the conversational agent. For assessing classifier performance (RQ1) in Task 1, we used the classifiers trained and developed in the previous chapter (Chapter 6). For each core component, a separate classifier was trained. We report the macro-average F1 score for each classifier over the 51 argumentative statements from Task 1 of the treatment group. The macro-average F1 score indicates the average of F1 scores of all classes. For assessing the coherence of tutorial conversations (RQ2), we report the number of coherent conversations in Task 1 of the treatment group expressed as an absolute number and as the ratio of all 42 conversations. (Section 7.4.4).

RQs 3 and 4 ask whether participants in the treatment group have learned Toulmin's argument structure. As the conversational agent in Task 1 specifically points out missing components from Toulmin's model, we hypothesised that students would learn to include all elements in the subsequent Tasks 2 and 3. In other words, we expect the treatment group to significantly more often use claims, warrants or evidence in their responses to questions in Tasks 2 and 3.

In order to measure this, we define that a sentence is argumentative if it contains at least one of the core components (a claim, warrant or evidence). We further define the ratio of argumentative sentences in the response to the question in either Task 2 or 3 to be the number of argumentative sentences in the user response divided by the overall number of sentences in a response. Using these definitions, we compare the treatment and the control group with a t-test on the ratio of argumentative sentences in Tasks 2 and 3. Additionally, to get more insights into the results, we do the same analysis for each Toulmin's core component separately. This means that we compare the treatment and the control group with a t-test on the ratio of sentences that contain a claim, warrant or evidence respectively. For RQ 4, which means to compare responses on Task 3, we only do this analysis for the claim component, since we did not reach sufficient inter-rater reliability for the warrant and evidence component (see Section 7.4.5).

7.5. Results

Concerning *RQ1 (classifier accuracy)*, we assess the classifier performance by computing the macro-average F1 score based on data from Task 1. The classifiers demonstrate promising performance in detecting the presence of claims, warrants, and evidence with macro-average F1 scores of 0.72, 0.91, and 0.77, respectively. These scores are generally considered indicative of excellent classifier performance. Moreover, they are comparable to the classifiers' performance on a dataset collected outside the conversational agent environment in the previous chapter, where the respective values were 0.77, 0.88, and 0.71 for claims, warrants, and evidence. Any misclassifications are addressed through the conversational agent's response formulation, ensuring a robust approach to handling classification errors.

Concerning *RQ2 (conversation coherence)*, our findings reveal that 85% (36 out of 42) of the dialogues exhibit coherence. Although this proportion is slightly lower, it remains comparable to the coherence rate observed in related work on a reflective conversational agent [WPSR20], where 97% (149 out of 153) of conversations were coherent. While there is a small difference in coherence percentages, the results indicate that our conversational agent's performance in maintaining coherence aligns well with existing research in the field.

Regarding *RQ3 (learning the argument structure, measured by a task within the same argumentation domain)*, we find that the treatment group ($N = 42, M = 0.83, SD = 0.18$) compared to the control group ($N = 53, M = 0.76, SD = 0.21$) has a significantly higher percentage of argumentative sentences in Task 2, $t(94) = 1.73, p = 0.042$ (see Table 7.5). To have more insight into this result, we compared the two groups in a more fine-granular manner per core component (claim, warrant, evidence). The only significant result is related to the claim component. The treatment group ($M = 0.26, SD = 0.25$) compared to the control group ($M = 0.16, SD = 0.1$) has a significantly higher percentage of sentences with a claim in their responses, $t(94) = 2.45, p = 0.007$. Overall, the results show that the treatment group has learned Toulmin's argument structure, but the effect is mainly visible on one of three argument elements, namely the claim.

Regarding *RQ4 (learning transfer, measured by a task within the same argumentation domain)*, we again compared the responses of both groups based on the proportion of argumentative sentences. The treatment group's responses ($M = 0.68, SD = 0.25$) compared to the control group's responses ($M = 0.59, SD = 0.27$) has a significantly higher ratio of argumentative sentences, $t(94) = 1.7, p = 0.045$. Because of the low inter-rater reliability on Task 3 for the components warrant and evidence, we only did the more fine-granular analysis for the component "claim". Different than for RQ3, however, the treatment group's responses do not contain a higher ratio of sentences with a claim than the control group's responses. Overall, these results show that the participants of the treatment group have transferred Toulmin's argument structure also to a new argumentation domain. Our results cannot be used to make a more fine-granular statement that distinguishes between the effects in terms of single core components.

Further, we see that in both Task 2 and Task 3, the control group wrote more sentences. Proportionally more of them were non-argumentative, however, i.e. not actually needed for the argument. In other words, the control group had more difficulties coming to the point and justifying it. Table 7.5 summarises the results for both RQ3 and RQ4.

7.6. Discussion

In summary, participants in the treatment group, who had an adaptive-tutoring conversation with the conversational agent (Task 1), wrote significantly more argumentative sentences, both in the same domain of argumentation (Task 2) in which the teaching took place, and in a different domain of argumentation (Task 3). This is encouraging,

Task 2					
Group	**# of sentences**	**% of arg. sentences**	**% of sent. with a claim**	**% of sent. with a warrant**	**% of sent. with evidence**
Treatment	358	**0.762**	**0.159**	**0.315**	**0.497**
Control	546	0.697	0.130	0.238	0.468
Task 3					
Treatment	252	**0.662**	**0.297**	-	-
Control	348	0.494	0.221	-	-

Table 7.1.: The ratio of argumentative sentences to all sentences based on each core component in Tasks 2 and 3.

as it shows the feasibility of using scalable tools such as a conversational agent to teach the basics of argumentation, as would be useful based on the understanding of the importance of argumentation in many areas of life. Our study further shows that it is possible to conduct a coherent conversation with a conversational tutorial agent. This is an important indicator of user experience and hence teaching quality.

We make two additional observations that highlight that future research is needed: For Task 2, at a fine-granular level, the treatment group and control group differed statistically significantly only in terms of the claim component. We interpreted this in the sense that the claim, in this task, was the easiest argumentative element. However, in Task 3, the difference between the proportion of sentences that contained claims was not significant. This could be due to the nature of the question, which was about an ethical dilemma situation in which at least one person would die and making one final decision is challenging. Further, we have noticed low inter-rater reliability in differentiating between warrant and evidence in Task 3. Overall, this leads us to suspect that the ethical dilemma is - in terms of Toulmin's structure of argument - substantially different from the discussion of which type of thing to consider intelligent and which not. In order to be able to effectively use Toulmin's model as a structure for a good argument in computational environments that teach argumentation - because such environments are in many ways more strictly bound to a pre-defined structure than human teachers - it will be necessary to develop systematic knowledge about what kinds of arguments Toulmin's model of argument is suitable for.
Secondly, our study illustrates that the tutorial conversation had the desired positive effect when compared with no intervention. However, our study cannot make comparative statements towards other interventions, amongst them technically simpler ones like asking study participants to re-work their answers after re-reading educational material about Toulmin's argument structure (comparable to the time spent on Task 1). Further experiments are needed to make statements about the efficiency of the agent in comparison to other interventions.

Finally, there is a range of wider-reaching directions for exciting future research. For instance, structural argument assessment, as is done by the conversational agent studied in this study, could be complemented by content-wise argument assessment. For this, additional approaches from argument mining such as clustering of arguments could be useful [Dax+20]. Such content-based assessment of arguments is the basis for further teaching strategies, such as teaching by giving counter-arguments.

7.7. Conclusion

In this Chapter, we presented a conversational agent that teaches Toulmin's model of argument via an argumentative task. By conducting a between-subjects experiment, we showed that study participants learned what they should - Toulmin's model of argument - and could apply it both in the same (Task 2) and in a new domain of argumentation (Task 3). This demonstrates the fundamental possibility that an AI-enabled tool can teach a structure for argumentation, which learners can then transfer to a setting outside the tutoring conversation as well as to a different argumentation domain. This study advances teaching argumentation in education as follows: This work indicates that Toulmin's model is a suitable structure to computationally analyse relatively short argumentation, as found in conversations. This work also constitutes one of the still few examples of an educational conversational agent that teaches a skill (argumentation), rather than facts.

8. Using Transformer-based Language Models within the Learning Engineering Process of Developing Educational Conversational Modules

In the previous chapter, we showed the positive impact of asking argumentative questions and providing argumentative feedback on the structural gap in learners' arguments. To have such conversational modules in various learning domains, this chapter, following the second goal of the thesis, focuses on conversational modules that can do the following: Given a definition, such as what constitutes a planet, when a situation falls under the GDPR (Europen General Data Protection Regulation), when an animal is considered a mammal, the conversational agent asks the learner to apply this definition to an example. As in a typical educational scenario, learners should apply the definition and reason about it, by making a claim, citing the given definition and giving evidence that the definition is fulfilled. The conversational agent asks follow-up questions if the learner does not reason sufficiently, i.e. if the claim, relationship to the definition, or evidence is missing. We consider this to be a conversational module that would be part of a larger educational agent (similar to the agents presented in Chapters 4 and 5), or part of a wider system rather than a stand-alone agent. We propose and investigate a systematic workflow that supports the learning engineering process of 1) formulating the starting question for such a conversational module based on existing learning materials, 2) specifying example statements as input to transformer-based language models (as classifiers) that will need to decide whether the claim, relationship to the definition and evidence are given in a user response, and 3) specifying an adaptive dialogue structure. Note that we specifically investigate transformer-based language models, such as underlie the widely discussed ChatGPT for instance, as enabling a relatively lightweight learning engineering process. In this chapter, we highlight 1) the potential of transformer-based language models to support learning engineers in developing dedicated conversational agents. As such models will continue to be improved, their benefits for learning engineering will rise; 2) the necessity to consider both classifier/adaptation mechanism quality and dialogue quality, because classifier quality is worse than the resulting dialogue coherence.

8.1. Introduction

An educational conversational agent needs content and an adaptation mechanism to be effective in educational scenarios and achieve an educational goal. The educational

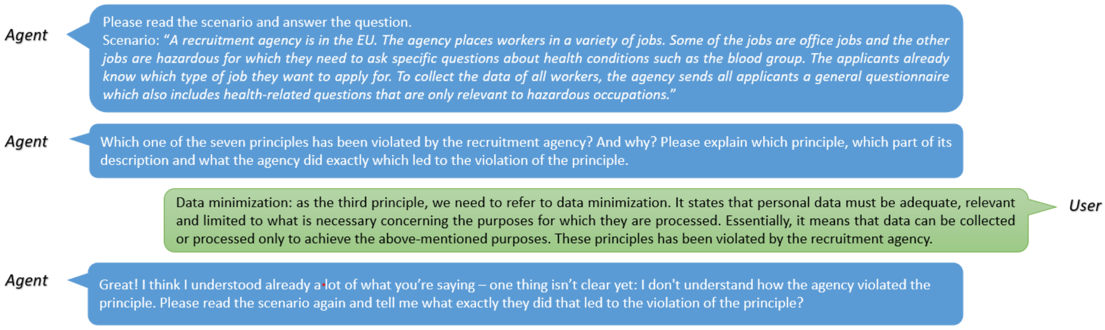


Figure 8.1.: A conversational module about the topic of the GDPR created with the workflow. The user's response is a real answer to the question.

content should be carefully structured and aligned with the adaptation mechanisms. From a technical perspective, the adaptation mechanisms behind an educational agent play a crucial role in its effectiveness. These mechanisms enable the agent to analyse and understand the user's inputs and provide personalised educational content.

In a typical educational scenario, it is essential for learners to not only understand a concept but also be able to apply it and argue about it effectively. In the field of education, argumentation tasks, such as posing argumentative questions, which require reasoning, play a crucial role in deepening students' knowledge in specific subject domains [VA+08]. Students not only need to develop valid arguments but also engage in scientific reasoning through argumentation [ESO04]. In general, argumentation serves as a heuristic method for developing a deeper understanding of scientific concepts [VA+08]. Effective questions should elicit argumentation, thereby stimulating creativity, critical thinking and boosting students' confidence [Chi07].

In this chapter, we are interested in conversational modules that can do the following: 1) provide a concept or a definition from the student's learning domain, 2) ask the learner to apply this definition to an example, and 3) give adaptive feedback to the learners on their reasoning. Such questions are central to students' cognitive development, and research evidence suggests that students' levels of achievement can be increased by regular access to higher-order thinking (e.g. [BH01]). For instance, in the learning domain of astronomy, given the definitions of a planet and a star, learners would be asked to apply the definitions to a specific example, such as Jupiter and reason about it. The question asked by the agent would be "*Based on the definitions, is Jupiter a planet or a star? Explain why?*" A complete answer from a learner needs to include a claim ("*Yes, Jupiter is a planet*"), making a connection between Jupiter and the definitions by citing the given definition and giving evidence related to Jupiter that fulfilled the definition. Fig. 8.1 shows another example, based on real user data, in the domain of GDPR (European General Data Protection Regulation). In Fig. 8.1, note that the last question of the agent responds adaptively to the user statement.

Note that we call this a conversational module to clarify that we think of it as needing to be part of a wider educational intervention - a lesson taught by a human teacher, a tutorial agent, or one element in a more traditional web-based learning (management) system.

The adaptation mechanism selects follow-up questions if the learner does not reason sufficiently. We base our understanding of what constitutes a sufficiently reasoned answer on Toulmin's model of argument [Tou03]. Concretely, we expect there to be a claim, i.e., whether the example fulfils a given definition, a warrant, i.e., an explicit mention of relevant parts of the definition, and evidence, i.e. an explicit relationship between the example and the definition. This corresponds to the core elements in Toulmin's model of arguments (ibid). Prior work shows this to be a suitable structure for argumentative answers or essays (e.g.,[HG17; WJL22]), even if other models of argument also exist (e.g. [Wha97]). Our prior work has shown that learners can learn how to argue with such a scaffold [MPS22b]. However, our prior work needed a complete machine-learning engineering process to develop the classifiers behind the adaptation mechanism. Such multi-step processes, typically including document annotation, iterative training and model calibration steps are still the dominant procedure in existing literature (e.g. [Wam+21; Wan+20; HG17]). As an alternative, related work knows semantic similarity-based approaches (e.g., a literature review [Amu+23]), which need, as the workflow proposed by us in this chapter based on transformer-based large language models, no multi-step training but rather example input statements (e.g., [Gra+01]). Transformer-based large language models refer to a type of deep learning models which contain encoders and decoders with self-attention capabilities [Vas+17]. These models are often pre-trained on massive datasets, allowing them to learn a broad understanding of language. These are, however, susceptible to differences in language. Transformer-based large language models now promise to combine the best of both worlds. These models have been analysed and discussed in other domains. For instance, in [Zha+23], the common challenges and opportunities of such models in the context of bioinformatics have been discussed.

In this work, we are now concerned with whether they indeed make the learning engineering process simple (which they do) whilst keeping up a reasonable technical and pedagogical quality. Specifically, we evaluated transformer-based large language models as classifiers which are responsible for identifying the components of Toulmin's model of argument. The evaluation was done through three different test cases to show the performance in different domains.

In this study, we propose a systematic workflow to provide content and adaptation mechanisms for a conversational module as described above. The workflow uses transformer-based language models, such as the now widely discussed large language models. Like this, the classical iterative machine-learning-based procedure can be avoided, and the process of learning engineering becomes viable for a larger group of potential learning engineers, such as teachers, educational technology consultants in educational institutions, or media and content developers. The workflow consists of three steps: 1) Defining the Initial Question - formulating the starting question for such a conversational module based on existing learning materials, 2) Defining Expected Phrases - specifying the input that transformer-based language models need to function as classifiers that decide about the subsequent turns that the conversational module takes 3) Defining the Dialogue Structure - specifying the adaptive dialogue structure, i.e., the turns the classifiers can choose between. This workflow is systematic, implying that the steps for developing a conversational module remain consistent across diverse learning domains.

The workflow aligns with the educational goals of teaching concepts, definitions, or terminologies through argumentation. Such a conversational module can be integrated, as one element, into a wider conversational interface or educational agent, as described in Chapter 4. Additionally, similar to the approach taken in [Dem+18], which focuses on structuring and guiding peer interaction with an emphasis on knowledge building, our proposed conversational module can be employed in Massive Open Online Courses (MOOCs) to support students and enhance knowledge acquisition by incorporating argumentative conversations through conversational agents. Following Weber et al.'s taxonomy of educational conversational agents [Web+21], we understand our conversational agent to be unspecific to different target groups (target group specificity would have to be achieved with the specification of the adaptive dialogue structure), to support learning factual knowledge and applying it, and thereby to support both practice at these levels, and preparation for subsequent learning phases.

This chapter is organised as follows. In Section 8.2 on "Background and Related Work", we elaborate on ongoing research on using pre-trained models in argument mining. In Section 8.3 on "Research Questions", we concretise the research questions that we ask and answer in this chapter. In Section 8.4 "The Systematic Workflow", we describe the workflow in detail. In Section 8.5 ("Test Cases"), we describe the three different conversational modules created as test cases. In "Evaluation Methodology" (Section 8.6), collecting data and the method used to answer the research questions are described. The results are shown in Section 8.7 ("Results"). We discuss the results in line with the research questions and conclude our work in Section 8.8, "Discussion and Conclusion".

8.2. Background and Related Work

8.2.1. Using Pre-trained Language Models in Argument Mining

The use of pre-trained language models has become popular in the natural language processing community. Converting words to numerical vector spaces that incorporate contextual information about words has been a prevalent approach which emerged due to rapid advances in neural networks. In these vector spaces, words or sentences with similar meanings are positioned closer to each other. This representation therefore captures semantic relationships between them, allowing algorithms to understand and work with the contextual meaning of them more efficiently. These word embedding models have outperformed traditional approaches in many natural language processing tasks, especially in argument mining. In the last few years, many different pre-trained language models have been created and used that have been leveraged differently in argument mining. For instance, in [HG17], they used word embedding vectors trained on part of the Google News dataset [Mik+13] to identify the argumentative components such as claims and premises. In [Wam+21], they trained a predictive model following BERT architecture to classify text tokens as claim, premise or non-argumentative following Toulmin's model of argument [Tou03] in students' essays about the topic of "*Does TV make stu-*

dents aggressive?" To train the model, they used a German corpus which contains 1000 business model peer reviews written by students [Wam+20a].
Transformer-based language models were also used for converting sentences to numerical vectors in order to be used as features or measuring the similarity. In [Xia+22], the authors employed BERT as a feature extractor to train a series of machine-learning models (e.g., Logistic Regression, Random Forest) for identifying the argumentative components and relations. As training data, they collected and annotated 1269 sentences including 164 discussion threads covering eight topics, including abortion, dating, eugenics, immortality, marriage, parenthood, pride, and suicide. In [Abr+22], they proposed a framework which contained two sub-models, namely intent classifier and argument similarity. In the latter, they looked for the most similar argument which referred to the user's utterances. To produce high-quality sentence representations, which are needed to measure the similarity, they combined contextualised word features from the BERT with some additional information and then used cosine similarity to compare them with the arguments in the system.

The implementation of machine learning and artificial intelligence techniques for argument mining relies heavily on the availability of annotated documents, which serve as a training set for predictive models. However, the process of constructing an annotated dataset is a complex and time-consuming endeavour, requiring substantial resources such as expert teams to ensure the acquisition of consistent and homogeneous annotations [LT16]. Furthermore, it is important to note that different datasets are often created with specific objectives or for particular genres, making them less suitable for all approaches or all stages of the argumentative tasks (see [Wam+21; Wam+22]).

One of the goals of the workflow presented in this chapter is to minimise the required engineering such as collecting huge amounts of data and annotating processes for identifying the core components in various learning domains. To this end, we utilise pre-trained large language models Sentence-BERT (SBERT) [RG19; RG20a], a modification of the pre-trained BERT network that use Siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. By utilising SBERT models and collecting a few reference samples for each Toulmin component, they can be identified.

To our knowledge, no prior studies have investigated how to have a systematic workflow for developing such adaptive-argumentative dialogues for conversational agents. To fill the gap, our goal has been to create a systematic workflow for creating a conversational module which consists of a dialogue structure and an adaptation mechanism for argumentative questions on different topics. In other words, by taking the steps of the workflow, we can ask an argumentative question, create machine learning classifiers with pre-trained models to identify Toulmin's core components and define the required branches in the dialogue based on the core components.

8.2.2. Synthesis

Building on the above-described prior work, in this work we are interested in the learning engineering process of building, for a specific learning topic, a conversational module that

supports learners in developing a full argumentation. This is different from what was done in [Lol+12] in which the focus was on the visualisation of arguments and collaboration. There are two main reasons why this process is labour-intensive: Firstly, *annotation of training data*. To have an adaptation mechanism that can support argumentation, computational argument-mining techniques are required to analyse and understand argumentation within different learning domains. Conducting argument mining through machine learning and artificial intelligence techniques needs the availability of annotated documents which serve as training sets for predictive models [LT16]. Creating an annotated dataset is a complex and time-consuming task that often requires substantial resources, including teams of experts, to ensure consistent and homogeneous annotations [Sch+18]. For instance, in the work by [LGP18], multiple calibration phases were conducted to enhance inter-annotator agreement, resulting in increased costs. This implies that creating a sufficiently large annotated dataset for new domains becomes impractical, leading to degraded performance in argument structure parsing due to limited data availability [LT16]. Additionally, different datasets are often constructed with specific objectives or for particular genres, making it difficult to find a single dataset that suits all approaches, domains, or stages of argument mining [LT16]. Subsequently, corpus annotation represents a severe knowledge engineering bottleneck for systems based on computational linguistics, specifically also true for argument mining-based systems [Sch+18]. Such annotation was done in much of the above-discussed prior work (e.g. [Wan+20; HG17]). Transformer-based large language models are pre-trained, on the other hand, and in this work, we investigate the performance achievable out-of-the-box with them.
Secondly, *content development*. Any educational systems, conversational agents or interfaces need educational content. As with any adaptive system, this goes beyond the main content and includes all kinds of feedback or responses given. In the case of educational conversational agents, this corresponds to the adaptive dialogue design. In the present work, we investigate a systematic - for one particular type of conversation - for building an adaptive dialogue structure.

Finally, learning engineering has a particular challenge at the intersection of technical and pedagogical design: Technological capability and pedagogical design need to be well aligned with each other. Subsequently, we investigate a workflow in which technical and pedagogical development of the conversational module are treated as tightly inter-related to the extent of being part of a single design workflow.

While learning and technologies literature does contain prior work on conversational educational agents; and there exists a substantial body of literature on argument mining techniques, literature on the learning engineering process of conversational educational agents is extremely scarce, to the extent that we identified only widely related literature on authoring for adaptive educational systems for instance [Spe+01; Bru+05; Man+23].

8.3. Research Questions

The workflow we investigate in this chapter has three main steps. Following the steps, we can create a conversational module containing a starting question based on existing learning materials, analysing the learners' answers using transformer-based language models and specifying the adaptive dialogue structure. We ask two research questions that target two different aspects of the quality of the conversational modules:

- RQ1 (classifier quality): How well do the pre-trained language models identify the selected Toulmin model components (claim, warrant, evidence) given manually pre-defined example input statements?
- RQ2 (conversational coherence): How coherent are agent follow-up questions to a user response based on classifier results and using the systematically defined adaptive dialogue structure?

Regarding RQ1, we use the F1-macro score as a measure of the classifier quality. RQ2 is based on the concept of "conversational coherence". This concept refers to the quality of turns in a dialogue to reasonably follow up on each other [VD77]. In previous works, this concept has been used to measure the quality of a user-agent conversation (e.g. [Wol+22]), as a factor that impacts the user experience of interaction with a conversational agent. In this work, we assess whether the next follow-up question of the agent is coherent (details in Section 8.6). Note that having a reasonable performance in the classifier quality is a prerequisite for having coherent conversations. However, ideally, dialogue design can serve as a cover-up for mis-classifications.

In order to evaluate the systematic workflow for developing a conversational agent module, firstly, we define three test cases (Section 8.5). For each test case, we apply the systematic workflow in order to develop a conversational agent module. Secondly, we collect data in each of the test cases in order to answer both RQ1 and RQ2 (Section 8.6).

8.4. The Systematic Workflow

The goal is to develop a conversational agent module in which the learner is asked to develop an argumentative answer to a question and is guided by the conversational agent towards developing a full argument if argumentative parts are missing (cp. Fig. 8.1). We assume that whoever creates the conversational agent module has both some learning domain knowledge and some technical knowledge. This could be for instance a single learning engineer who combines instructional and technical expertise, or a team consisting of an instructor and an engineer. For simplicity, below we always write "the learning engineer".

The workflow consists of three main steps. Firstly, the learning engineer formulates the starting question for such a conversational module based on existing learning materials (Section 8.4.1). Secondly, the learning engineer needs to specify the input or the expected answers that transformer-based language models need to function as classifiers that decide about the subsequent turns that the conversational module takes (Section 8.4.2). The output of the second step contains example phrases that represent what reasonable claim, warrant and evidence (following Toulmin's model of argumentation) to the agent's question defined in the first step. In the third step, the learning engineer specifies the adaptive dialogue structure, i.e. the turns the classifiers can choose between.

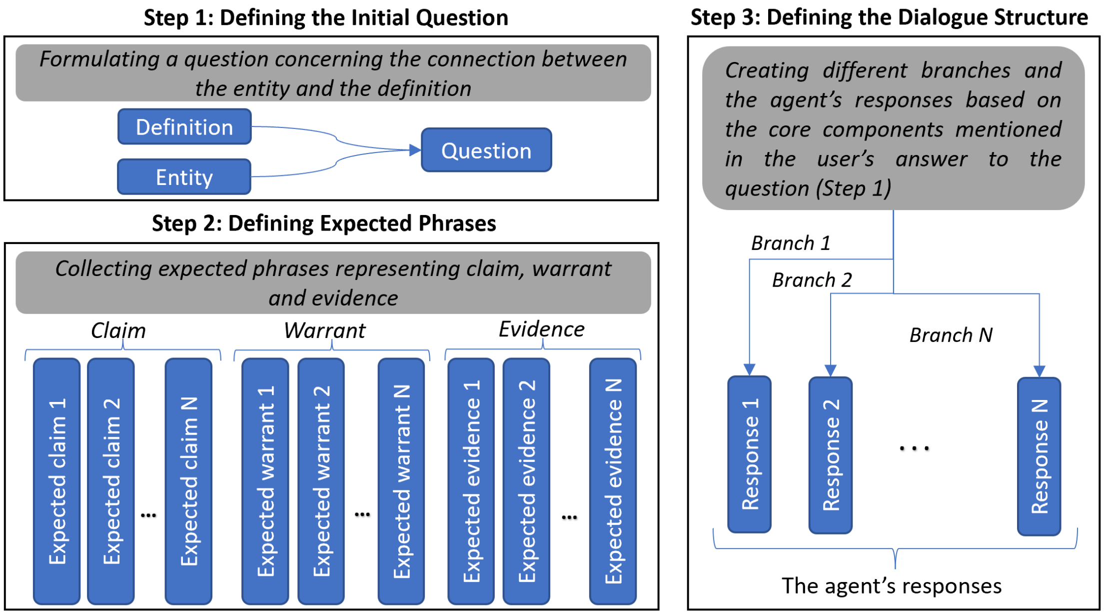


Figure 8.2.: The outline of the workflow for engaging in an argumentative conversation. The steps correspond to tasks assigned to a learning engineer. The pre-trained transformer models handle the classification of learners' responses, utilising expected phrases (it is not depicted separately in the figure).

It includes follow-up questions in cases where the learner's response misses core elements of Toulmin's model and thus does not provide a fully developed answer to the agent's question (Section 8.4.3). Fig. 8.2 summarises the workflow.

8.4.1. Step 1: Defining the Initial Question

The initial question of the conversational module consists of two main parts, 1) a definition or term mentioned in learning materials, and 2) an entity or a specific example. Given the definition and the example, the question should ask the learner to explain why/why not the given example fulfils the definition or not. Table 8.1 shows examples of such questions, including a definition and a given example, in various learning domains. For instance, in the topic of Biology, the definitions of the animal types could be as follows: *the mammal is a type of animal with warm blood and a hairy body. Examples of mammals are cows and elephants. Reptiles are types of animals with cold blood and dry scaly skin. Examples of Reptiles are snakes or crocodiles.* Given the definition, the learners are asked which definition can be fulfilled by lions as an example. All the required information to answer the question can be provided by the conversational agent during the conversation or by teachers beforehand.

Depending on the learning domain, the example is very short (e.g., "a lion", "Jupiter" in Table 8.1) or is longer (e.g., an elaborate scenario for which applicability of the GDPR should be decided (ibid). Different methods can be used by the learning engineer to sup-

Table 8.1.: Example study topics, definitions and argumentative questions that could be asked of students. The *entities in italic* in the example questions of course are variable.

Topic	Definition	Example question
Biology	The definitions of five types of animals: mammals, reptiles, birds, amphibians, insects	Is *a lion* a mammal or not? Explain Why?
Astronomy	The definition of a planet	Is *Jupiter* a planet or not? Explain Why?
Intelligence	The definition of intelligence	Is *a monkey* intelligent or not? Explain Why?
Physics	The definition of the state of matter	Is *a bottle of milk* solid, liquid or gas? Explain Why?
The GDPR	The definition of the GDPR principles	Based on *the scenario*, Which one of the seven principles has been violated? Explain Why?

port argumentation such as providing extra direction or prompting questions [JK10b]. The inclusion of "*Explain why?*" or similar phrases can be used to ask learners to construct an argument and to justify their claim using the provided definitions and the example [JK10b].

8.4.2. Step 2: Defining Expected Phrases

We view each user's response as an argument, employing Toulmin's framework, wherein the direct response serves as the claim, the relevant information related to the entity (or the example) is regarded as the evidence, while the segment of information within the definitions that establishes a connection between the evidence and the claim represents the warrant component. To give adaptive feedback, the learning engineer generates representative phrases for claim, warrant, and evidence. These are used as inputs for pre-trained transformer-based language models in the conventional module. If the learners' responses lack sufficient reasoning, such as a missing claim, warrant, or evidence, the conversational agent asks follow-up questions to support the learners in filling in the missing argumentative components in their answers.

The diversity of the expected phrases determines the number of branches in the dialogue and the agent's feedback. For instance, in this question "*What is the state of matter (solid, liquid or gas) of honey?*", if the learning engineer aims to assess only the correctness of the learners' answers, the expected phrases of the claim component should cover the correct answer such as "*Honey's state of matter is liquid*". However, to offer specific feedback on common errors and misconceptions, the learning engineer needs to provide expected phrases that address these typical mistakes, such as "*I think honey should be solid*".

Besides the claim component, the learning engineer should define the expected phrases of warrant and evidence component. In this example, the expected phrases of the warrant consist of the phrases which cover the key points within the definition of each state of matter such as "*Particles roll over each other and settle on the bottom.*" The expected phrases of evidence include the relevant information about the honey (used as an example by the learning engineer in the first step of the workflow) to support the claim such as "*it takes the shape of the container.*" In the case of using a scenario to describe an entity, the learning engineer restricts the number of expected phrases of evidence to the pieces of information mentioned in the description that can support the claim.

8.4.3. Step 3: Defining the Dialogue Structure

In this step, the learning engineer first defines the number of branches based on the expected phrases for each core component and then defines the agent's responses or the follow-up questions for each branch, A complete answer should consist of all three core components. By focusing only on the existence of each core component in the learners' answers, the learning engineer is capable of covering eight different branches as follows: "*with_claim*" and "*without_claim*" for the claim component, "*with_warrant*" and "*without_warrant*" for the warrant component and "*with_evidence*" and "*without_evidence*" for the evidence component.
As mentioned in the previous step, the conversational module can be more adaptive by having specific branches which deal with the wrong claims or common mistakes of learners. In this case, a ternary value is needed for the claim component, "*correct_claim*" for the correct answers, "*incorrect_claim*" for incorrect answers such as "*honey is solid*", and "*without_claim*" for answers in which the claim is missing. Having a ternary value for the claim and two binary values for the warrant and the evidence, twelve different branches can be generated. The twelve branches can be reduced to only six branches because providing feedback on the warrant and evidence when the claim is incorrect or missing is not reasonable and may cause misunderstandings. In other words, the agent's feedback should address first the missing or incorrect claim component, as the other two components are meant to support the correct claim. Therefore, the initial twelve branches can be reduced to six branches, as outlined below:

1. *incorrect_claim*
2. *without_claim*
3. *correct_claim, with_warrant, with_evidence*
4. *correct_claim, with_warrant, without_evidence*
5. *correct_claim, without_warrant, with_evidence*
6. *correct_claim, without_warrant, without_evidence*

After determining the required branches, the learning engineer needs to define the agent's feedback and/or the follow-up questions for each branch. For example, a possible

Table 8.2.: The summary of the test cases.

Test case	Question (Step 1)	# of expected phrases (Step 2)	# of branches in the dialogue structure (Step 3)
TC1	Based on the scenario, which one of the seven principles has been violated by the recruitment agency? And why?	38	6
TC2	Based on the scenario, which one of the seven principles has been violated by the hospital? And why?	31	6
TC3	Based on the definitions, do you think monkeys are intelligent or not? And why?	78	5

answer to the question about the state of matter of honey (see above) might be this: "*I think honey is liquid because it can flow and takes the shape of its container.*" This response includes a claim ("*I think honey is liquid*") and evidence ("*it can flow and takes the shape of its container.*"), and the warrant component that establishes the connection between the claim and evidence is missing. Based on the different values of expected phrases for each component, the corresponding branch to such answer is "*correct_ claim, without_ warrant, with_ evidence*". Due to the goal of the agent which is supporting the learners to write a complete answer, a possible agent's response could be: "*Great! I think I understood already a lot of what you're saying – one thing is not clear yet: I don't understand based on which part of the definition of states of matter, you think honey is liquid. Please read the definitions again and tell me which part of it supports your claim.*". In the next section, we created three different test cases following the steps of the workflow.

8.5. Test Cases

8.5.1. Test Case 1

Test Case 1 concerns the European General Data Protection Regulation (GDPR). The GDPR introduces new definitions and frameworks for the handling and management of personal data. As a result, organisations are required to adapt to the concepts outlined in the GDPR. This topic holds significance for a wide range of professions, making it a typical subject covered in Massive Open Online Courses (MOOCs) at an introductory level. By addressing the GDPR in our test case, we aimed to assess the applicability and effectiveness of our systematic workflow in a context that is relevant and valuable for various professional domains.

Table 8.3.: The definitions of GDPR's principles

The definition of GDPR's principles
Data minimisation: As the third principle, we need to refer to data minimisation. It states that personal data must be adequate, relevant and limited to what is necessary concerning the purposes for which they are processed. Essentially, it means that data can be collected or processed only to achieve the above-mentioned purposes.
Accuracy: Accuracy is the fourth principle meaning that it is required to ensure that personal data are accurate and are kept up to date where it is necessary. The data should also accurately reflect the order of events. Inaccurate personal data – considering the purposes for their processing – must be deleted or rectified without any delay.

Defining the Initial Question

To define the initial question, we used the seven GDPR principles (Lawfulness, fairness and transparency, Purpose limitation, Data minimisation, Accuracy, Storage limitation, Integrity and confidentiality and Accountability - cp. Table 8.3). In addition to the definitions, we needed an entity or an example to which these definitions could be applied. We defined example scenarios based on existing learning materials around GDPR. The example scenario used in our evaluation is as follows:

> Question 1: *A recruitment agency is located in the EU. The agency places workers in a variety of jobs. Some of the jobs are office jobs and the other jobs are hazardous for which they need to ask specific questions about health conditions such as the blood group. The applicants already know which type of job they want to apply for. To collect the data of all workers, the agency sends all applicants a general questionnaire which also includes health-related questions that are only relevant to hazardous occupations. Which one of the seven principles has been violated by the recruitment agency? And why? Please explain which principle, which part of its description and what the agency did exactly which led to the violation of the principle.*

Based on the scenario, the correct answer or the violated principle is the data minimisation principle (cp. 8.3).

Defining Expected Phrases

In Step 2 (defining expected phrases), we created a list of expected phrases for each component. Considering the question, learners can select any of the seven GDPR principles as their claim. To cover all principles, we added seven statements with the pattern of "*X has been violated*" to the expected list of claims. Here, "*X*" represents the name of each principle, such as "*data minimisation principle has been violated*". All the expected phrases related to each test case are listed in Appendix A.4.

Regarding the creation of the expected list of warrants, our focus was primarily on the definition of the violated principle. The list consisted of six statements, including examples such as "*personal data must be adequate, relevant and limited to what is necessary in relation to the purposes*".
Given the scenario describing the imaginary entity (the recruitment agency), the evidence component of users' answers should consist of statements or phrases extracted from the scenario that highlight the factors leading to the violation. We included 13 statements in the expected list of evidence, such as "*the blood groups were not related to all applicants*".

Defining the Dialogue Structure

In this step, we defined the dialogue branches and the agent's responses for each branch. As explained in Section 8.4.3, these responses should assist users in mentioning all the essential components and contribute to a natural and coherent conversation. For example, in case of missing evidence in the user's response, the corresponding branch would be "*correct_claim, with_warrant, without_evidence*". In this branch, the agent's feedback would be: "*Great! I think I understood already a lot of what you're saying – one thing isn't clear yet: I don't understand how the agency violated the principle. Please read the scenario again and tell me what exactly they did that led to the violation of the principle.*" The feedback starts with a positive statement to acknowledge that the claim was correctly mentioned, followed by a request for the missing evidence component. Similar feedback statements have been crafted for the other branches as well. These statements aim to assist users in addressing any structural gaps in their answers and maintaining a coherent and meaningful conversation. All the agent's responses were listed in Appendix A.5.

8.5.2. Test Case 2

Defining the Initial Question

For the second test case, we continued with the topic of GDPR and used a different scenario in which another principle was violated. The question and the scenario are as follows:

> Question 2: "*In a hospital, for each patient, only the last diagnosis of a medical condition continues to be held and the previous diagnoses are deleted. Now Bob who has been in the hospital for six months wants to know why his treatments have been changed monthly. The hospital cannot answer his question because they just keep the last diagnosis of each patient and delete the old ones. Which one of the seven principles has been violated by the recruitment agency? And why? Please explain which principle, which part of its description and what the agency did exactly which led to the violation of the principle.*

In this particular example, the principle that has been violated is the accuracy principle, which is explained in Table 8.3.

Defining Expected Phrases

Similar to the question in Test Case 1, an answer needs to have a claim which refers to one of the GDPR principles violated in the scenario by the entity (the hospital). The expected phrases for the claim component were identical to those created for Test Case 1. However, the correct or violated claim for Test Case 2 was the accuracy principle. The expected warrants included phrases in the violated principle's definitions such as "*the data should also accurately reflect the order of events*". Similarly, the expected evidence for this question consisted of nine statements mentioned in the scenario, which led to the violation of the accuracy principle, for instance, "*only the last diagnosis of a medical condition continues to be held*".

Defining the Dialogue Structure

In Test Case 2, we utilised the same branches defined for Test Case 1. However, in Test Case 2, the "*correct_claim*" referred to the accuracy principle. For each branch, a suitable response for the agent was defined. The agent's responses should help learners to mention all the required components in their answers, and also, the agent's responses should keep the conversation coherent. For example, in the "*without_claim*" branch in which the claim component is missing, the agent would respond with: "*Mmmm, could you clearly specify which principle was violated? And explain why?*" Since the claim component is the most crucial one, if the correct claim is missing, the response should solely focus on obtaining the claim, regardless of the status of the other components.

8.5.3. Test Case 3

Defining the Initial Question

In the third test case, we explored the topic of intelligence, which has been previously addressed in our previous chapters. The selection of this topic was based on the aim of enhancing AI literacy [LM20]. We focused on the definitions of intelligence and utilised five definitions of intelligence: acting rationally, acting humanly, thinking rationally, thinking humanly, and the ability to learn from experiences. The first four definitions are connected to intelligence and its impact on the development of artificial intelligence, and they have influenced various directions of AI research [Rus10]. However, the fifth definition aligns more closely with the understanding of learning in psychology and learning sciences. As for the entity, we used monkeys, and the question for this test case is as follows:

> Question 3: *Based on the definitions, do you think monkeys are intelligent or not? Please explain why?*

Defining Expected Phrases

To generate the expected statements for each core component, we took a different approach than in Test Cases 1 and 2. We leveraged a dataset that was presented in Chapter

6. This dataset contains annotated answers for the question specified in Test Case 3. The dataset encompasses 1337 answers related to 12 different entities. From the 155 available answers about monkeys, we randomly selected 20 answers and utilised them to form the expected phrases for each component.
The resulting list of expected claims, warrants, and evidence consists of 20, 35, and 39 statements, respectively. For example, an expected phrase for the claim could be "*monkeys are intelligent*", for the warrant it could be "*it can be observed thinking and behaving like a human*", and for the evidence, it could be "*monkeys have acquired the skills to use basic tools*".

Defining the Dialogue Structure

In Test Case 3, we employed similar branches to those defined in Test Cases 1 and 2, with two modifications. Firstly, we removed the "*incorrect_claim*" branch, as there was no specific correct answer to the question. Given different definitions, the claim could be different. Secondly, we renamed the "*correct_claim*" branch to "*with_claim*". Since there is no specific correct answer (claim), the focus should be on the presence or absence of a claim rather than its correctness. Therefore, we defined only five branches, one related to the answers without a claim, and four more branches for answers with a claim and different values of warrant and evidence.

Once the branches have been finalised, it is important to specify a meaningful response for each branch. For example, in this branch ("*with_claim, with_warrant, without_evidence*"), we defined the agent's feedback as "*Interesting! But the evidence part is missing! Could you explain why you think monkeys are intelligent based on the definition(s) that you mentioned?*". This feedback aims to prompt the user to provide additional information or reasoning to support their claim, highlighting the importance of including evidence in their response.

8.6. Evaluation Methodology

Apparatus

The three above-described test cases were operationalised as web-based, interactive conversational agents using the Bazaar framework [Ada+14; KBR11][1], and inserting the classifiers developed via our systematic workflow.

Data Collection

We collected two different datasets (Datasets 1 and 2) for each test case which contained the sample answers to the initial questions. To collect data for Dataset 1, we asked our research team ($n = 7$) to answer the initial question related to each test case. All the answers were split into sentences and, overall, 89 sentences were collected, 31 sentences belonged to Test Case 1, 32 sentences belonged to Test Case 2 and 26 sentences belonged

[1] https://github.com/DANCEcollaborative/bazaar

to Test Case 3. Dataset 1 was utilised to compare pre-trained transformer-based language models (see Section 8.6.1).

We also asked the same questions to 100 unique MTurk workers to create Dataset 2 for Test Cases 1 and 2. For Test Case 3, we used parts of the data collected in Chapter 6: This dataset contains 156 answers to questions that correspond to Test Case 3. The remainder of the overall $1,335$ answers in the full dataset (ibid) are answers to questions that concern different examples. We used 136 out of 156 answers to create Dataset 2 and the rest (20 out of 156) were used to create the expected phrases for Test Case 3. All the answers were split into sentences and overall 700 sentences were collected, 199 sentences for Test Case 1, 209 sentences for Test Case 2 and 292 sentences for Test Case 3. Note that *this dataset was only used to evaluate the workflow, not to design the conversational modules.*

Annotation of Datasets and Inter-rater Agreement

Both Datasets 1 and 2 were coded for Toulmin's model's core elements, except for the subset related to Test Case 3, which was already annotated. New annotations done for this publication were done by two annotators (the first author included) and they coded the datasets used in Test Cases 1 and 2. Each annotation stated whether the three relevant Toulmin model elements (claim, warrant, evidence) are present in a user statement. In addition, the correctness of claim components for Test Cases 1 and 2 were also considered.

The annotation process consisted of three steps. Firstly, the annotators engaged in discussions to establish a consensus on the definition of each component (claims, warrants, and evidence). This step ensured a shared understanding among the annotators. Secondly, 50 sentences from Dataset 2 (25 from each test case) were randomly selected and independently annotated by the two annotators to assess inter-rater agreement. Finally, any disagreements that arose during the second step were discussed, and the remaining data were annotated by the first author. Cohen's kappa (κ) value was used to quantify the inter-rater agreement, taking into account the possibility of agreement occurring by chance. This statistical measure helped assess the level of agreement between the annotators. The κ values for the claim, warrant and evidence were 0.90, 0.62 and 0.71 respectively. The κ value for the claim was almost perfect, however, for the warrant and evidence, we achieved a substantial agreement. Table 8.4 represents the collected data for all three test cases based on the number of each argumentative component.

8.6.1. RQ1 (classifier quality): Comparing Pre-Trained Language Models

We assumed that a pre-trained large language model is chosen by the learning engineer based on some decision criteria (e.g., performance vs. available hardware, pricing, existing models that are already used within the socio-technical system of the learning engineer, etc.). We also assumed that the choice might matter a bit but not substantially. We tested this assumption informally for each test case and the three core components separately. To this end, for each test case and each component, we compared 16 different

Table 8.4.: The number of components in each dataset. For Question 3, we used one label, "*with_claim*", for both "*correct_claim*" and "*incorrect_claim*" labels. Dataset 1 was used to compare pre-trained models and it was not part of the workflow. Dataset 2 was used to evaluate the whole workflow (answering RQs 1 and 2).

Label	Test Case 1		Test Case 2		Test Case 3	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
# of *correct_claim*	10	33	9	13	5	139
# of *incorrect_claim*	5	66	5	83		
# of *without_claim*	16	100	18	113	21	153
# of *with_warrant*	8	15	5	10	8	111
# of *without_warrant*	23	184	27	199	18	181
# of *with_evidence*	9	38	8	54	15	119
# of *without_evidence*	22	161	24	155	11	173

pre-trained language models (SBERT models) such as *all-mpnet-base-v2* and *paraphrase-multilingual-mpnet-base-v2* [RG20b]. In addition to SBERT models, various other models and architectures can be employed. However, to minimise engineering efforts and model complexity, we opted for SBERT models, which come pre-trained, tuned, and well-documented, aligning with one of the workflow's objectives. Analysing the details of the models in terms of how they were trained and their structure was not in the scope of the current study. The list of all models and more details are described in Appendix A.6.

The selection of the best model involved comparing different models with a range of similarity thresholds based on the F1 score. A similarity threshold needed to be defined for each model by which the agent can decide about the final label of the users' statements. To this end, we defined a range of thresholds from 0.3 to 0.95 with a 0.05 increment for each model. For each model and its threshold value, we used the expected phrases (defined in the second step of the workflow) to identify the components in Dataset 1. For components with ternary values (e.g. the claim component in Test Cases 1 and 2 which includes "*correct_claim*","*incorrect_claim*" and "*without_claim*"), we used the F1 macro score. However, for binary components (e.g. warrants and evidence in all test cases, and claims in Test Case 3), we compared models based on the F1 score of the main class, such as "*with_warrant*", "*with_evidence*", and "*with_claim*".

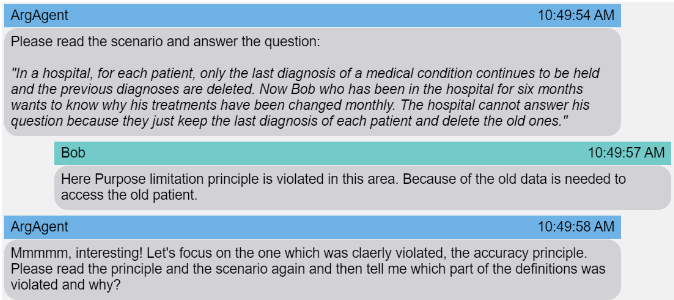


Figure 8.3.: A user response from Dataset 2, and the follow-up question that the selected classifier would choose. The correct answer to the initial question from Test Case 2 (the claim) is the accuracy principle. This example shows how the agent (ArgAgent)'s response can be coherent even though a misclassification happened. The user's response does not contain evidence, however, the agent classifies the second sentence as evidence which is incorrect.

8.6.2. RQ1 (classifier quality) on Dataset 2

Following the above comparison of available transformer-based large language models, we chose the best-performing model and the similarity threshold value for each test case and each component and then assessed its performance on Dataset 2. We chose the best-performing model as the LLM performance, in general, is still on the rise, so the resulting performance on Dataset 2 would still not be overly optimistic as an outlook for the future. To assess the performance of the best model on Dataset 2, we utilised Precision, Recall, F1 score and Accuracy. In addition, we employed the macro and weighted F1 scores.

8.6.3. RQ2 (conversational coherence)

In order to assess conversational coherence (RQ2), we coded the sequence of "the initial agent question - user response - follow-up agent question" for each test case, separately. The agent selected the follow-up questions or responses (defined by the learning engineer in the third step of the workflow) based on the classifier results from the selected/best classifier according to results from RQ1. The whole coding process was done by the first author. As a result, we give the percentage of coherent sequences in the overall number of statements in Dataset 2 ($n = 335$).

For example, Fig. 8.3 and 8.4 show such sequences that were coded for coherence. User responses to the initial questions of Test Cases 2 and 3 respectively from Dataset 2 are shown, as well as the agent response that would be made with the classifier selected for the evaluation results in Section 8.6.2.

8.7. Results

In this section, we show the results. All the experiments were run on a computer with an *Intel i7 (11800H)* processor running at $2.3 MHz$ using 32 GB of RAM, running on

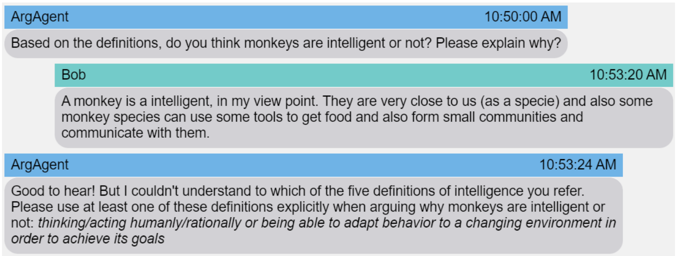


Figure 8.4.: A user response from Dataset 2 for the initial question from Test Case 3, and the follow-up question that the selected classifier would choose. The user has included a claim and evidence. The classifiers identified the claim and evidence correctly. Therefore, the agent (ArgAgent) only addresses the missing component and asks the user to use one of the definitions of intelligence as a warrant.

Table 8.5.: The mean (M) and standard deviation (SD) of F1 scores of all pre-trained models on Dataset 1 based on each component.

Dataset 1	**Claim**	**Warrant**	**Evidence**
Test Case 1	$M = 0.78$ $(SD = 0.07)$	$M = 0.76$ $(SD = 0.05)$	$M = 0.86$ $(SD = 0.08)$
Test Case 2	$M = 0.74$ $(SD = 0.04)$	$M = 0.88$ $(SD = 0.08)$	$M = 0.80$ $(SD = 0.07)$
Test Case 3	$M = 0.97$ $(SD = 0.03)$	$M = 0.88$ $(SD = 0.06)$	$M = 0.69$ $(SD = 0.07)$

Windows 11. The most time-consuming part of the computational time was related to downloading and loading the SBERT models which should be done once for each model. The details of each model can be found on the SBERT official website[2].

As it stands, the workflow was carried out, for each test case, once. Based on this experience, each test case took around a few hours to create a conversational module in a new learning domain. It is under the assumption that an operative large language model is available and selected, and the learning engineer has learning materials in traditional formats at hand (i.e. only needs to choose between definitions, and examples, not author them from scratch).

8.7.1. RQ1 (classifier quality): Comparing Pre-Trained Language Models

In each test case, we compared the F1 score of all 16 SBERT models on Dataset 1. The average performance of the models for each test case and component is shown in Table 8.5.

In Test Cases 1 and 2, we selected the best model for each core component based on

[2] `https://www.sbert.net/docs/pretrained_models.html`

Table 8.6.: The best model and its threshold (Th.) for each core component (Co.) based on Dataset 1.

Test Case 1			
Comnent	**Model**	**Th.**	**F1 score**
Claim	*msmarco-MiniLM-L12-cos-v5*	0.35	0.92
Warrant	*distiluse-base-multilingual-cased-v1*	0.50	0.87
Evidence	*all-mpnet-base-v2*	0.50	0.96
Test Case 2			
Claim	*paraphrase-multilingual-mpnet-base-v2*	0.55	0.84
Warrant	*all-distilroberta-v1*	0.60	0.95
Evidence	*paraphrase-MiniLM-L3-v2*	0.50	0.92
Test Case 3			
Claim	*all-distilroberta-v1*	0.50	0.94
Warrant	*msmarco-MiniLM-L6-cos-v5*	0.60	0.95
Evidence	*all-distilroberta-v1*	0.50	0.81

the F1 scores. Notably, for Test Case 3, we selected the second-best models for claims and warrants, as all metrics for the best model were 1.00, which could indicate a potential bias towards Dataset 1. Table 8.6 shows the best models and their associated similarity threshold for each core component. For instance, in Test Case 3, using the pre-trained model of *all-distilroberta-v1* with a similarity threshold of 0.5 resulted in the F1 score of 0.94 in identifying evidence on Dataset 1.

Overall, we note that differences between models are not substantial, which supports our assumption that model comparison and selection need not necessarily be part of the learning engineering process.

8.7.2. RQ1 (classifier quality) on Dataset 2

The best model for identifying the claim component in Test Case 1 was *msmarco-MiniLM-L12-cos-v5*, which achieved the highest F1 score (0.92) among the other models on Dataset 1. When applied to Dataset 2, this model achieved precision, recall, and F1 score values of 0.74, 0.91, and 0.82, respectively, for identifying incorrect claims. In Test Case 1, the macro average F1 scores of all the components on Dataset 2 ranged from 0.75 to 0.86, which is a promising result. Table 8.7 displays the results of identifying the core components in Test Case 1 on Dataset 2.

In Test Case 2, *paraphrase-multilingual-mpnet-base-v2* achieved the best F1 score on identifying claims in Dataset 1. By applying the model on Dataset 2, the model achieved precision, recall, and F1 score values of 0.87, 0.73, and 0.80, respectively, for identifying incorrect claims. Similar to Test Case 1, the results were encouraging. The macro average of F1 scores of all the components on Dataset 2 was between 0.76 and 0.83. In Table 8.8, the results related to Test Case 2 were shown.

The best model in Test Case 3 was *all-distilroberta-v1* which achieved the highest score on Dataset 1. Using the model on Dataset 2, it achieved the macro average F1 scores of 0.77, 0.74 and 0.66 for claims, warrants and evidence, respectively. The results for Test Case 3 are shown in Table 8.9. By analysing the performance of the classification of the core components in three different test cases, we showed how well the components can be identified in conversational modules created based on the workflow.

When we compare these results to results reported in Chapter 6, they are extremely encouraging: In Chapter 6, we had developed Random Forest classifiers based on a training data set with 1337 annotated user statements and achieved F1 scores of 0.77, 0.88, and 0.71 for claims, warrants, and evidence, respectively. Also in further related work in argumentation mining, the F1 scores we achieve in the here described work are absolutely comparable (e.g., F1 scores in the ranges of $60\% - 90\%$ in [SG17a; Moe+07; HG17; Wam+21; Wam+20b]).

These findings indicate that the systematic workflow, with its focus on leveraging pre-trained language models and minimising engineering efforts, can yield comparable performance to more complex models trained with specific data and fine-tuning. This demonstrates the potential of the workflow to streamline the development process and reduce the required engineering work to have such a conversational module in various learning domains.

8.7.3. RQ2: How coherent are agent follow-up questions to a user response based on classifier results?

In this section, we address RQ2, which focuses on the coherence of conversations produced through our systematic workflow. For each test case, we calculated the percentage of coherent dialogue turns including the agent's initial question, the user's response and the agent's follow-up question or response. The percentages of coherent conversations in Test Cases 1,2 and 3 are 84%, 79% and 80%, respectively.

8.8. Discussion and Conclusion

Our work has been motivated by the potential of modern conversational agents to support learning to argue, and learning through argumentation on the one hand; and the technology of transformer-based language models. The latter holds the promise to alleviate an acknowledged bottleneck in the creation of language-based adaptive systems: Creating substantial and domain-specific datasets (data collection and annotation). This bottleneck can be addressed with transformer-based language models, as they allow training classifiers only with very few training examples ("example phrases" in Step 2 of the workflow we presented, cp. Fig 8.2). This bottleneck comes on top of the challenge in adaptive educational systems to align well technology and (interactive) content.

In this work we investigate the promise of transformer-based language models by proposing and investigating a systematic learning engineering process for a conversational module that can do the following: 1) provide a concept or a definition from the student's learning domain, 2) ask the learner to apply this definition to an example, and 3) give adaptive feedback to the learners on their reasoning.

Table 8.7.: Test Case 1: The results of the selected models for Question 1 on Dataset 2.

Claim component			
Model	*msmarco-MiniLM-L12-cos-v5*		
Label	**Precision**	**Recall**	**F1 score**
incorrect_claim	0.74	0.91	0.82
without_claim	0.94	0.77	0.85
correct_claim	0.89	0.97	0.93
accuracy	0.85		
macro avg.	0.86	0.88	0.86
weighted avg.	0.86	0.85	0.85
Warrant component			
Model	*distiluse-base-multilingual-cased-v1*		
Label	**Precision**	**Recall**	**F1 score**
without_warrant	0.98	0.95	0.97
with_warrant	0.57	0.80	0.67
accuracy	0.94		
macro avg.	0.78	0.88	0.82
weighted avg.	0.95	0.94	0.94
Evidence component			
Model	*all-mpnet-base-v2*		
Label	**Precision**	**Recall**	**F1 score**
without_evidence	0.96	0.80	0.87
with_evidence	0.50	0.87	0.63
accuracy	0.81		
macro avg.	0.73	0.83	0.75
weighted avg.	0.87	0.81	0.83

Our findings show that the average of F1 scores of classifying Toulmin's core components - claim, warrant and evidence - in all three test cases was $M = 0.79$ ($SD = 0.06$). As described in the results Section 8.7.2, these results are absolutely comparable to the performance reported in prior literature *without necessitating the collection and annotation of training data.*
Further, our findings show that the percentage of coherent dialogue turns was 84%, 79% and 80% for Test Cases 1, 2 and 3 respectively.

These findings have implications for the development of educational technology, and teaching practice that makes use of educational technology. Firstly, for the development of educational technology, the results emphasise the boost that transformer-based language models give to adaptive learning technology, such as conversational modules: Without the necessity to go through the process of engineering a machine-learning-based system, a very reasonable performance of the overall educational system can be achieved. As transformer-based models continue to advance and improve, their benefits for learning

Table 8.8.: Test Case 2: The results of the selected models for Question 2 on Dataset 2.

Claim component			
Model	*paraphrase-multilingual-mpnet-base-v2*		
Label	**Precision**	**Recall**	**F1 score**
incorrect_claim	0.87	0.73	0.80
without_claim	0.79	0.91	0.85
correct_claim	0.78	0.54	0.64
accuracy	0.82		
macro avg.	0.81	0.73	0.76
weighted avg.	0.82	0.82	0.84
Warrant component			
Model	*all-distilroberta-v1*		
Label	**Precision**	**Recall**	**F1 score**
without_warrant	0.98	0.99	0.98
with_warrant	0.75	0.60	0.67
accuracy	0.97		
macro avg.	0.87	0.79	0.83
weighted avg.	0.97	0.97	0.97
Evidence component			
Model	*paraphrase-MiniLM-L3-v2*		
Label	**Precision**	**Recall**	**F1 score**
without_evidence	0.94	0.86	0.90
with_evidence	0.69	0.85	0.76
accuracy	0.86		
macro avg.	0.82	0.86	0.83
weighted avg.	0.88	0.86	0.87

engineering are expected to increase. These models offer powerful natural language processing capabilities, enabling more sophisticated and context-aware interactions between conversational agents and learners.

Secondly, the study highlights the importance of considering both classifier or adaptation mechanism quality and dialogue coherence when evaluating the performance of conversational agents. While the quality of the classifier or adaptation mechanism is crucial for accurate understanding and response generation, the resulting dialogue coherence plays a vital role in ensuring meaningful and engaging interactions with learners. This finding suggests that future research and development efforts should focus on optimising both aspects to enhance the overall effectiveness of educational conversational agents. For the development of educational technology, this means a shift from technology development towards learning engineering, in which technical competencies are important when developing educational technology, but major efforts required are knowledge engineering efforts that require domain and didactical domain knowledge. This is what is generally

Table 8.9.: Test Case 3: The results of the selected models for Question 3 on Dataset 2.

Claim component			
Model	*all-distilroberta-v1*		
Label	**Precision**	**Recall**	**F1 score**
without_claim	0.90	0.64	0.75
with_claim	0.70	0.92	0.80
accuracy	0.77		
macro avg.	0.80	0.78	0.77
weighted avg.	0.80	0.77	0.77
Warrant component			
Model	*msmarco-MiniLM-L6-cos-v5*		
Label	**Precision**	**Recall**	**F1 score**
without_warrant	0.76	0.94	0.84
with_warrant	0.85	0.51	0.64
accuracy	0.78		
macro avg.	0.81	0.73	0.74
weighted avg.	0.79	0.78	0.77
Evidence component			
Model	*all-distilroberta-v1*		
Label	**Precision**	**Recall**	**F1 score**
without_evidence	0.73	0.71	0.72
with_evidence	0.59	0.61	0.60
accuracy	0.67		
macro avg.	0.66	0.66	0.66
weighted avg.	0.67	0.67	0.67

understood as "learning engineering" then.
As the development of adaptive conversational modules becomes less of a technical, and more of a domain and didactical effort, creating specific conversational modules becomes realistic for individual educators. This will be an important shift in teaching practice if educators can easily learning-engineer their own conversational modules.

From a pedagogical perspective, asking argumentative questions and providing argumentative feedback on learners' responses not only leads to learning to argue but also arguing to learn. By leveraging transformer-based language models as used in the proposed workflow, argumentation can be injected into various learning domains. Using transformer-based language models enables us to reduce the time and resources traditionally required for developing argumentative conversational modules, making it a practical and efficient solution for modern classrooms. Furthermore, teachers can leverage conversational modules as dynamic tools for presenting concepts, prompting students to apply knowledge, and providing tailored feedback on their reasoning.

A main direction for future research following on this work of ours, as well as in connection to other ongoing research, is to carry out more user-oriented research. By this we mean, to evaluate a workflow like ours in user studies with people who would act as learning engineers (educators, or support staff close to them) who could realise this "ad-hoc creation" of adaptive learning technology. While our proposed workflow has the goal to support such learning engineers, the evaluation presented in this paper was still technical, by considering the qualities of the resulting agent, and not the ease of use of the workflow.

Furthermore, we highlight again that we understand our workflow to produce a “conversational module", i.e. something that per definition does not constitute a full educational system. Rather, this should be one piece in a larger educational ensemble: The conversational module could be a very specific interactive exercise given by a human teacher as part of a lesson; it could be embedded in a larger tutorial agent that teaches on a range of subjects and includes exercises, as given by the conversational module we investigate here, or this module could simply be part of a more traditional, web-based interactive learning system with some conversational-interface-exercises like ours included. We highlight this to point out at the end of our work that this underlines the argument for technology-enhanced learning research to be introduced in processes of streamlined production of adaptive technology-content modules: Real classrooms - at whatever level of education - need an enormous amount of content. On the other hand, technology-enhanced learning literature is full of very promising, very well-engineered systems that cover a tiny piece of the curriculum. We do not even want to argue that our workflow is the best, the only one, or solves a substantial part of the challenge of doing the learning engineering for providing adaptive technology-content ensembles for whole curricula, however, our work is one of very few that addresses the learning engineering process at all. We hope that this work does spark interesting future work that builds on existing technological advances, and solid learning sciences foundation.

9. Discussion and Conclusion

In this thesis, we followed two main objectives: First, we showed what and how full-fledged or full-tutorial educational conversational agents should look like and their affordance (in Chapters 4 and 5). Second, we investigated how argumentation can be leveraged and how it enhances the learning experience in such agents (in Chapters 6, 7 and 8).

9.1. Educational conversational agents in full-fledged learning environments

The first goal aimed to investigate the perceptions surrounding an educational conversational agent with which the learners encounter a complete learning experience including covering the learning materials, having quizzes and receiving personalised feedback. Such agents fill the gap in the literature in which the focus was mainly on the agents that dealt with a specific task and supported learners to finish the task (e.g. [Gro+19; Gra+05]). In Chapter 4, we showed two prototype agents and what such agents need to do to be considered full-tutorial conversational agents. Then, in Chapter 5, we developed a full-tutorial conversational agent, called DIGIBOT, and compared it with the DIGIVIDget, which presents the same learning content in the style of a state-of-the-art web-based learning management system.

9.1.1. Findings, Limitations and Future Works

The findings offer valuable insights into the capability of two different educational technologies which materialise two different interaction metaphors and how learners perceive each technology. The data reveals a generally positive reception towards using each educational tool. In addition, the results highlight the usefulness of each tool in inclusive education or for specific users with special needs. For example, the DIGIVIDget as a learning management system, is more suitable for users with self-regulated competence and also users who want to learn at their own pace. Furthermore, it is also suitable for learners who have concentration problems. However, DIGIBOT, the educational conversational agent, is shown to be useful for learners who have less self-regulated guidance. The study participants consistently highlighted the agents' ability to provide immediate and personalised feedback, fostering a sense of talking to a real person. Moreover, participants expressed appreciation for the interactive nature of the agents, which not only increased concentration but also enhanced motivation, especially in learners with special needs.

The main limitation is associated with the study participants, who were enrolled in a master's program specializing in inclusive education. The findings presented in Chapter 5 demonstrate the expert opinions regarding the perceptions of individuals with special needs towards each technology. Despite their experience in working with learners with special needs, it is imperative to acknowledge that they do not constitute authentic end users. Our findings can be enhanced by shifting the focus towards understanding the perspectives of real students or learners with special needs. This entails a comprehensive exploration of how these users experience each tool, with a specific aim to determine differences in their perceptions. The conclusions drawn in Chapter 5 remain valid when conceptualizing each educational tool as a generic, one-size-fits-all learning support system. However, in inclusive education, the student's needs vary significantly and educational tools should be adapted to their special needs [GC22]. To analyse the suitability of each tool with real learners with special needs, we can replicate the experiment and recruit learners with specific needs. For example, a detailed examination could be conducted to understand how visually impaired learners perceive web-based interactive learning materials in comparison to a conventional agent when both tools are equipped with voice modality. The experiment could be as follows:

1. Recruiting study participants with special educational needs such as students with visual or hearing impairments.

2. All the students are divided into different groups. In each group, the students try both educational tools in different order.

3. After trying both tools, they are asked to fill out a questionnaire regarding how they received each tool.

4. To have more insights, conducting a pre-and post-test to assess the learning gain and also a focus group discussion could be useful.

Conducting such an experiment would yield more valuable and nuanced insights into the usability and acceptance of each tool within the context of inclusive education.

9.2. Enhancing Learning Experience with Conversational Agents Using Argumentation

The second research objective delved into leveraging argumentation in conversational agents to enhance the learning experience. In Chapter 6, first, we showed the usefulness of Toulmin's model of argument [Tou03] and how the core components, claims, warrants, and evidence can be identified in learners' responses. Second, in Chapter 7, we investigated the impact of giving personalised feedback on the existence of the core components in the learners' responses. We also analysed the transferability of the acquired knowledge to writing an argumentative essay on a completely different topic. Finally in Chapter 8, we presented a workflow to be able to ask argumentative questions and provide adaptive feedback on learners' responses based on the existence of argumentative components in various learning domains. Following the workflow, the learning engineer can develop a conversational module in which the agent analyses the learners' responses and gives adaptive feedback on the structural wrongness in their arguments. Such conversational modules can be used as a dialogue turn in fully-tutorial conversational agents presented in Chapters 4 and 5.

9.2.1. Findings, Limitations and Future Works

The findings reveal acceptable results in identifying Toulmin's core components (Chapter 6), an encouraging improvement in writing more argumentative essays (Chapter 7) and promising results in identifying Toulmin's core components in various topics using pre-trained transformers models (Chapter 8).

Our work focused on Toulmin's model of argument [Tou03], and more specifically, we identified the three main components, claims, warrants and evidence. The first limitation is derived from the model and the components. There are additional argumentative components such as counterarguments which are an important element in actual human argumentation. The literature on students and argumentation revealed that the main problem in students' arguments is the lack of counterarguments [Lei03; Kos03; ECT17]. To enhance learners' argumentation, as a following work, we can focus on the existence of counterarguments in students' answers. To support learners in mentioning counterarguments in their essays, three different methods have been suggested in [JK10b]. These methods are used to push learners to write counterarguments in their responses. The methods are as follows:

- The first method is called Direction which provides a set of directions for constructing arguments.
- The second method is called Question Prompts which supports learners by prompting a set of following questions.
- The third method enables learners to see the structure of the argument (Graphical argumentation aid). By showing the structural gaps in students' arguments, they will notice what is missing in their arguments.

The mentioned approaches were used in real education settings such as classrooms and supervised by teachers, however, due to progress in educational conversational agents, the mentioned approaches can be implemented and handled by agents. Having an agent with the capability of helping users to write counterarguments using the mentioned methods can fill the gap in the argumentation literature. Comparing the impact of the three approaches on the number of written counterarguments can be considered a research question which can be answered by conducting a between-subject experiment in four steps as follows:

1. In a conversation with an agent, all students are asked to answer one specific argumentative question.

2. All the students are divided into different groups. In each group, the students receive a different type of feedback from the agent. For instance, the first group received such feedback "*Provide as many as reasons. Then discuss two reasons why others might disagree with you and why those reasons are wrong*" which is the direction method [JK10b]. The second group are asked to answer some questions such as "*How could somebody else disagree with you? How could you show that he or she is wrong?*" which is the question prompt method [JK10b]. Finally, the third group are shown a figure in which the missing argumentative components are illustrated (the graphical argumentation aid) [JK10b].

3. After receiving feedback, all students are asked to write an essay or answer an argumentative question mentioning all the argumentative components.

4. In the last step, we can compare the students' essays or answers written in the third step based on whether a counterargument is mentioned or not.

The second limitation of the thesis is focusing only on the structure of arguments and providing feedback on the existence of the core components. Addressing the structural wrongness in argumentation does not per se allow us to assess the content-wise plausibility of the made argument. However, analysing the content-wise plausibility of learners' argument would enhance the learners' experience in many directions. Based on the literature on educational conversational agents, analysing content-wise students' responses leads to more helpful and realistic feedback [Ara+23; MM11]. One way to analyse students' arguments content-wise is using large language models such as DialoGPT [Zha+19] or generative conversational agents such as ChatGPT [Ope23]. Using such models, the agents can provide feedback based on the existence of the core component and the content. Content analysis allows us to fulfil the lack of counterarguments in students' arguments as one of the main problems [Lei03; Kos03; ECT17]. We can leverage large language models in order to show what would be a counterargument to their own claim. To be able to generate a counterargument by a large language model, we need to collect a list of arguments and counterarguments and then fine-tune the large language model. Besides fine-tuning large language models, we can also use prompting methods with zero- or few-shot learning. For example, here, we used a real student answer and used it in the following prompt in OpenAI's ChatGPT [Ope23]:

> **Prompt:** Based on the question, the definitions and the student's answer, provide a counter-argument or an opposite answer to the student's answer. The counter-argument should be short and suitable in a conversation with a chatbot.
>
> The question: "*Are monkeys intelligent or not? Why?*"
>
> The definition: "*An entity is intelligent if it can act humanly, act rationally, think humanly, think rationally and be able to learn from experience towards its goals.*"
>
> The student's answer: "*No monkeys are not intelligent. They do not act rationally in stressful situations like humans are forced to. They throw feces.*"

In the prompt, we asked ChatGPT to generate a counter-argument based on the provided student's answer and the definitions of intelligence and we received this response:

> **ChatGPT:** While it's true that monkeys may not always exhibit rational behavior in stressful situations, it's important to note that intelligence can manifest in various ways. Monkeys demonstrate remarkable problem-solving abilities, social skills, and even exhibit learning from experiences in their natural habitats. Their cognitive abilities may not align with human standards of intelligence, but they display a unique form of intelligence suited to their environment.

The generated response by ChatGPT can be used as our agent's response. It can be even more tuned by adding sample responses and counter-arguments as examples (few-shot learning). Providing such feedback teaches the learners how a counterargument should be mentioned. The impact of such feedback on the argumentative skills of students can be measured by analysing essays written by students who received different types of feedback.

9.3. Conclusion

Conversational agents as a computer-supported educational technology simulate human conversations using language technology and AI algorithms. They engage in dialogues, interpret user input, and provide relevant responses, improving communication and offering personalized experiences. As an educational technology, they can provide learning in scale in various domains and teaching strategies. Furthermore, conversational agents enhance education accessibility for students with disabilities. Offering different communication modes, like voice commands or text-to-speech, breaks down barriers and ensures equal educational opportunities. Integrating conversational agents in education poses challenges. Designing agents to understand and respond accurately to diverse student needs is complex.

A key benefit of using conversational agents in education is personalized feedback and support. Students can get instant feedback, ask questions, and receive guidance, helpful for those struggling or needing extra help. Learning to argue and arguing to learn can also be supported by educational conventional agents. As for supporting argumentation, conversational agents can also cultivate argumentation and critical thinking. Engaging in debates or discussions with agents helps students practice constructing and defending arguments, evaluating evidence, and engaging with diverse viewpoints. In conclusion, conversational agents hold promise in education and argumentation. By providing feedback, promoting critical thinking, and enhancing accessibility, they can enhance learning and prepare students for a complex world. Continued advancements will likely yield even greater benefits in the future.

Bibliography

[Abr+22] Waheed Ahmed Abro et al. "Natural language understanding for argumentative dialogue systems in the opinion building domain". In: *Knowledge-Based Systems* 242 (2022). ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.108318.

[Ada+14] David Adamson et al. "Towards an agile approach to adapting dynamic collaboration support to student needs". In: *International Journal of Artificial Intelligence in Education* 24.1 (2014), pp. 92–124.

[Afr+21] Tazin Afrin et al. "Effective Interfaces for Student-Driven Revision Sessions for Argumentative Writing". In: *Proc. of the CHI*. 2021, pp. 1–13.

[AK01] Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2001.

[Akc+18] Damla Ezgi Akcora et al. "Conversational support for education". In: *International Conference on Artificial Intelligence in Education*. Springer. 2018, pp. 14–19.

[Ale17] K. R Aleven, V., McLaughlin, E. A., Glenn, R. A., Koedinger. *Instruction Based on Adaptive Learning Technologies*. 2017, pp. 522–560.

[AM20] Eleni Adamopoulou and Lefteris Moussiades. "Chatbots: History, technology, and applications". In: *Machine Learning with Applications* 2 (2020), p. 100006.

[Amu+23] Zaira Hassan Amur et al. "Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives". In: *Applied Sciences* 13.6 (2023), p. 3911.

[AP08] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics". In: *Computational linguistics* 34.4 (2008), pp. 555–596.

[Ara+23] Adelson de Araujo et al. "Automated coding of student chats, a trans-topic and language approach". In: *Computers and Education: Artificial Intelligence* 4 (2023), p. 100123.

[AS14] Mustafa Al Emran and Khaled Shaalan. "A Survey of Intelligent Language Tutoring Systems". In: *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014* (2014), pp. 393–399. DOI: 10.1109/ICACCI.2014.6968503.

[ATR14] Eliane Vigneron Barreto Aguiar, Liane M. Rockenbach Tarouco, and Eliseo Reategui. “Supporting problem-solving in Mathematics with a conversational agent capable of representing gifted students’ knowledge”. In: *47th HICSS*. IEEE. IEEE, 2014, pp. 130–137. ISBN: 9781479925049.

[Auf+08] Claudia von Aufschnaiter et al. “Arguing to learn and learning to argue: Case studies of how students’ argumentation relates to their scientific knowledge”. In: *Journal of Research in Science Teaching* 45.1 (2008), pp. 101–131. ISSN: 00224308. DOI: `10.1002/tea.20213`. URL: `http://doi.wiley.com/10.1002/tea.20213`.

[Bas+16] Pierpaolo Basile et al. “Argument mining on italian news blogs”. In: *Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. 2016.

[BCD07] Trevor JM Bench-Capon and Paul E Dunne. “Argumentation in artificial intelligence”. In: *Artificial intelligence* 171.10-15 (2007), pp. 619–641.

[BDA22] Abednego Kofi Bansah and Douglas Darko Agyei. “Perceived convenience, usefulness, effectiveness and user acceptance of information technology: evaluating students’ experiences of a Learning Management System”. In: *Technol. Pedagogy Educ* 31.4 (2022), pp. 431–449.

[Bea50] Monroe Curtis Beardsley. “Practical logic”. In: (1950).

[BH01] Paul Black and Christine Harrison. “Feedback in questioning and marking: The science teacher’s role in formative assessment”. In: *School science review* 82.301 (2001), pp. 55–61.

[Blo+56] Benjamin S Bloom et al. *Taxonomy of educational objetives: the classification of educational goals: handbook I: cognitive domain.* Tech. rep. New York, US: D. Mckay, 1956.

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[Bri+14] Christopher G Brinton et al. “Individualization for education at scale: MIIC design and preliminary evaluation”. In: *IEEE Transactions on Learning Technologies* 8.1 (2014), pp. 136–148.

[Bro+96] John Brooke et al. “SUS-A quick and dirty usability scale”. In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[Bru+05] Peter Brusilovsky et al. “Interactive Authoring Support for Adaptive Educational Systems.” In: *AIED*. 2005, pp. 96–103.

[BŠ14] Filip Boltužić and Jan Šnajder. “Back up your stance: Recognizing arguments in online discussions”. In: *Proceedings of the First Workshop on Argumentation Mining*. 2014, pp. 49–58. DOI: `10.3115/v1/w14-2107`.

[Cai+19] William Cai et al. “MathBot: A personalized conversational agent for learning math”. In: *ACM* (2019).

[CH20] Lisa A. Chalaguine and Anthony Hunter. "A persuasive chatbot using a crowd-sourced argument graph and concerns". In: *Frontiers in AI and Applications* 326 (2020), pp. 9–20. ISSN: 09226389.

[Cha+02] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[Cha+19] Lisa Andreevna Chalaguine et al. "Impact of argument type and concerns in argumentation with a chatbot". In: *2019 IEEE 31st ICTAI*. IEEE. 2019, pp. 1557–1562.

[Cha+20] Tuhin Chakrabarty et al. "Ampersand: Argument mining for persuasive online discussions". In: *arXiv preprint arXiv:2004.14677* (2020).

[Chi07] Christine Chin. "Teacher questioning in science classrooms: Approaches that stimulate productive thinking". In: *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 44.6 (2007), pp. 815–843.

[Cic+18] Analía Cicchinelli et al. "Finding traces of self-regulated learning in activity streams". In: *Proc. of the 8th inter. conf. on learning analytics and knowledge*. 2018, pp. 191–200.

[Cla+18] Fabio Clarizia et al. "Chatbot: An education support system for student". In: *International Symposium on Cyberspace Safety and Security*. Springer. 2018, pp. 291–302.

[Col+72] Kenneth Colby et al. "Experimental validation of a computer simulation of paranoid processes". In: *Mathematical Biosciences* 15 (Oct. 1972), 187–191. DOI: 10.1016/0025-5564(72)90073-9.

[CT17] Lucas Carstens and Francesca Toni. "Using argumentation to improve classification in natural language problems". In: *ACM Transactions on Internet Technology (TOIT)* 17.3 (2017), pp. 1–23.

[CV13] Elena Cabrio and Serena Villata. "A natural language bipolar argumentation approach to support users in online debate interactions†". In: *Argument and Computation* 4.3 (2013), pp. 209–230. ISSN: 19462166. DOI: 10.1080/19462166.2013.862303.

[CV18] Elena Cabrio and Serena Villata. "Five years of argument mining: A Data-driven Analysis". In: *IJCAI International Joint Conference on Artificial Intelligence* 2018-July (2018), pp. 5427–5433. ISSN: 10450823. DOI: 10.24963/ijcai.2018/766.

[Dax+20] Johannes Daxenberger et al. "Argumentext: argument classification and clustering in a generalized search scenario". In: *Datenbank-Spektrum* 20.2 (2020), pp. 115–121.

[DCV17] Mihai Dusmanu, Elena Cabrio, and Serena Villata. "Argument mining on Twitter: Arguments, facts and sources". In: *Proc. of the 2017 Conference on EMNLP*. 2017, pp. 2317–2322.

[Dem+18] Stavros Demetriadis et al. "Conversational agents as group-teacher interaction mediators in MOOCs". In: *LWMOOCS*. IEEE. 2018, pp. 43–46.

[DK20] Eva Durall and Evangelos Kapros. "Co-design for a competency self-assessment chatbot and survey in science education". In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 13–24.

[DNO00] Rosalind Driver, Paul Newton, and Jonathan Osborne. "Establishing the norms of scientific argumentation in classrooms". In: *Science education* 84.3 (2000), pp. 287–312.

[DO02] Richard Alan Duschl and Jonathan Osborne. "Supporting and promoting argumentation discourse in science education". In: 38.1 (2002), pp. 39–72.

[DV09] Vaille Dawson and Grady Jane Venville. "High-school Students' Informal Reasoning and Argumentation about Biotechnology: An indicator of scientific literacy?" In: *International Journal of Science Education* 31.11 (2009), pp. 1421–1445.

[DV10] Vaille Maree Dawson and Grady Venville. "Teaching strategies for developing students' argumentation skills about socioscientific issues in high school genetics". In: *Research in Science Education* 40.2 (2010), pp. 133–148.

[Dzi+10] Myroslava O. Dzikovska et al. "Beetle II: a system for tutoring and computational linguistics experimentation". In: *Proceedings of the ACL 2010 System Demonstrations*. 2010, pp. 13–18.

[EAYG06] Sibel Erduran, Dilek Ardac, and Buket Yakmaci-Guzel. "Learning to teach argumentation: Case studies of pre-service secondary science teachers". In: *Eurasia Journal of Mathematics, Science and Technology Education* 2.2 (2006), pp. 1–14.

[ECT17] Ibrahim Erdogan, Ayse Ciftci, and Mustafa Sami Topcu. "Examination of the questions used in science lessons and argumentation levels of students". In: *Journal of Baltic Science Education* 16.6 (2017), p. 980.

[EDG17] Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. "Neural end-to-end learning for computational argumentation mining". In: *arXiv preprint arXiv:1704.06104* (2017).

[EKKG15] Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. "On the role of discourse markers for discriminating claims and premises in argumentative discourse". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. September. 2015, pp. 2236–2242. ISBN: 9781941643327. DOI: 10.18653/v1/d15-1267.

[Erd06] Sibel Erduran. "Promoting ideas, evidence and argument in initial science teacher training". In: *School Science Review* 87.321 (2006), p. 45.

[Erd07] Sibel Erduran. "Methodological foundations in the study of argumentation in science classrooms". In: *Argumentation in science education*. Springer, 2007, pp. 47–69.

[ESO04] Sibel Erduran, Shirley Simon, and Jonathan Osborne. "TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse". In: *Science education* 88.6 (2004), pp. 915–933. ISSN: 00368326. DOI: 10.1002/sce.20012.

[Fes+19] Angela Fessl et al. ""Mirror, mirror on my search...": Data-Driven Reflection and Experimentation with Search Behaviour". In: *Transforming Learning with Meaningful Technologies: 14th EC-TEL, 2019, Delft, The Netherlands, Proc. 14*. Springer. 2019, pp. 83–97.

[Fes+21] Angela Fessl et al. "The Impact of Explicating Learning Goals on Teaching and Learning in Higher Education: Evaluating a Learning Goal Visualization". In: *EC-TEL*. Springer. 2021, pp. 1–15.

[FF00] Monique Frize and Claude Frasson. "Decision-support and intelligent tutoring systems in medical education". In: *Clinical and investigative medicine* 23.4 (2000), pp. 266–269.

[Fle43] R Flesch. *Marks of readable style: a study in adult education. , no. 897*. Teachers College Contributions to Education, 1943.

[Fle71] Joseph L. Fleiss. "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5 (1971), p. 378.

[FMP17] Alfio Ferrara, Stefano Montanelli, and Georgios Petasis. "Unsupervised detection of argumentative units though topic modeling techniques". In: *Proc. of the 4th Workshop on Argument Mining*. 2017, pp. 97–107.

[Fre11] James B Freeman. "Dialectics and the Macrostructure of Arguments". In: *Dialectics and the Macrostructure of Arguments*. De Gruyter Mouton, 2011.

[GC22] Sambhav Gupta and Yu Chen. "Supporting inclusive learning using chatbots? A chatbot-led interview study". In: *Journal of Information Systems Education* 33.1 (2022), pp. 98–108.

[Geo+20] Martha Georgiou et al. "Investigating the Impact of the Duration of Engagement in Socioscientific Issues in Developing Greek Students' Argumentation and Informal Reasoning Skills". In: *American Journal of Educational Research* 8.1 (2020), pp. 16–23.

[Gou+14] Theodosis Goudas et al. "Argument extraction from news, blogs, and social media". In: *Hellenic Conference on Artificial Intelligence*. Springer. 2014, pp. 287–299.

[Gra+00] Arthur C. Graesser et al. "Using latent semantic analysis to evaluate the contributions of students in AutoTutor". In: *Interactive learning environments* 8.2 (2000), pp. 129–147. ISSN: 1049-4820. DOI: 10.1076/1049-4820(200008)8.

[Gra+01] Arthur C Graesser et al. "Intelligent tutoring systems with conversational dialogue". In: *AI magazine* 22.4 (2001), pp. 39–39.

[Gra+05] Arthur C Graesser et al. "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue". In: *IEEE Transactions on Education* 48.4 (2005), pp. 612–618.

[Gra+99] Arthur C Graesser et al. "AutoTutor: A simulation of a human tutor". In: *Cognitive Systems Research* 1.1 (1999), pp. 35–51.

[Gro+19] Joshua Grossman et al. "MathBot: Transforming online resources for learning math into conversational interactions". In: *AAAI Story-Enabled Intelligence* (2019).

[GV00] Abigail S Gertner and Kurt VanLehn. "Andes: A coached problem solving environment for physics". In: *International conference on intelligent tutoring systems*. Springer. 2000, pp. 133–142.

[HG17] Ivan Habernal and Iryna Gurevych. "Argumentation mining in user-generated web discourse". In: *Computational Linguistics* 43.1 (2017), pp. 125–179.

[Hie+18] Ho Thao Hien et al. "Intelligent assistants in higher-education environments: the FIT-EBot, a chatbot for administrative and learning support". In: *Proc. of the ninth international symposium on information and communication technology*. 2018, pp. 69–76.

[HSA19] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. *A Survey on Conversational Agents / Chatbots Classification and Design Techniques A Survey on Conversational Agents / Chatbots Classi fi cation and Design Techniques*. March. Springer International Publishing, 2019. ISBN: 9783030150358. DOI: 10.1007/978-3-030-15035-8. URL: http://dx.doi.org/10.1007/978-3-030-15035-8_93.

[HW07] Hans Hansen and Douglas Walton. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press, 2007.

[IL17] H. N. Io and C. B. Lee. "Chatbots and conversational agents: A bibliometric analysis". In: *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. Vol. 2017-Decem. IEEE. 2017, pp. 215–219. ISBN: 9781538609484. DOI: 10.1109/IEEM.2017.8289883.

[Jag+21] Rohan Jagtap et al. "Healthcare conversational chatbot for medical diagnosis". In: *Handbook of research on engineering, business, and healthcare applications of data science*. IGI Global, 2021, pp. 401–415.

[JFY09] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. “Cutting-plane training of structural SVMs”. In: *Machine learning* 77.1 (2009), pp. 27–59.

[JK10a] David H. Jonassen and Bosung Kim. “Arguing to learn and learning to argue: Design justifications and guidelines”. In: *Educational Technology Research and Development* 58.4 (2010), pp. 439–457. ISSN: 10421629. DOI: 10.1007/s11423-009-9143-8.

[JK10b] David H Jonassen and Bosung Kim. “Arguing to learn and learning to argue: Design justifications and guidelines”. In: *Educational Technology Research and Development* 58 (2010), pp. 439–457.

[Joa98] Thorsten Joachims. “Text categorization with support vector machines: Learning with many relevant features”. In: *European conference on machine learning*. Springer. 1998, pp. 137–142.

[KBR11] Rohit Kumar, Jack Beuth, and Carolyn Rosé. “Conversational strategies that support idea generation productivity in groups”. In: (2011).

[Kim01] Beaumie Kim. “Social constructivism”. In: *Emerging perspectives on learning, teaching, and technology* 1.1 (2001), p. 16.

[Koe+13] Kenneth R Koedinger et al. “New potentials for data-driven intelligent tutoring system development and optimization”. In: *AI Magazine* 34.3 (2013), pp. 27–41.

[Koe+97] Kenneth R Koedinger et al. “Intelligent tutoring goes to school in the big city”. In: (1997).

[Kos03] Timothy Koschmann. “CSCL, ARGUMENTATION, AND DEWEY AN INQUIRY”. In: *Arguing to learn* (2003), p. 261.

[KSC06] Paul A. Kirschner, John Sweller, and Richard E. Clark. “Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching”. In: *Educ. Psychol.* 41.2 (2006), pp. 75–86.

[Kuc18] Udo Kuckartz. *Qualitative Inhaltsanalyse : Methoden, Praxis, Computerunterstützung*. ger. 2018.

[Kuh92] Deanna Kuhn. “Thinking as argument”. In: *Harvard Educational Review* 62.2 (1992), pp. 155–179.

[Kuh93] Deanna Kuhn. “Science as argument: Implications for teaching and learning scientific thinking”. In: *Science education* 77.3 (1993), pp. 319–337.

[KW04] Tim Kelly and Rob Weaver. “The goal structuring notation–a safety argument notation”. In: *Proc. of the dependable systems and networks workshop on assurance cases*. Citeseer. 2004, p. 6.

[Lei03] Selma Leitão. “Evaluating and selecting counterarguments: Studies of children’s rhetorical awareness”. In: *Written Communication* 20.3 (2003), pp. 269–306.

[Leo+21] Pui Huang Leong et al. “The Evaluation of User Experience Testing for Retrieval-based Model and Deep Learning Conversational Agent”. In: *Inter. Jour. of Advanced Computer Science and Applications* 12.4 (2021).

[LGP18] Anne Lauscher, Goran Glavaš, and Simone Paolo Ponzetto. “An argument-annotated corpus of scientific publications”. In: Association for Computational Linguistics. 2018.

[LK77] J. Richard Landis and Gary G. Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* 33.1 (1977), pp. 159–174. ISSN: 0006341X.

[LM20] Duri Long and Brian Magerko. “What is AI literacy? Competencies and design considerations”. In: *Proc. of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–16.

[LNN18] Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. “Dave the debater: a retrieval-based and generative argumentative dialogue agent”. In: 2018, pp. 121–130.

[Lol+12] Frank Loll et al. “How tough should it be? Simplifying the development of argumentation systems using a configurable platform”. In: *Educational Technologies for Teaching Argumentation Skills//Pinkwart, N., McLaren, B.(eds). Bentham Science Publishers, Sharjah, United Arab Emirates* (2012), pp. 169–197.

[LPS22] Allison Littlejohn and Viktoria Pammer-Schindler. “Technologies for Professional Learning”. In: *Research Approaches on Workplace Learning: Insights from a Growing Field*. Ed. by Christian Harteis, David Gijbels, and Eva Kyndt. Springer, 2022, pp. 321–346.

[LT16] Marco Lippi and Paolo Torroni. “Argumentation mining: State of the art and emerging trends”. In: *ACM Transactions on Internet Technology (TOIT)* 16.2 (2016), pp. 1–25. ISSN: 15576051. DOI: 10.1145/2850417.

[Lyt+19] Anastasios Lytos et al. “The evolution of argumentation mining: From models to social media and emerging tools”. In: *Information Processing & Management* 56.6 (2019), p. 102055.

[Man+23] Amogh Mannekote et al. “Exploring Usability Issues in Instruction-Based and Schema-Based Authoring of Task-Oriented Dialogue Agents”. In: *Proc. of the 5th International Conference on Conversational User Interfaces*. 2023, pp. 1–6.

[Man84] William C Mann. *Discourse Structures for Text Generation.* Tech. rep. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1984.

[Mar+09] Brent Martin et al. “An intelligent tutoring system for medical imaging”. In: *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE). 2009, pp. 502–509.

[MCd19] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. "Survey of conversational agents in health". In: *Expert Systems with Applications* 129 (2019), pp. 56–67. ISSN: 09574174. DOI: 10.1016/j.eswa.2019.03.054. URL: https://doi.org/10.1016/j.eswa.2019.03.054.

[Mik+13] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013).

[Mir+23] Behzad Mirzababaei et al. "Interactive Web-Based Learning Materials Vs. Tutorial Chatbot: Differences in User Experience". In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 213–228.

[Mit03] Antonija Mitrovic. "An intelligent SQL tutor on the web". In: *International Journal of Artificial Intelligence in Education* 13.2-4 (2003), pp. 173–197.

[MM11] Raquel Mochales and Marie-Francine Francine Moens. "Argumentation mining". In: *Artificial Intelligence and Law* 19.1 (2011), pp. 1–22. ISSN: 09248463. DOI: 10.1007/s10506-010-9104-x.

[Moe+07] Marie-Francine Francine Moens et al. "Automatic detection of arguments in legal texts". In: *Proceedings of the 11th international conference on Artificial intelligence and law*. 2007, pp. 225–230. ISBN: 1595936807. DOI: 10.1145/1276318.1276362.

[MPS21a] Behzad Mirzababaei and Viktoria Pammer-Schindler. "Developing a Conversational Agent's Capability to Identify Structural Wrongness in Arguments Based on Toulmin's Model of Arguments". In: *Frontiers in Artificial Intelligence* 4 (2021). ISSN: 2624-8212.

[MPS21b] Behzad Mirzababaei and Viktoria Pammer-Schindler. "Developing a Conversational Agent's Capability to Identify Structural Wrongness in Arguments Based on Toulmin's Model of Arguments". In: *Frontiers in artificial intelligence* 4 (2021).

[MPS22a] Behzad Mirzababaei and Viktoria Pammer-Schindler. "An Educational Conversational Agent for GDPR". In: *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th EC-TEL 2022, September 12–16, 2022, Proceedings*. Springer. 2022, pp. 470–476.

[MPS22b] Behzad Mirzababaei and Viktoria Pammer-Schindler. "Learning to Give a Complete Argument with a Conversational Agent: An Experimental Study in Two Domains of Argumentation". In: *ECTEL*. Springer. 2022, pp. 215–228.

[MS04] Erica Melis and Jörg Siekmann. "Activemath: An intelligent tutoring system for mathematics". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer. 2004, pp. 91–101.

[Nak+18] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. 2018.

[NDO99] Paul Newton, Rosalind Driver, and Jonathan Osborne. "The place of argumentation in the pedagogy of school science". In: *International Journal of science education* 21.5 (1999), pp. 553–576.

[OLB14] Nathan Ong, Diane Litman, and Alexandra Brusilovsky. "Ontology-based argument mining and automatic essay scoring". In: *Proceedings of the First Workshop on Argumentation Mining*. 2014, pp. 24–28.

[Ope23] OpenAI. *ChatGPT*. (Sep 25 version) [Large language model] `https://chat.openai.com/chat`. 2023.

[Ped+11] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[PGQ16] Prakash Poudyal, Teresa Goncalves, and Paulo Quaresma. "Experiments on identification of argumentative sentences". In: *2016 10th SKIMA*. IEEE. 2016, pp. 398–403.

[Pin04] Paul R Pintrich. "A conceptual framework for assessing motivation and self-regulated learning in college students". In: *Educational psychology review* 16.4 (2004), pp. 385–407.

[Pin99] Paul R Pintrich. "The role of motivation in promoting and sustaining self-regulated learning". In: *International journal of educational research* 31.6 (1999), pp. 459–470.

[PKY15] Joonsuk Park, Arzoo Katiyar, and Bishan Yang. "Conditional random fields for identifying appropriate types of support for propositions in online user comments". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. 2015, pp. 39–44.

[PM09] Raquel Mochales Palau and Marie-Francine Moens. "Argumentation mining: the detection, classification and structure of arguments in text". In: *Proc. of the 12th IAAIL*. 2009, pp. 98–107.

[PM21] Diana Pérez-Marín. "A review of the practical applications of pedagogic conversational agents to be used in school and university classrooms". In: *Digital* 1.1 (2021), pp. 18–33.

[PMB13] Diana Pérez-Marín and Antonio Boza. "A procedure to create a pedagogic conversational agent in secondary physics and chemistry education". In: *International Journal of Information and Communication Technology Education* 9.4 (2013), pp. 94–112. ISSN: 15501337. DOI: `10.4018/ijicte.2013100107`.

[PN16] Isaac Persing and Vincent Ng. "End-to-end argumentation mining in student essays". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 1384–1394.

[PN20] Isaac Persing and Vincent Ng. “Unsupervised argumentation mining in student essays”. In: *Proc. of the 12th Language Resources and Evaluation Conference*. 2020, pp. 6795–6803.

[PS15] Andreas Peldszus and Manfred Stede. “Joint prediction in MST-style discourse parsing for argumentation mining”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015, pp. 938–948.

[PTT22] Jan Wira Gotama Putra, Simone Teufel, and Takenobu Tokunaga. “Annotating argumentative structure in English-as-a-Foreign-Language learner essays”. In: *Natural Language Engineering* 28.6 (2022), pp. 797–823.

[Rak+19] Geetanjali Rakshit et al. “Debbie, the debate bot of the future”. In: *Lecture Notes in Electrical Engineering* 510 (2019), pp. 45–52. ISSN: 18761119. arXiv: 1709.03167.

[RCD11] Alan Ritter, Colin Cherry, and Bill Dolan. “Data-driven response generation in social media”. In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2011.

[Red+17] Christine Redecker et al. *European framework for the digital competence of educators: DigCompEdu*. Tech. rep. Joint Research Centre (Seville site), 2017.

[RG07] Chris Reed and Floriana Grasso. “Recent advances in computational models of natural argument”. In: *International Journal of Intelligent Systems* 22.1 (2007), pp. 1–15. ISSN: 08848173. DOI: 10.1002/int.20187.

[RG19] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proc. of the 2019 Conference on EMNLP*. Association for Computational Linguistics, Nov. 2019.

[RG20a] Nils Reimers and Iryna Gurevych. “Making monolingual sentence embeddings multilingual using knowledge distillation”. In: *arXiv preprint arXiv:2004.09813* (2020).

[RG20b] Nils Reimers and Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. In: *Proc. of the 2020 Conference on EMNLP*. Association for Computational Linguistics, Nov. 2020.

[Rin+15] Ruty Rinott et al. “Show me your evidence-an automatic method for context dependent evidence detection”. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2015, pp. 440–450.

[RN02] Stuart Russell and Peter Norvig. “Artificial intelligence: a modern approach”. In: (2002).

[RSY22] Vuorikari R, Kluzer S, and Punie Y. *DigComp 2.2: The Digital Competence Framework for Citizens - With new examples of knowledge, skills and attitudes*. Scientific analysis or review KJ-NA-31006-EN-N (online),KJ-NA-31006-EN-C (print). Luxembourg (Luxembourg), 2022. DOI: 10.2760/115376(online),10.2760/490274(print).

[Rua+19] Sherry Ruan et al. "Quizbot: A dialogue-based adaptive learning system for factual knowledge". In: *Proc. of the 2019 CHI*. 2019, pp. 1–13.

[Rus10] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[RWB12] Niall Rooney, Hui Wang, and Fiona Browne. "Applying Kernel Methods to Argumentation Mining." In: *FLAIRS Conference*. Vol. 172. 2012.

[Sab+13] Kent E. Sabo et al. "Searching for the two sigma advantage: Evaluating algebra intelligent tutors". In: *Computers in Human Behavior* 29.4 (2013), pp. 1833–1840. ISSN: 07475632. DOI: 10.1016/j.chb.2013.03.001. URL: http://dx.doi.org/10.1016/j.chb.2013.03.001.

[Sar+15] Christos Sardianos et al. "Argument extraction from news". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. 2015, pp. 56–66. DOI: 10.3115/v1/w15-0508.

[Sch+18] Claudia Schulz et al. "Multi-task learning for argumentation mining in low-resource settings". In: *arXiv preprint arXiv:1804.04083* (2018).

[SG14] Christian Stab and Iryna Gurevych. "Identifying argumentative discourse structures in persuasive essays". In: *Proc. of the 2014 conference on EMNLP*. 2014, pp. 46–56.

[SG17a] Christian Stab and Iryna Gurevych. "Parsing argumentation structures in persuasive essays". In: *Computational Linguistics* 43.3 (2017), pp. 619–659.

[SG17b] Christian Stab and Iryna Gurevych. "Recognizing insufficiently supported arguments in argumentative essays". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 2017, pp. 980–990.

[SH04] Siriwan Suebnukarn and Peter Haddawy. "A collaborative intelligent tutoring system for medical problem-based learning". In: *Proceedings of the 9th international conference on Intelligent user interfaces*. 2004, pp. 14–21.

[Shi20] Jennifer Shirk. "Designing a self-paced learning experience to support learner self-regulation". In: *Journal of Teaching and Learning with Technology* 9.1 (2020).

[Shn+17] Eyal Shnarch et al. "GrASP: Rich Patterns for Argumentation Mining". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 1345-1350)*. 2017, pp. 1345–1350.

[Sim08] Shirley Simon. “Using Toulmin’s argument pattern in the evaluation of argumentation in school science”. In: *International Journal of Research & Method in Education* 31.3 (2008), pp. 277–289.

[Slo+21] Noam Slonim et al. “An autonomous debating system”. In: *Nature* 591.7850 (2021), pp. 379–384.

[SM23] Michael David Sankey and Stephen James Marshall. “Perspective Chapter: The Learning Management System of 2028 and How we Start Planning for this Now”. In: *Higher Education-Reflections From the Field*. IntechOpen, 2023.

[Spe+01] Marcus Specht et al. “Authoring adaptive educational hypermedia in WINDS”. In: *Proc. of ABIS2001, Dortmund, Germany* 3.3 (2001), pp. 1–8.

[ST08] Steven E. Stemler and Jessica Tsai. “Best practices in interrater reliability: Three common approaches”. In: *Best practices in quantitative methods* (2008), pp. 29–49.

[TCL21] Darren Turnbull, Ritesh Chugh, and Jo Luck. “Issues in learning management systems implementation: A comparison of research perspectives between Australia and China”. In: *Education and Information Technologies* 26.4 (2021), pp. 3789–3810.

[TG17] Daniel Toniuc and Adrian Groza. “Climebot: An argumentative agent for climate change”. In: *Proc. of the ICCP 2017*. Vol. 2. IEEE. 2017, pp. 63–70. ISBN: 9781538633687.

[Tou03] Stephen E. Toulmin. *The uses of argument*. Cambridge university press, 2003.

[VA+08] Claudia Von Aufschnaiter et al. “Arguing to learn and learning to argue: Case studies of how students’ argumentation relates to their scientific knowledge”. In: *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 45.1 (2008), pp. 101–131.

[Van+02] Kurt VanLehn et al. “The architecture of Why2-Atlas: A coach for qualitative physics essay writing”. In: *International Conference on Intelligent Tutoring Systems*. Springer. 2002, pp. 158–167.

[Vas+17] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[VD77] Teun Adrianus Van Dijk. “Text and context: Explorations in the semantics and pragmatics of discourse”. In: (1977).

[VGK19] Frans H. Van Eemeren, Rob Grootendorst, and Tjark Kruiger. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*. Vol. 7. De Gruyter Mouton, 2019.

[VR14] George Veletsianos and Gregory S. Russell. “Pedagogical agents”. In: *Handbook of research on educational communications and technology*. Springer, 2014, pp. 759–769. ISBN: 9781461431855. DOI: 10.1007/978-1-4614-3185-5_61.

[VRKP22] Riina Vuorikari Rina, Stefano Kluzer, and Yves Punie. *DigComp 2.2: The Digital Competence Framework for Citizens-With new examples of knowledge, skills and attitudes*. Tech. rep. Joint Research Centre (Seville site), 2022.

[Wac+17] Henning Wachsmuth et al. “Argumentation quality assessment: Theory vs. Practice”. In: *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2 (2017), pp. 250–255. DOI: 10.18653/v1/P17-2039.

[Wal95a] R Wallace. “Alice-artificial linguistic internet computer entity-the ALICE AI. foundation”. In: *Disponıvel em http://www. alicebot. org. Acesso em* 18 (1995).

[Wal95b] Richard Wallace. “Artificial linguistic internet computer entity (alice)”. In: *City* (1995).

[Wam+20a] Thiemo Wambsganss et al. “A corpus for argumentative writing support in German”. In: *arXiv preprint arXiv:2010.13674* (2020).

[Wam+20b] Thiemo Wambsganss et al. “AL: An adaptive learning support system for argumentation skills”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–14.

[Wam+20c] Thiemo Wambsganss et al. “AL: An Adaptive Learning Support System for Argumentation Skills”. In: *Proceedings of the 2020 CHI conference on human factors in computing systems* 20 (2020), pp. 1–14. DOI: 10.1145/3313831.3376732. URL: http://dx.doi.org/10.1145/3313831.3376732.

[Wam+21] Thiemo Wambsganss et al. “ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills”. In: *Proc. of the CHI*. 2021, pp. 1–13. ISBN: 9781450380966.

[Wam+22] Thiemo Wambsganss et al. “Improving students argumentation learning with adaptive self-evaluation nudging”. In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–31.

[Wan+15] Dongqing Wang et al. “A problem solving oriented intelligent tutoring system to improve students’ acquisition of basic computer skills”. In: *Computers & Education* 81 (2015), pp. 102–112.

[Wan+20] Wenting Wang et al. “ArguLens: Anatomy of Community Opinions On Usability Issues Using Argumentation Models”. In: *In Proc. of the ACM 2020* 20 (2020), pp. 1–14.

[Web+21] Florian Weber et al. "Pedagogical agents for interactive learning: A taxonomy of conversational agents in education". In: *Forty-Second International Conference on Information Systems. Austin, Texas.* 2021, pp. 1–17.

[Wei66] Joseph Weizenbaum. "ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine". In: *Commun. ACM* 9.1 (1966), 36–45. ISSN: 0001-0782. DOI: 10.1145/365153.365168. URL: https://doi.org/10.1145/365153.365168.

[WF06] Armin Weinberger and Frank Fischer. "A framework to analyze argumentative knowledge construction in computer-supported collaborative learning". In: *Computers and Education* 46.1 (2006), pp. 71–95. ISSN: 03601315. DOI: 10.1016/j.compedu.2005.04.003.

[Wha97] Richard Whately. *Elements of logic.* Longman, Green, Longman, Roberts and Green, 1897.

[WJL22] Thiemo Wambsganss, Andreas Janson, and Jan Marco Leimeister. "Enhancing argumentative writing with automated feedback and social comparison nudging". In: *Computers & Education* 191 (2022), p. 104644.

[WM11] Amali Weerasinghe and Antonija Mitrovic. "Facilitating adaptive tutorial dialogues in EER-tutor". In: *International Conference on Artificial Intelligence in Education.* Springer. 2011, pp. 630–631.

[Wol+21] Sebastian Wollny et al. "Are we there yet?-a systematic literature review on chatbots in education". In: *Frontiers in artificial intelligence* 4 (2021), p. 654924.

[Wol+22] Irmtraud Wolfbauer et al. "A Script for Conversational Reflection Guidance: A Field Study on Developing Reflection Competence with Apprentices". In: *IEEE Transactions on Learning Technologies* 15.5 (2022).

[WPSR20] Irmtraud Wolfbauer, Viktoria Pammer-Schindler, and C. Rosé. "Rebo Junior: Analysis of Dialogue Structure Quality for a Reflection Guidance Chatbot". In: *Proc. of the EC-TEL 15th.* 2020, pp. 14–18.

[WRM08] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation schemes.* Cambridge University Press, 2008.

[WSA17] Henning Wachsmuth, Benno Stein, and Yamen Ajjour. *PageRank" for argument relevance.* Tech. rep. 2017, pp. 1117–1127. DOI: 10.18653/v1/e17-1105. URL: www.bbc.co.uk/ethics/capitalpunishment/against_1.shtml..

[Xia+22] Meng Xia et al. "Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–30.

[Zha+16] Guangxuan Zhang et al. "Argument detection in online discussion: A theory based approach". In: *22nd Americas Conference on Information Systems: Surfing the IT Innovation Wave, AMCIS 2016*. Association for Information Systems. 2016.

[Zha+19] Yizhe Zhang et al. "Dialogpt: Large-scale generative pre-training for conversational response generation". In: *arXiv preprint arXiv:1911.00536* (2019).

[Zha+23] Shuang Zhang et al. "Applications of transformer-based language models in bioinformatics: a survey". In: *Bioinformatics Advances* 3.1 (2023), vbad001.

[ZJ17] Bilan Zhang and Jiyou Jia. "Evaluating an intelligent tutoring system for personalized math teaching". In: *2017 International Symposium on Educational Technology (ISET)*. IEEE. 2017, pp. 126–130.

A. Appendix

A.1. Chapter 4: The GDPR Agent's Dialogue Structure

In this appendix section, all the GDPR Agent's messages and all the possible responses to users' answers are listed. Each dialogue turn (DT) starts with the agent's message and then all possible user's responses (U) with related agent's feedback (F) are shown.

DT1: Hello! I'm GDPR Agent. I'm here to talk to you about the General Data Protection Regulation (GDPR). Are you ready to learn about it?

- U: Yes
  - F: Great! Now, let's go ...
- U: No
  - F: It's OK. We can talk later :)

DT2: First, let's deal with some of the basics of the GDPR. Here, we are going to talk about: What is the GDPR and its scope? Personal data V.S. Sensitive data, Seven data protection principles. The information was collected from FUTURE.

Do you know, what the scope of GDPR is? By this I mean: To whom and what types of activities does it apply? What is regulated by the GDPR?

- U: I don't know.
  - F: Don't worry! We are here to learn.
- U: _*anything else*_
  - F: GDPR contains rules concerning the protection of natural persons (in other words, individuals) when their personal data are processed and rules on the free movement of personal data, as stressed in Article 1(1). Article 1(2) provides that the GDPR seeks to protect the fundamental rights and freedoms of natural persons and, more specifically, their right to the protection of personal data. It means that, as such, the regulation does not deal with the rights and freedoms of legal persons, such as companies.

DT3: Regarding its scope, the GDPR rules are applied to any data processing activities conducted by organisations established in the EU. In addition, it also applies to any business activities relating to monitoring or offering of goods (or services) to data subjects situated in the EU.

Now, the question is: What is personal data? Can you try to give a definition?

U: _*anything*_

F: Based on GDPR, personal data can be considered as any piece of information relating to a person by which we can identify the person, simple as that! The identification of people could be based on just a piece of information (directly) or a combination of other information (indirectly). The data can appear in various forms such as names, emails, location data, home addresses and IP addresses.

Each of these on their own does not necessarily classify as personal data, because they don't clearly identify an individual. For example, it's very unlikely that nobody else in the world has the same name as you and as such, you are not easily identifiable by your name alone. However, when this is combined with your email or home address, this information is sufficient to clearly identify you as an individual.

DT4: The GDPR differentiates between personal data and sensitive personal data. What kind of data is sensitive?

U: _*anything*_

F: Sensitive personal data is a specific set of "special categories" that must be treated with extra security. This includes information pertaining to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, and biometric data (where processed to uniquely identify someone).

The difference between personal data and sensitive personal data is that processing sensitive personal data requires additional protection granted by the GDPR since processing those types of data can involve severe and unacceptable risks to fundamental human rights and freedoms.

DT5: Now it's time to answer three questions about personal and sensitive data (*).

Which one of these is personal data when considered separately from other possible additional data? (Please select your answer and press "Send Message")

1. Your first and last names are written on a piece of paper
2. A location of your suitcase
3. Age and gender of a dead person
4. A brand of your car

U: Answer 1

F: Good job :) Your first and last names can be used to establish who you are as an identified or identifiable natural person.

U: Answer 2

F: You're on the right track, but not there yet. While the location of your suitcase, just as any other object you use, can in theory be used to determine who you are as an identified or identifiable natural person, as such this location does not constitute personal data.

The correct answer is your first and last names. They can be used to establish who you are as an identified or identifiable natural person.

U: Answer 3

F: Information about dead persons does not fall within the scope of personal data.

The first is the correct answer. Your first and last names can be used to establish who you are as an identified or identifiable natural person.

U: Answer 4

F: _*goto DT5.1*_

DT5.1 Why do you think the brand of your car is personal data? If you want to change your answer, please write "CHANGE".

U: CHANGE

F: _*goto DT5*_

U: _*anything else*_

F: _*goto DT5.2*_

DT5.2 Let's assume I know someone whose car's brand is Mazda. Could you identify him or her by knowing just the brand?

U: Yes

F: There are millions of people whose car is a Mazda. It's not possible to identify a person just by knowing his or her brand's car. Information such as the brand of the car you drive cannot be considered as some information that relates to you as an identified or identifiable natural person.

The first is the correct answer. Your first and last names can be used to establish who you are as an identified or identifiable natural person.

U: _*anything else*_

F: As such, information such as the brand of the car you drive cannot be considered as data that relates to you as an identified or identifiable natural person.

The first is the correct answer. Your first and last names can be used to establish who you are as an identified or identifiable natural person.

DT6 Which of these is the processing of personal data? (Please select your answer and press "Send Message")

1. Remembering from a piece of paper the names and addresses of people who participated in a local food festival

2. Deleting from a computer a PDF document with sick leaves and employee numbers by an HR department of a company

3. Creating on a computer a Word file with job titles from a university

4. Restructuring an Excel sheet with genders of 5.000 employees working for a bank

U: Answer 1

F: Merely remembering something that appears to be personal data cannot be said to constitute the actual processing of these data.

The correct answer is the second one. Sick leave and employee numbers can be considered personal data and deleting them is one of the operations that falls under the scope of the processing of personal data.

U: Answer 2

F: Great!!! Sick leave and employee numbers can be considered personal data and deleting them is one of the operations that falls under the scope of the processing of personal data.

U: Answer 3

F: While there are not too many job titles at the universities, they cannot be regarded as personal data that can be used to identify certain individuals and therefore there is no processing of these data.

The correct answer is the second one. Sick leave and employee numbers can be considered personal data and deleting them is one of the operations that falls under the scope of the processing of personal data.

U: Answer 4

F: An Excel sheet with genders of 5000 employees does not contain personal data as such given that specific individuals cannot be identified on the basis of this information and therefore there is no processing of personal data.

The correct answer is the second one. Sick leave and employee numbers can be considered personal data and deleting them is one of the operations that falls under the scope of the processing of personal data.

DT7 Which of the following can be concluded to belong to special categories of (sensitive) personal data when considered separately from other possible additional data? (Please select your answer and press "Send Message")

1. Your initials

2. The name of the university you studied at

3. Your PC's IP address

4. An audio recording of your voice

U: Answer 1

F: This information can be seen as personal data but it does not belong to special categories of personal data that concern racial or ethnical origin, political opinions, health and other aspects.

The correct answer is the last one. Audio recordings of voices are considered biometric data relating to physical, physiological or behavioural characteristics of individuals that can be used to uniquely identify them. Other examples of biometric data are, for instance, fingerprints, facial images and iris scans.

U: Answer 2

F: This information cannot be considered as information relating to you as an identified or identifiable natural person when other data are not used to specifically determine your identity. As such, it cannot be seen as belonging to special categories of personal data that concern racial or ethnical origin, political opinions, health and other aspects.

The correct answer is the last one. Audio recordings of voices are considered biometric data relating to physical, physiological or behavioural characteristics of individuals that can be used to uniquely identify them. Other examples of biometric data are, for instance, fingerprints, facial images and iris scans.

U: Answer 3

F: In certain circumstances, IP addresses can be considered personal data but they cannot be said to fit into special categories of personal data that concern racial or ethnical origin, political opinions, health and other aspects.

The correct answer is the last one. Audio recordings of voices are considered biometric data relating to physical, physiological or behavioural characteristics of individuals that can be used to uniquely identify them. Other examples of biometric data are, for instance, fingerprints, facial images and iris scans.

U: Answer 4

F: Exactly!!! Audio recordings of voices are considered biometric data relating to physical, physiological or behavioural characteristics of individuals that can be used to uniquely identify them. Other examples of biometric data are, for instance, fingerprints, facial images and iris scans.

DT8 Which of these are personal data when considered separately from other possible additional data? (Please select your answer and press "Send Message")

1. Name of your dog

2. Photos of you on Facebook
3. Music you listen to
4. The fact that you frequently visit the gym

U: Answer 1

F: The names of your pets cannot be regarded as information relating to you as an identified or identifiable natural person.

The correct answer is the second option. Photos, videos and audio recordings you post on the social media platforms are personal data that specifically relate to you as an identified or identifiable natural person.

U: Answer 2

F: Perfect! Photos, videos and audio recordings you post on the social media platforms are personal data that specifically relate to you as an identified or identifiable natural person.

U: Answer 3

F: As typically many people share your musical preferences, this information isn't considered as relating only to you as an identified or identifiable natural person.

The correct answer is "your photos on Facebook". Photos, videos and audio recordings you post on the social media platforms are personal data that specifically relate to you as an identified or identifiable natural person.

U: Answer 4

F: The fact that you are a frequent gym visitor does not make it possible to identify you as a natural person.

The answer is your photos on Facebook. Photos, videos and audio recordings you post on the social media platforms are personal data that specifically relate to you as an identified or identifiable natural person.

DT9 A shopkeeper wants to register how many visitors enter his shop every day. A system detects the MAC address of each visitor's smartphone. It is impossible for the shopkeeper to identify the owner of the phone from this signal, but telephone providers can link the MAC address to the owner of the phone. According to the GDPR, is the shopkeeper allowed to use this method? (ref:Exin.com) (Please select your answer and press "Send Message")

1. Yes, because the shopkeeper cannot identify the owner of the telephone.
2. Yes, because the visitor has automatically consented by connecting to the Wi-Fi.
3. No, because the telephone's MAC address must be regarded as personal data.
4. No, because the telephone providers are the owners of the MAC addresses.

U: Answer 1

F: Unfortunately, this isn't the correct answer. The issue is not whether the shopkeeper can identify the visitor, but that it is technically possible to do so.

The correct answer is the third one. The phone's signal is a unique code that can be linked to the owner of the phone. The data must be regarded as personal data because it is technically possible to identify the visitor. (Chapter 3, Article 26, Article 30)

U: Answer 2

F: This is a common misunderstanding. Consent must be an active, informed and free act of agreement to the processing. To see a MAC address, the visitor does not need to be logged onto the Wi-Fi.

The correct answer is the third one. The phone's signal is a unique code that can be linked to the owner of the phone. The data must be regarded as personal data because it is technically possible to identify the visitor. (Chapter 3, Article 26, Article 30)

U: Answer 3

F: Excellent! The phone's signal is a unique code that can be linked to the owner of the phone. The data must be regarded as personal data because it is technically possible to identify the visitor (Chapter 3, Article 26, Article 30)

U: Answer4

F: Unfortunately, this isn't the correct answer. The shopkeeper is not allowed to keep the data or process it because it must be regarded as personal data. The telephone provider is not the owner of the MAC address, nor is the telephone provider protected by the GDPR.

The correct answer is the third one. The phone's signal is a unique code that can be linked to the owner of the phone. The data must be regarded as personal data because it is technically possible to identify the visitor. (Chapter 3, Article 26, Article 30)

DT10 I have a figure that summarizes all that we have learned so far about the scope of GDPR and personal vs. sensitive data. Take a minute to review it (Figure A.1).

Please let me know when you want to continue and talk about the GDPR's principles.

U: _*anything*_

F: Let's talk about the GDPR principles

DT11 We are now moving on from the scope of GDPR to seven data protection principles but before that: Do you know a few of them already and can you name them?

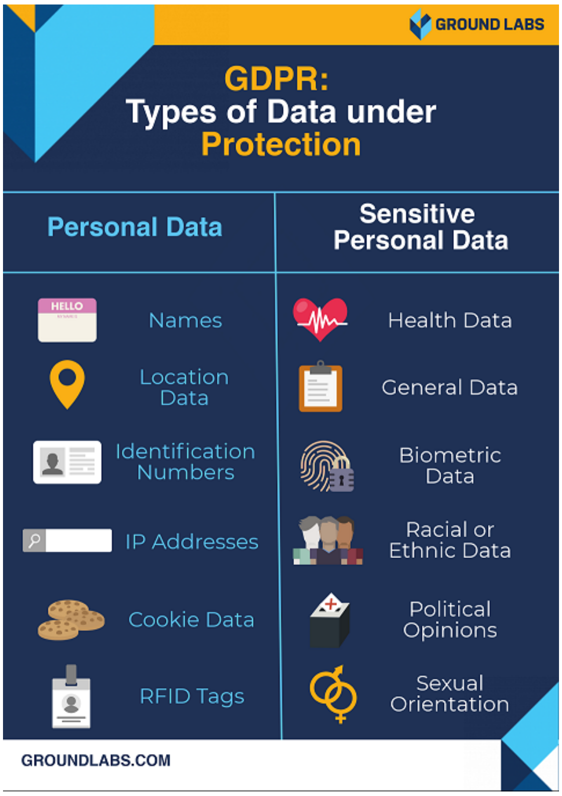


Figure A.1.: The summary of personal and sensitive data

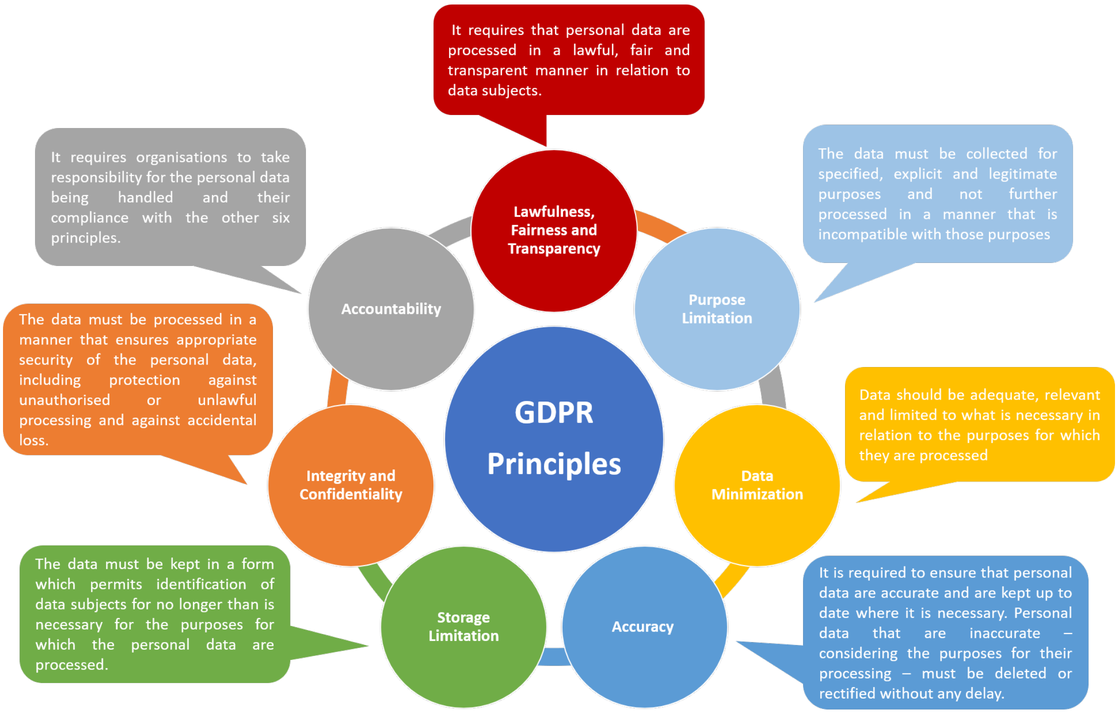


Figure A.2.: The seven GDPR's principles

U: _*mentioned 3 or less*_

F: Good job! I've summarised all GDPR principles in this figure.

U: _*mentioned more than 3*_

F: Perfect! Now let's deal with them. I've summarised all GDPR principles in this figure.

U: I don't know.

F: It's OK! We are here to learn. I've summarised all GDPR principles in this figure.

U: _*anything else*_

F: I've summarised all GDPR principles in this figure.

DT12 I summarized all GDPR principles in this figure. Please take a look at the seven GDPR principles and let me know when you finish.

Now I have prepared three questions about how these principles can be violated. Now, let's go for the questions.

Please, indicate which of the seven data protection principles have been violated in the following scenario. A travel agency secretly processes the personal data of individuals in the form of their names and home addresses. It appears that the actual data subjects whose data are being processed are not even aware of this. (Please select your answer and press "Send Message")

1. Data minimisation
2. Integrity and confidentiality
3. Lawfulness, fairness and transparency
4. Purpose limitation

U: Answer 1

F: Unfortunately, this scenario is not about data minimisation. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

This scenario deals with the lack of transparency in the processing of personal data. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

U: Answer 2

F: Unfortunately, this scenario does not concern integrity and confidentiality. This principle means that in the processing of personal data, appropriate security of personal data is ensured.

This scenario deals with the lack of transparency in the processing of personal data. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

U: Answer 3

F: Excellent! This scenario deals with the lack of transparency in the processing of personal data. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

U: Answer 4

F: Unfortunately, this scenario is not about purposes of data processing and does not concern the principle of purpose limitation. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

This scenario deals with the lack of transparency in the processing of personal data. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

DT13 Please, indicate which of the seven data protection principles have been violated in the following scenario. A bank has collected and stored the personal data of its clients in order to open their bank accounts. Then these data were used to send these individuals information about mortgages offered by the bank. (Please select your answer and press "Send Message")

1. Accuracy
2. Lawfulness, fairness and transparency
3. Purpose limitation
4. Storage limitation

U: Answer 1

F: Unfortunately, this scenario does not concern accuracy. This principle means that personal data are accurate and are kept up to date where it is necessary.

This scenario is based on the violation of the principle of purpose limitation given that personal data had been processed for one purpose and then processed again for a completely different purpose. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

U: Answer 2

F: Unfortunately, this scenario is not about lawfulness, fairness and transparency. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

This scenario is based on the violation of the principle of purpose limitation given that personal data had been processed for one purpose and then processed again for a completely different purpose. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

U: Answer 3

F: Fantastic! This scenario is based on the violation of the principle of purpose limitation given that personal data had been processed for one purpose and then processed again for a completely different purpose. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

U: Answer 4

F: Unfortunately, this scenario does not deal with storage limitations. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

This scenario is based on the violation of the principle of purpose limitation given that personal data had been processed for one purpose and then processed again for a completely different purpose. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

DT14 Please, indicate which of the seven data protection principles have been violated in the following scenario. A person wants to become a member of an archery club. The membership form requires him to provide not only his name and home address but also his social security number and political beliefs that will be stored in a digital database of the club. (Please select your answer and press "Send Message")

1. Data minimisation
2. Integrity and confidentiality
3. Lawfulness, fairness and transparency
4. Storage limitation

U: Answer 1

F: Good job! This scenario concerns the data minimisation principle because some personal data that are collected, such as social security numbers and political beliefs, are not relevant and are not limited to what is necessary for the purpose of their processing. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

U: Answer 2

F: I hate to say it, but that's the wrong answer. The integrity and confidentiality principle means that in the processing of personal data, appropriate security of personal data is ensured.

This scenario concerns the data minimisation principle because some personal data that are collected, such as social security numbers and political beliefs, are not relevant and are not limited to what is necessary for the purpose of their processing. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

U: Answer 3

F: I hate to say it, but that's the wrong answer. The lawfulness, fairness and transparency principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

This scenario concerns the data minimisation principle because some personal data that are collected, such as social security numbers and political beliefs, are not relevant and are not limited to what is necessary for the purpose of their processing. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

U: Answer 4

F: I hate to say it, but that's the wrong answer. The storage limitation principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

This scenario concerns the data minimisation principle because some personal data that are collected, such as social security numbers and political beliefs, are not relevant and are not limited to what is necessary for the purpose of their processing. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

DT15 U: Answer 1

F: Awesome! This scenario is about the violation of the principle of accuracy due to the fact that information about companies is not accurate anymore and is not updated by the municipality. This principle means that personal data are accurate and kept up to date where necessary.

U: Answer 2

F: Unfortunately, this scenario is not about integrity and confidentiality. This principle means that in the processing of personal data, appropriate security of personal data is ensured.

This scenario is about the violation of the principle of accuracy due to the fact that information about companies is not accurate anymore and is not updated by the municipality. This principle means that personal data are accurate and kept up to date where necessary.

U: Answer 3

F: Unfortunately, this scenario is not about purpose limitation. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

This scenario is about the violation of the principle of accuracy due to the fact that information about companies is not accurate anymore and is not updated by the municipality. This principle means that personal data are accurate and kept up to date where necessary.

U: Answer 4

F: Unfortunately, this scenario does not deal with storage limitations. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

This scenario is about the violation of the principle of accuracy due to the fact that information about companies is not accurate anymore and is not updated by the municipality. This principle means that personal data are accurate and kept up to date where necessary.

DT16 Please, indicate which of the seven data protection principles have been violated in the following scenario. For the purposes of holding a sports event, an organisation collects, stores and processes the personal data of those who participate in this event. After the event is finished, these data are kept in the system for years. (Please select your answer and press "Send Message")

1. Storage limitation
2. Lawfulness, fairness and transparency
3. Data minimisation
4. Accuracy

U: Answer 1

F: Great! This scenario concerns the principle of storage limitation that has been violated. Personal data are stored in the system for many years while it is no longer necessary for the purposes of the processing. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

U: Answer 2

F: This scenario is not about lawfulness, fairness and transparency. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

This scenario concerns the principle of storage limitation that has been violated. Personal data are stored in the system for many years while it is no longer necessary for the purposes of the processing. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

U: Answer 3

F: It's not the correct answer. This scenario is not about data minimisation. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

This scenario concerns the principle of storage limitation that has been violated. Personal data are stored in the system for many years while it is no longer necessary for the purposes of the processing. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

U: Answer 4

F: Unfortunately, it's not the correct answer. This scenario is about the violation of the principle of accuracy due to the fact that information about companies is not accurate anymore and is not updated by the municipality. This principle means that personal data are accurate and kept up to date where necessary.

This scenario concerns the principle of storage limitation that has been violated. Personal data are stored in the system for many years while it is no longer necessary for the purposes of the processing. This principle means that personal data must be kept in a form that makes it possible to identify data subjects for no longer than is necessary for the purposes of the processing.

DT17 Please, indicate which of the seven data protection principles have been violated in the following scenario. The staff intranet of a university holds much personal data of its employees. Unfortunately, the server does not rely on the use of digital security mechanisms to protect this information and any cyber-attack can cause significant damage (Please select your answer and press "Send Message.")

1. Data minimisation
2. Integrity and confidentiality
3. Lawfulness, fairness and transparency
4. Purpose limitation

U: Answer 1

F: Unfortunately, it's not the correct answer. This scenario is not about purpose limitation. This principle means that personal data must be collected only for specified, explicit and legitimate purposes and it is not allowed to process them further in a way that is not compatible with those purposes.

This scenario describes the violation of the principle of integrity and confidentiality because the system holding personal data is not sufficiently protected and a possible data breach can take place. This principle namely means that in the processing of personal data, appropriate security of personal data is ensured.

U: Answer 2

F: Unfortunately, this scenario is not about lawfulness, fairness and transparency. This principle means that personal data are processed in a lawful, fair and transparent manner in relation to data subjects.

This scenario describes the violation of the principle of integrity and confidentiality because the system holding personal data is not sufficiently protected and a possible data breach can take place. This principle means that in the processing of personal data, appropriate security of personal data is ensured.

U: Answer 3

F: Perfecto! This scenario describes the violation of the principle of integrity and confidentiality because the system holding personal data is not sufficiently protected and a possible data breach can take place. This principle means that in the processing of personal data, appropriate security of personal data is ensured.

U: Answer 4

F: Unfortunately, it's not the correct answer. This scenario is not about data minimisation. This principle means that personal data must be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

This scenario describes the violation of the principle of integrity and confidentiality because the system holding personal data is not sufficiently protected and a possible data breach can take place. This principle namely means that in the processing of personal data, appropriate security of personal data is ensured.

DT18 You did it! You have now worked through the basics of GDPR, its scope, its fundamental understanding of what personal and sensitive data is, and its seven principles "lawfulness, fairness and transparency", "purpose limitation", "data minimization", "accuracy", "storage limitation", "integrity and confidentiality", and "accountability".

This could be the basis for discussing technologies that allow for preserving people's privacy whilst still allowing them to benefit from data analytics. These are very novel technologies, right now at research. I'm currently preparing to teach about this as well, do check in with me!

A.2. Chapter 4: The PETs Agent's Dialogue Structure

In this appendix section, all the PETs Agent's messages and all the possible responses to users' answers are listed. Each dialogue turn (DT) starts with the agent's message and then all possible user's responses (U) with related agent's feedback (F) are shown.

DT1: Hello! I'm PETs Agent. I'm here to talk to you about the privacy-enhancing technologies. Are you ready to learn about it?

U: Yes

F: Great! Now, let's go ...

U: No

F: OK, no problem; I currently am not able to discuss anything else - if you are interested in talking to me any time in the future, you know how to find me ;) Bye!

DT2: Our conversation has two main phases. The first one is a teaching phase in which we discuss privacy-enhancing technologies (PETs). In the second phase, we talk about which technologies can be applied to your potential use-cases.

If you are already familiar with PETs, you can skip the introductory part of the course (Teaching phase). How would you like to continue?

1. Teaching phase
2. Use-case phase

U: Answer 1

F: So let's start with what these technologies are and why they are important for businesses.

U: Answer 2

F: *_goto DT16_*

DT3: Before we continue, I would like to know your opinion about the PETs. So, could you briefly tell me what you already know about it?

And don't worry if you know nothing because we are here to learn :)

U: I don't know.

F: That's okay, I'm going to tell you.

U: *_found keywords about PETs_*

F: Great! So you know what we're going to talk about.

U: *_anything_*

F: Let me tell you.

DT4: PETs could be any software or hardware which we use to keep the data safe. With leverage PETs, we can use or share our data in order to extract information without revealing our personal or sensitive data which may be in our data.

As for its importance for businesses, they need to protect their data to stick to the GDPR in some cases, otherwise, they have to pay serious fines due to data breaches. Besides the GDPR, the companies may need to protect their own personal data when they share the data with a third-party partner to do some analysis.

Speaking of GDPR, do you know what the GDPR is? What is the definition of personal and sensitive data based on the GDPR?

U: I don't know.

F: That's okay, I'm going to tell you :)

U: *_found relevant keywords_*

F: WOW! You already know. But let me briefly explain it.

U: _*anything*_

F:

DT5: Here you can see the brief definitions of the GDPR, personal data and sensitive data. You can find the definitions by clicking on the titles. Please write "Done" when you read the explanations to continue our conversation.

The GDPR stands for the General Data Protection Regulation. It contains rules concerning the protection of natural persons (in other words, individuals) when their personal data are processed and rules on the free movement of personal data.

Based on the GDPR, personal data can be considered as any piece of information relating to a person by which we can identify the person, simple as that! The identification of people could be based on just a piece of information (directly) or a combination of other information (indirectly). The data can appear in various forms such as names, emails, location data, home addresses, and IP addresses.

Sensitive personal data is a specific set of “special categories” that must be treated with extra security. This includes information pertaining to: racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, and biometric data (where processed to uniquely identify someone).

U: Done

F: OK, let’s continue.

U: _*anything else*_

F: I consider it as done :)

DT6 Do you think besides enhancing privacy for your data it’s possible to protect the intellectual property of your algorithms or evaluation as well with PETs?

U: _*anything*_

F: That’s okay, I’m going to tell you :)

PETs can be categorized into two groups.

1. Crypto-based methods: It refers to methods which use encryption to enhance privacy. These methods are more secure but often slower than other methods. Final speed performance very much depends on actual use-case and security guarantees.

2. ML-based methods: The methods which focus on machine learning techniques to enhance privacy, such as federated learning. Some models offer privacy and data protection, but usually, the security is enhanced when ML models are combined with differential privacy or crypto methods. On the other hand, pure ML models usually perform much better.

In general, there are various aspects which we need to consider to decide which method we want to apply, such as what we want to protect (our data or our algorithms), the number of involved parties, how urgent do we need results and do we want to outsource anything or not?

DT7 Before I list the methods, could you name some of them? I already mentioned two of them :)

U: I don't know.

F: That's okay, I'm going to tell you :)

U: _*found relevant keywords*_

F: Good job!

U: _*anything else*_

F: _*goto DT8*_

DT8 Here you can see the most common technologies or methods of PETs. You can find the explanations by clicking on the titles. Please write "Done" when you read the explanations to continue our conversation.

FD: The basic idea of FD is training a model on a server/cloud with data from different clients, but not transmitting any actual data, only parameters of the model. A typical situation in which we can apply FL is the participation of many clients who want to keep their data local e.g. input from mobile phones, different measuring devices, etc.

DP: The basic idea of DP is adding noise to the data to enhance security. Ideally, if you evaluate an ML model on a data set, the final result does not change if some parts of the data (i.e. containing my personal information) is left out or changed in the original data. Can be applied in many different scenarios, and works better for larger data sets (Works whenever data can be blurred.)

HE: The basic idea of HE is doing calculations and analysing encrypted data. Typically, one party sends encrypted data to a 2nd party where this data is evaluated without the 2nd party being able to learn anything about this data. A common situation in which we can apply HE is when we want to outsource computation and analysis of data without giving up any privacy, typically one client and data evaluation on the cloud/server.

MPC: It is a cryptographic method, where several parties (at least two) can compute a common function on their data without learning anything about the data from the other participants. Quite a complex approach with different possibilities, one way is that you create so-called secret shares (pieces of your data) that are transmitted to the others (they don't learn anything from these pieces) and all together compute a function/ML model with their received pieces. Efficiency and speed depend on security guarantees. More secure than FD, but slower.

U: Done

F: OK, let's go. So far, we talked about GDPR and some privacy-preserving methods. The next step is checking some imaginary use cases to find out which privacy-preserving methods could be applied.

U: _*anything else*_

F: I consider it as done :) So far, we talked about GDPR and some privacy-preserving methods. The next step is checking some imaginary use cases to find out which privacy-preserving methods could be applied.

DT9 Someone wants to outsource computation and evaluate sensitive and private data on a remote model, e.g. by sending it to an ML model on a cloud. What method can be applied here? (Please select your answer and press "Send Message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption
4. Differential privacy

U: Answer 1

F: You're on the right track, but not there yet. With MPC all participants have to perform local computations and there is a lot of communication necessary between all parties.

The correct answer is HE. It allows you to outsource your computation while keeping your data private, under the assumption that the remote device can handle encrypted data. With MPC and FL usually all participants need to perform local computations, DP helps to protect your data set but does not encrypt it (also not the result of the evaluation). With HE you encrypt your data, send it to the device where the calculation takes place and receive the result – your data and the calculation result never get revealed to someone else. Sometimes also combinations of methods are necessary, one concrete example is COVID HEATMAP, where data from two parties is evaluated without revealing the content to each other based on HE, DP and MPC.

U: Answer 2

F: I'm afraid FL isn't the correct answer. With FL you do the computation locally on your device, you bring the algorithm to your data and do not outsource the computation directly.

The correct answer is HE. It allows you to outsource your computation while keeping your data private, under the assumption that the remote device can handle encrypted data. With MPC and FL usually all participants need to perform local computations, DP helps to protect your data set but does not encrypt it (also not the result of the evaluation). With HE you encrypt your data, send it to the device where the calculation takes place and receive the result – your data and the calculation result never get revealed to someone else. Sometimes also combinations of methods are necessary, one concrete example is COVID HEATMAP, where data from two parties is evaluated without revealing the content to each other based on HE, DP and MPC.

U: Answer 3

F: Good job! HE allows you to outsource your computation while keeping your data private, under the assumption that the remote device can handle encrypted data. With MPC and FL usually all participants need to perform local computations, DP helps to protect your data set but does not encrypt it (also not the result of the evaluation). With HE you encrypt your data, send it to the device where the calculation takes place and receive the result – your data and the calculation result never get revealed to someone else. Sometimes also combinations of methods are necessary, one concrete example is COVID HEATMAP, where data from two parties is evaluated without revealing the content to each other based on HE, DP and MPC.

U: Answer 4

F: Unfortunately not. With DP you add noise to your data to obscure some of the content, so it enhances your privacy and in principle could also be applied to this scenario, but you have stronger security guarantees with HE and also protect the outcome of your remote evaluation.

The correct answer is HE. It allows you to outsource your computation while keeping your data private, under the assumption that the remote device can handle encrypted data. With MPC and FL usually all participants need to perform local computations, DP helps to protect your data set but does not encrypt it (also not the result of the evaluation). With HE you encrypt your data, send it to the device where the calculation takes place and receive the result – your data and the calculation result never get revealed to someone else. Sometimes also combinations of methods are necessary, one concrete example is COVID HEATMAP, where data from two parties is evaluated without revealing the content to each other based on HE, DP and MPC.

DT10 Suppose there are two companies each having a private data set and it would be beneficial to them to find common entries in these sets without having to show the other party the whole data. What method might offer such a functionality? (Please select your answer and press "Send message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption
4. Differential privacy

U: Answer 1

F: Fantastic! This can be realized with a 2-party MPC protocol for example, in the simplest case by just finding common entries on two private lists (e.g. detecting money laundering or fraud, etc.). This method is called private set intersection and is one of the most mature and fast MPC applications. Besides MPC, there are HE solutions available that can be applied to this task.

U: Answer 2

F: I'm afraid FL isn't the correct answer. FL is typically applied in ML scenarios with many clients, where a usually public algorithm should be executed over the private input data. In this use-case, there is no need for ML, but for a private matching of the two data sets instead.

This can be realized with a 2-party MPC protocol for example, in the simplest case by just finding common entries on two private lists (e.g. detecting money laundering or fraud, etc.). This method is called private set intersection and is one of the most mature and fast MPC applications. Besides MPC, there are HE solutions available that can be applied to this task.

U: Answer 3

F: Good job! There are HE solutions available that can be applied to this task. Solutions based on 2-party MPC protocols are also possible and usually very efficient.

U: Answer 4

F: Unfortunately not. DP enhances the privacy of your data but does not allow you to compare two data sets without revealing their content.

This can be realized with a 2-party MPC protocol for example, in the simplest case by just finding common entries on two private lists (e.g. detecting money laundering or fraud, etc.). This method is called private set intersection and is one of the most mature and fast MPC applications. Besides MPC, there are HE solutions available that can be applied to this task.

DT11 A company has sensitive data on several devices, among them also 500 mobile measurement units with computer processors. They want to train and evaluate a machine learning model based on the combined data, but without revealing the input data to everyone. What method could they use? (Please select your answer and press "Send message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption

4. Differential privacy

U: Answer 1

F: Unfortunately not. The computational complexity of MPC scales drastically with the number of participants in the protocol, especially if there is also training of an ML model involved. Currently, MPC is not feasible due to the computational overhead for such a use case.

With FL they can train a global model without having to transmit the local sensitive data from the devices. However, keep in mind that the level of security is different compared to MPC or HE, the central server or a third party could still learn something about the training or evaluation data, but this is not possible with MPC or HE.

U: Answer 2

F: You're right. With FL they can train a global model without having to transmit the local sensitive data from the devices. However, keep in mind that the level of security is different compared to MPC or HE, the central server or a third party could still learn something about the training or evaluation data, but this is not possible with MPC or HE.

U: Answer 3

F: Unfortunately not. Training an encrypted ML model with so many inputs is currently not feasible due to performance issues of HE (and also MPC).

With FL they can train a global model without having to transmit the local sensitive data from the devices. However, keep in mind that the level of security is different compared to MPC or HE, the central server or a third party could still learn something about the training or evaluation data, but this is not possible with MPC or HE.

U: Answer 4

F: Unfortunately not. DP enhances the privacy of your data and can be combined with other methods for example, but it usually also comes at an accuracy price in ML applications. If it's important to have exact values in the evaluation, FL is the prime example in this scenario.

With FL they can train a global model without having to transmit the local sensitive data from the devices. However, keep in mind that the level of security is different compared to MPC or HE, the central server or a third party could still learn something about the training or evaluation data, but this is not possible with MPC or HE.

DT12 Assume it would massively improve your workflow if you could provide statistical summaries of sensors in your production to your suppliers. However, such summaries might also reveal some private information of specific sensors that could be leveraged by your competitors. What could you do? (Please select your answer and press "Send message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption
4. Differential privacy

U: Answer 1

F: I'm afraid MPC isn't the correct answer. Here you don't want to do a joint evaluation with multiple parties, you want to publish aggregated data while ensuring that the original input cannot be reconstructed.

With DP you would add random noise to your summaries to such a degree that your suppliers could still do meaningful evaluations with it but an identification of individual sensor data would be prohibited.

U: Answer 2

F: I'm afraid FL isn't the correct answer. There is no ML model that should be evaluated with client input, you only want to protect the aggregated summaries.

With DP you would add random noise to your summaries to such a degree that your suppliers could still do meaningful evaluations with it but an identification of individual sensor data would be prohibited.

U: Answer 3

F: I'm afraid HE isn't the correct answer. In general encrypted queries of your suppliers or similar forms of encrypted evaluations might also be possible here, but DP alone would already ensure that the individual sensor inputs are hidden and could not be identified.

With DP you would add random noise to your summaries to such a degree that your suppliers could still do meaningful evaluations with it but an identification of individual sensor data would be prohibited.

U: Answer 4

F: Right, with DP you would add random noise to your summaries to such a degree that your suppliers could still do meaningful evaluations with it but an identification of individual sensor data would be prohibited.

DT13 Employees in a company want to know if they earn more or less than the average employee without revealing how much they earn by themselves to anyone. (Please select your answer and press "Send message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption
4. Differential privacy

U: Answer 1

F: Right, this is a prime example of an MPC protocol, the employees do not need to share their private income data but can compute the overall average salary together based on MPC.

U: Answer 2

F: L could be applied if there are thousands of people working in the company and a central ML algorithm is trained with all inputs, for a simple average of private input you don't need an ML algorithm and can apply MPC here.

This is a prime example of an MPC protocol, the employees do not need to share their private income data but can compute the overall average salary together based on MPC.

U: Answer 3

F: With HE you typically encrypt your input data and send it to a remote evaluation. There are also methods that allow combining data from multiple users (so-called Multi-Key HE), so yes, given a low number of employees HE could also be used, but in general, this use case is a prime example for MPC.

The employees do not need to share their private income data but can compute the overall average salary together based on MPC.

U: Answer 4

F: DP enhances the privacy of your data based on added random noise, but with MPC you can calculate the result based on the exact data without revealing it to the others.

This is a prime example of an MPC protocol, the employees do not need to share their private income data but can compute the overall average salary together based on MPC.

DT14 A company wants to use health data from smartwatches to train and evaluate a global ML model without violating the privacy of the individual watch owners. How could this be realized? (Please select your answer and press "Send message")

1. Multiparty computation
2. Federated learning
3. Homomorphic encryption
4. Differential privacy

U: Answer 1

F: Interesting! but with MPC the watches would need to heavily communicate with each other to compute a joined function, also the number of smartwatch owners will easily surpass the maximal number of MPC clients in current protocols.

With FL, the model would be trained locally on the smartwatch and only the model parameters are then updated in the cloud. This will lead to a perfectly trained ML model that can handle data from thousands of input clients. However, if someone has access to the trained model on the server or intercepts the communication between the watch and server a malicious adversary could reconstruct the original health data. A combination with encryption methods would allow to enhance the privacy guarantees, but typically at the cost of performance.

U: Answer 2

F: You're right. With FL, the model would be trained locally on the smartwatch and only the model parameters are then updated in the cloud. This will lead to a perfectly trained ML model that can handle data from thousands of input clients. However, if someone has access to the trained model on the server or intercepts the communication between the watch and server a malicious adversary could reconstruct the original health data. A combination with encryption methods would allow to enhance the privacy guarantees, but typically at the cost of performance.

U: Answer 3

F: With HE you would encrypt data on your personal smartwatch, send the encrypted data to the server and get back the encrypted evaluation result (e.g. a prediction about my health, fitness, etc.). The server never knows my actual health data and only I can decrypt the evaluation result. However, if the ML model on the server should also be trained with the combined input of all watches the high number of participants makes it unrealistic to apply HE in this case.

With FL, the model would be trained locally on the smartwatch and only the model parameters are then updated in the cloud. This will lead to a perfectly trained ML model that can handle data from thousands of input clients. However, if someone has access to the trained model on the server or intercepts the communication between the watch and server a malicious adversary could reconstruct the original health data. A combination with encryption methods would allow to enhance the privacy guarantees, but typically at the cost of performance.

U: Answer 4

F: DP enhances the privacy of your data and can be combined with other methods for example, but it usually also comes at an accuracy price in ML applications. If it's important to have exact values in the evaluation, FL is the prime example in this scenario.

With FL, the model would be trained locally on the smartwatch and only the model parameters are then updated in the cloud. This will lead to a perfectly trained ML model that can handle data from thousands of input clients. However, if someone has access to the trained model on the server or intercepts the communication between the watch and server a malicious adversary could reconstruct the original health data. A combination with encryption methods would allow to enhance the privacy guarantees, but typically at the cost of performance.

DT15 The teaching part is over! Now, I'm going to start asking some general questions regarding your use-case where I try to narrow down the best method. Sometimes more than one method could be used or also combinations of them, by asking some more specific questions I will try to give you my best guess.

U: _*anything*_

F: _*goto DT17*_

DT16 OK. So we skip the teaching phase and start talking about your use-case.

I'm going to start asking some general questions regarding your use-case where I try to narrow down the best method. Sometimes more than one method could be used or also combinations of them, by asking some more specific questions I will try to give you my best guess.

U: _*anything*_

F: _*goto DT17*_

DT17 First, it makes a difference what kind of information you want to protect. Some methods allow to protect your algorithm as well as your data for example, do you know what you want to protect exactly?

1. Data
2. Algorithm
3. Both

U: Answer 1

F: Uhum, so you want to keep the data safe. _*goto DT17.1*_

U: Answer 2

F: Uhum, so you want to keep the algorithm(s) safe. _*goto DT17.2*_

U: Answer 3

F: Uhum, so you want to keep both data and algorithm(s) safe. _*goto DT17.3*_

DT17.1 Adding privacy to your evaluation usually comes at a performance cost, so it is important to know how complex your evaluation is and how many parties will be involved. Will there be more than 10 different parties or clients participating in the computation at the same time?

1. Yes
2. No

U: Answer 1

F: _*goto DT17.1.1*_

U: Answer 2

F: _*goto DT17.1.2*_

DT7.1.1 Do you want to outsource your computation, e.g. to a remote server or cloud?

1. Yes
2. No
3. I don't know.

U: Answer 1

F: That was my last question! Please have a closer look at FL and MPC. With MPC and FL you typically have to participate in the training and evaluation phase with local computation power, for smaller groups I would recommend MPC solutions here. FL will probably also work for you, but it has its advantages, especially for the larger number of clients where MPC becomes infeasible. _*goto DT18*_

U: Answer 2

F: That was my last question! Please have a closer look at DP, HE and MPC. Outsourcing computations for smaller groups or individuals with the strongest privacy guarantees calls for HE. In MPC protocols the parties providing data can also differ from the parties doing the computation, e.g. three data providers secretly share their data with two computational nodes where the evaluation is performed. If you don't want to encrypt your data the application of DP might also be an option if you have large data sets and don't depend on the evaluation of unperturbed values. _*goto DT18*_

U: Answer 3

F: That was my last question! Outsourcing means that you send your data to a remote location where the computation takes place. If you are not sure here, in principle all mentioned PETs offer a privacy advantage. Since you are dealing with a small number of involved parties I suggest taking a closer look at MPC and HE. _*goto DT18*_

DT17.1.2 OK, it is also important to know how fast you need your results since the computational overhead for some methods grows massively with the number of involved parties. Do you need your results in real-time (let's say within seconds or minutes)?

1. Yes
2. No

U: Answer 1

F: That was my last question! To the best of my knowledge, FL and DP might be useful. Large groups and real-time results currently limit the application of cryptographic methods to some degree, I suggest you take a closer look at FL. You could maybe also combine your evaluation with DP. *_goto DT18_*

U: Answer 2

F: That was my last question! Based on your situation, all methods could be feasible. Large groups of participants currently limit the application of cryptographic methods to some degree, but since your computation can also take some time, some clever MPC approaches might still work in your case. I suggest taking a closer look at FL and possible combinations with MPC or HE in order to exploit the efficiency of FL for large groups while enhancing the security of the method with cryptographic approaches. *_goto DT18_*

DT17.2 Adding privacy to your evaluation usually comes at a performance cost, so it is important to know how complex your evaluation is and how many parties will be involved. Will there be more than 10 different parties or clients participating in the computation at the same time?

1. Yes
2. No

U: Answer 1

F: Protecting algorithms and having more than 10 clients is quite difficult since the performance of the cryptographic methods becomes very challenging and methods like FL or DP usually don't protect algorithms. Let's try with this explanation and another question regarding the complexity of the calculation:

The performance overhead of cryptographic methods like HE or MPC might be very challenging in your case, but ML-based approaches like FL usually don't protect algorithms.

U: Answer 2

F: That was my last question! I suggest taking a closer look at the HE and MPC. Cryptographic methods like HE or MPC are very likely the best methods for your use case. With HE you can encrypt your algorithm and still allow someone to perform evaluations. An MPC-based approach could be private classification, where one party owns a private model, and the other party wants to use this model to classify some private input. *_goto DT18_*

DT17.2.1 On a scale from 1-10, how complex is your evaluation going to be (e.g. 10 being the training of a very complex and deep neural network, 1 being a simple calculation like the mean of the combined input)?

U: 1 - 5

F: Take a closer look at MPC and Multikey-HE, I assume they are the best options in your case. _*goto DT18*_

U: 6 - 10

F: It seems you are looking for a hybrid solution where you exploit the efficiency of ML approaches and combine them with the protection possibilities of cryptographic methods, which is at the forefront of current research. You need to talk to one of our experts in the field to discuss your options in detail. _*goto DT18*_

DT17.3 Adding privacy to your evaluation usually comes at a performance cost, so it is important to know how complex your evaluation is and how many parties will be involved. How many different parties or clients will be participating in the computation at the same time?

U: 2

F: HE, MPC (e.g. PSI)

Private calculations between two parties where both data and algorithms are protected are possible with cryptographic methods like HE or MPC protocols. A very efficient MPC solution to find common entries in two private databases is a private set intersection for example. HE on the other hand allows you to perform evaluations where both, data and algorithm, are encrypted. _*goto DT18*_

U: 3 - 10

F: MPC

MPC typically offers efficient solutions for a smaller number of involved clients and allows for example collaborative learning on combined private datasets or private classification, where one party owns a private model, and the other party wants to use this model to classify some private input. Data providers and computation nodes also don't have to be identical, three data providers can for example secretly share their data with two computational nodes where the evaluation is performed. HE solutions for more than two parties are currently also investigated in our research group. _*goto DT18*_

U: more than 10

F: it gets very difficult with both methods (HE, MPC).

The performance overhead of cryptographic methods like HE or MPC might be very challenging in your case, but ML-based approaches like FL usually don't protect algorithms. Maybe a hybrid solution where you exploit the efficiency of ML approaches and combine them with the protection possibilities of cryptographic methods is possible in your case, please talk to one of our experts in the field to discuss your options in detail. _goto DT18_

DT18 The conversation is over! I hope you find out how to apply PETs to your use-case.

A.3. Chapter 6: The Dialogue Structure

In Section 6, an agent with the capability of identifying Toulmin's core components [Tou03] is described. In Table A.1, all the steps and conditions that the agent reacts based on are illustrated.

Table A.1.: The whole dialogue structure and branches for the agent introduced in Chapter 6.

Agent's response	**Condition/Expalnation**
Hi, I'm Rebo4AI, a prototypical conversational agent. With me you can currently discuss only about intelligence. Is that OK with you; do you want to start a conversation? (please answer "yes" to continue)	The greeting and introduction of the dialogue.
Do you remember the definitions given in XXX – what are they?	Remembering the definitions of intelligence (this part would need to be changed depending on the educational context in which this agent is embedded)
What do you think, what does entity – chosen randomly from the list of definitions actually mean?	Understanding of the definitions of intelligence
So, I'm curious: Is an entity now intelligent or not? Please tell me whether you think it is intelligent or not, which definition you use to make this claim, and which evidence of background knowledge makes you think so?	Applying the definitions of intelligence to an example. This is the first chance for users to mention claims, warrants and evidence.
Claim conditions: Looking for the missing claim component	
Mmmm, I don't get your opinion about it. Could you tell me clearly whether it is intelligent or not? And then explain why you think like that?	*Condition C2*: This is the second try for getting the user's claim. The concretely missing core component (claim) is pointed out.
Unfortunately, I cannot understand what you just said. Please try to answer the original question of whether an entity is intelligent or not, and why, by starting the argumentation with: "The entity is (not) intelligent because ..."	*Condition C3*: This is the third and last try for getting the user's claim. A sentence starter is given to scaffold the correct argument.

I could not understand your assertion regarding the entity. I think I need to improve my understanding to find out what you are saying. So, let's talk later.	*Condition C4*: After the third chance, the agent will end the conversation, if until here we haven't been able to understand the claim, either the user isn't engaging with the agent, or the classifiers are substantially failing.
Warrant conditions: Looking for the missing warrant component	
(1) I feel it should be intelligent too. (2) Interesting! You think it is intelligent. (3) I feel it should not be intelligent too. (4) Interesting! You think it isn't intelligent. But I couldn't understand to which of the five definitions of intelligence you refer to. Please use at least one of these definitions explicitly when arguing why an entity is (not) intelligent: thinking/acting humanly/rationally or being able to adapt behaviour to a changing environment in order to achieve its goals.	*Condition W2*: Based on the user's claim and the entity, the agent gives different feedback. The first line is actually adaptive to the direction of the user's claim and whether this agrees with what is laid down as the direction of the claim that was defined as most reasonable by the authors. This is the second try to get the user's warrant.
Mmmm, I'm sorry, I still don't understand which of the five definitions you refer to. Could you please try and phrase your answer like this: <the entity> is (not) intelligent in the sense of (not) being able to act or think rationally." You should mention some of the definitions and then explain why they (don't) fit to the entity.	*Condition W3*: This is the third try for getting the user's warrant. Based on the user's claim, the agent tries to help the user by saying how to write. A sentence starter is given to scaffold the correct argument. The second part of the response will change based on the user's claim.
Unfortunately, I could not understand based on which definitions you said the entity is intelligent. Anyway, my developer is already working on giving me more knowledge, so in the future, I will be able to discuss more about what intelligence is, and in what sense different pieces of AI-based technology are intelligent.	*Condition W4*: After the third chance, the agent will end the conversation.
Condition 3: Looking for the missing evidence component	

Great, I think I understand already a lot of what you're saying – one thing isn't clear yet: I don't understand which evidence or background knowledge you use in order to decide that <the entity> fits to the definition of intelligence you used above. Could you explain why you think like this? You can talk about the characteristics of <the entity> to show that how <the entity> (doesn't) fits to the definitions.	*Condition E2*: This is the second try for getting the user's evidence. The second part of the response will be changed based on the user's claim.
Mmmm, I'm sorry, I still don't understand how you argue that <the entity> (doesn't) fits to the definitions. For instance, you can say monkey can act rationally because they know how to take care of themselves. Or they can make tools which is related to learn new things to achieve goals. So, how do you justify the mentioned definitions fit to <the entity>? For instance, you can say since a pen is an inanimate object, it cannot act or think. That is why a pen cannot be called intelligent. So, how do you justify the mentioned definitions fit to <the entity>?	*Condition E3*: This is the third try for getting the user's warrant. Based on the user's claim, the explanation about the evidence will be changed.
Unfortunately, I could not understand your evidence or observation related to the definitions of intelligence you used. Anyways, my developer is already working on giving me more knowledge, so in the future, I will be able to discuss more about what intelligence is, and in what sense different pieces of AI-based technology are intelligent.	*Condition E4*: This is the last chance for the user to talk about his or her observation or experience regarding his or her claim. After the third chance, the agent will end the conversation
All the core components are mentioned.	
OK, that makes sense. My developer is already working on giving me more knowledge, in this case about what makes the entity intelligent or not. So, in the future, I will be able to discuss more about what intelligence is, and in what sense different pieces of AI-based technology are intelligent.	This is the agent response when all the core components were mentioned by the user.

A.4. Chapter 8: Expected Phrases

Here the expected phrases for each component are listed. These phrases and the user responses were used as input to the pre-trained transformer models. The models used the inputs to identify the core components, claims, warrants and evidence, in the users' responses based on semantic similarity.

A.4.1. Test Case 1

In Test Case 1, the question was about the GDPR's principles. In the question, the correct answer or the violated principle was the data minimisation which is also the third principle.

Claim Component

The claim component corresponds with the direct answer to the question or the violated principle. In this test case, three different states were defined for the claims, *incorrect_ claim*, *correct_ claim* and *without_ claim*. The expected phrases of the claim component are as follows:

1. "*lawfulness, fairness and transparency principle has been violated*"
2. "*fairness principle has been violated*"
3. "*transparency principle has been violated*"
4. "*lawfulness principle has been violated*"
5. "*first principle has been violated*"
6. "*purpose limitation principle has been violated*"
7. "*second principle has been violated*"
8. "*minimization principle has been violated*"
9. "*third principle has been violated*"
10. "*Accuracy principle has been violated*"
11. "*forth principle has been violated*"
12. "*storage limitation principle has been violated*"
13. "*fifth principle has been violated*"
14. "*integrity and confidentiality principle has been violated*"
15. "*sixth principle has been violated*"

16. "*accountability principle has been violated*"
17. "*seventh principle has been violated*"
18. "*last principle has been violated*"

Warrant Component

The warrant component points to the definition of the violated or the correct principle. Here we only looked for the existence of the correct warrant which was the definition of the data minimisation principle. Therefore, we defined two cases for the warrant, *with_warrant* and *without_warrant*. The expected phrases of the warrant component are as follows:

1. "*personal data must be adequate, relevant and limited to what is necessary in relation to the purposes*"
2. "*data cannot be collected or processed unless it is needed*"
3. "*data should be necessary in relation to the purposes*"
4. "*adequate, relevant and limited*"
5. "*it stores data which may never be needed*"
6. "*adequate, relevant and limited to the purposes of the processing*"

Evidence Component

The evidence component refers to the information about the recruitment agency. Similar to the warrant component, only the existence of the evidence is checked. The expected phrases of the evidence component are as follows:

1. "*the agency sends to all applicants a general questionnaire, which also includes specific questions about health conditions, that only relevant to particular occupations.*"
2. "*They sends the same survey to all applicants*"
3. "*They asked the same question from all workers*"
4. "*they collected the same information from all workers*"
5. "*there was no need to collect the blood groups of all workers*"
6. "*there was no need to collect the health-related data of all workers*"
7. "*there was no need to collect the data on the health conditions of all workers*"
8. "*the blood groups were not related to all applicants*"
9. "*the health conditions were not related to all applicants*"

10. "*the blood groups are not relevant to all workers*"
11. "*the health condition data are not relevant to all workers*"
12. "*By storing data on health conditions by all applicants that are only relevant to particular occupations*"
13. "*As the collection of this health-related data is not for all workers*"

A.4.2. Test Case 2

Similar to Test Case 1, the question was about the GDPR's principles, however, the example and the correct answer was different. In the question, the correct answer or the violated principle was the accuracy principle which is also the fourth principle.

Claim Component

The claim component corresponds with the direct answer to the question or the violated principle. In this test case, we used the same states and the expected phrases of Test Case 1 for the claim component but *correct_ claim* refers to the accuracy principle.

Warrant Component

The warrant component corresponds with the definition of the correct principles, the accuracy principle. The expected phrases of the warrant component are as follows:

1. "*the data should also accurately reflect the order of events*"
2. "*reflect the order of events*"
3. "*Accuracy is the fourth principle meaning that it is required to ensure that personal data are accurate and are kept up to date where it is necessary. The data should also accurately reflect the order of events. Personal data that are inaccurate – considering the purposes for their processing – must be deleted or rectified without any delay*"

Evidence Component

The evidence component refers to the information about the hospital. Similar to the warrant component, only the existence of the evidence is checked. The expected phrases of the evidence component are as follows:

1. "*only the last diagnosis of a medical condition continues to be held*"
2. "*the previous diagnoses are deleted*"
3. "*just keep the last diagnosis of each patient and delete the old ones.*"
4. "*delete the old data.*"

5. "*delete the old treatments.*"
6. "*just keep the last diagnosis and delete the old ones*"
7. "*store the last information and delete the old ones.*"
8. "*keep all the records of their patients*"
9. "*they deleted the medical data of patients*"

A.4.3. Test Case 3

Test Case 3 was about the intelligence. The question was about whether monkey are intelligent. The learners were asked to used the offered definitions to justify their answer.

Claim Component

The claim component corresponds with the direct answer to the question. In this test case, we defined two cases for the claim component, *with_ claim* and *without_ claim*. The expected phrases of the claim component are as follows:

1. "*A monkey is intelligent.*"
2. "*a monkey is an intelligent being*"
3. "*Yes, I think a monkey is intelligent.*"
4. "*A monkey is intelligent because*"
5. "*Yes, I would say that a monkey is intelligent*"
6. "*Monkeys are intelligent because*"
7. "*Therefore it is intelligent.*"
8. "*Yes, I think a monkey is really intelligent because*"
9. "*they have the characteristics of intelligence*"
10. "*Absolutely*"
11. "*Yes I think monkeys are intelligent*"
12. "*A monkey is intelligent as they are*"
13. "*Monkeys are intelligent because*"
14. "*A monkey would certainly be considered intelligent.*"
15. "*A monkey is definitely intelligent considering the 5 criteria for intelligence provided.*"

16. "*The monkey is almost certainly intelligent*"
17. "*yes*"
18. "*it is generally considered as intelligent*"
19. "*Yes, a money is intelligent*"
20. "*A monkey is very intelligent*"

Warrant Component

In Test Case 3, the warrant component was about the definitions offered by the agent. Since the learners were expected to mention at least one of the definitions, we only looked for the existence of them in the learners' responses. The expected phrases of the warrant component are as follows:

1. "*it things rationally and acts humanly.*"
2. "*It has a goal and always learn from previous experiences.*"
3. "*the monkey thinks, analyzes, acts rationally*"
4. "*learning from their mistakes.*"
5. "*This helps them better reach their goals.*"
6. "*it can be observed thinking and behaving like a human*"
7. "*it can think and act rationally and also can learn from experience to reach*"
8. "*because it acts like a human*"
9. "*can think and act rationally.*"
10. "*they are able to learn from experience in order to better reach their goals*"
11. "*they act humanly is certain circumstances*"
12. "*Whilst it can't be said to think or act humanly (by definition only humans can do that)*"
13. "*it can certainly think and act both intelligently and rationally, and most certainly learns from experiences.*"
14. "*it acts almost as clever as a human being and their amazing thinking skills are beyond our wildest hopes and dreams as an animal.*"
15. "*It doesn't think humanely or act humanely*"
16. "*it acts rationally*"

17. “*so they obviously learn from experience.*”
18. “*act according to the outcomes that they want to achieve.*”
19. “*they can rationally act and work towards attaining desired aims.*”
20. “*they are able to think rationally, act rationally and learn from experience in order to better reach its goals.*”
21. “*use rational thinking in order to*”
22. “*a monkey would not be considered to think or act humanly*”
23. “*he is able to think rationally, act rationally and learn from experience.*”
24. “*This proves that is is able to learn from his experiences.*”
25. “*he can rationally think*”
26. “*they share many human instincts, actions, and mannerisms with our species.*”
27. “*Monkeys may not always act rationally,*”
28. “*ability to learn from their experiences*”
29. “*Monkeys are shown to learn from experience, adapt*”
30. “*as it can learn from past experiences to better reach it's goals.*”
31. “*the monkey also acts rationally most of the time for it's own goals*”
32. “*You could argue that the monkey doesn't think humanly*”
33. “*it would be hard to argue that a monkey doesn't think humanly, as humans act in a wide variety of ways.*”
34. “*It learns from experience, reasons, plans, adapts to its environment*”
35. “*it is able to think and act rationally*”

Evidence Component

The question in this test case was about monkeys, therefore, the evidence component refers to any information about the entity. Similar to the warrant component, only the existence of the evidence is checked. The expected phrases of the evidence component are as follows:

1. “*It's able to make decisions*”
2. “*learns from things or tests to which they are subjected*”
3. “*I think monkeys have a way of passing their knowledge to their one another*”

4. "*its goals of survival.*"
5. "*it seeks to have a relationship with its mate and community members*"
6. "*it spends time relaxing as well as working on tasks like getting food.*"
7. "*A lot of monkeys have been shown to be able to learn*"
8. "*communicate with sign language, and do many complex tasks*"
9. "*Monkeys have brains*"
10. "*this is evidenced by how they raise and care for their young and the fact that some monkeys have acquired the skills to use basic tools.*"
11. "*Monkeys and humans are evolutionary speaking very close*"
12. "*Humans are very similar to monkeys in many ways*"
13. "*We have seen, for example, monkeys constructing tools and using them,*"
14. "*implying a grasp of cause and effect and goal-motivated thinking*"
15. "*monkeys are profoundly trainable*"
16. "*There are many examples of monkeys being challenged with puzzles that reward them with food.*"
17. "*The monkeys can become frustrated, or happy depending on their success of solving the puzzle.*"
18. "*once they learn the key to solving a puzzle, the next time they are challenged with a similar problem they can use their past experience to solve it more easily.*"
19. "*they are able to learn new behaviors, alter old ones, and repeat others, based on negative and/or positive reinforcement*"
20. "*altering their actions to match the situation*"
21. "*monkeys can learn to play games*"
22. "*problem solve and achieve in a game*"
23. "*monkeys can be taught to recognize symbols, and words.*"
24. "*Monkeys can be taught to do certain tricks and skills and they will change what they are doing to improve the trick or skill.*"
25. "*A monkey will rationally find food when he needs to eat.*"
26. "*He knows what methods of hunting work and what methods don't work*"

27. "*If he is hungry he can rationally think about how to go about rectifying that need.*"

28. "*food puzzle experiments where monkeys need to solve puzzles to reach a food goal.*"

29. "*Monkeys demonstrate that they know how to find food, where to look, and avoid repeating the same mistakes a lot of the time.*"

30. "*they make tools*"

31. "*behaviour and responses reflect its past experiences*"

32. "*It is able to detect dangerous situations*"

33. "*warn the rest of the group*"

34. "*act together to protect the group*"

35. "*able to search for food and find it*"

36. "*Monkeys have similar manners and mannerisms to humans including eating, grooming, and socially.*"

37. "*Monkeys often form families and social groups like humans.*"

38. "*A gorilla was even taught the English language.*"

39. "*They also invent tools to help them with goals such as using sharp rocks to beak open hard fruit*"

A.5. Chapter 8: Agent's Responses

In this section, we listed the agent responses or the follow-up questions in each branch of the dialogue.

A.5.1. Test Case 1

In this test case, the agent reacted differently when the claim was incorrect or there was no claim. In case of identifying the correct claim, the agent gives feedback based on the other two components, the warrant and evidence. The branches and the agent's responses are as follows:

1. "*incorrect_claim*": "*Mmmmm, interesting! Let's focus on the one which was clearly violated, the data minimization principle. Please read the principle and the scenario again and then tell me which part of the definitions was violated and why?*"

2. "*without_claim*": "*Mmmm, could you clearly specify which principle was violated? And explain why?*"

3. "*correct_claim, with_warrant, with_evidence*": "*Good job! Your answer is complete and correct.*"
4. "*correct_claim, with_warrant, without_evidence*": "*Great! I think I understood already a lot of what you're saying – one thing isn't clear yet: I don't understand how the agency violated the principle. Please read the scenario again and tell me what exactly they did that led to the violation of the principle.*"
5. "*correct_claim, without_warrant, with_evidence*": "*I understood already a lot of what you're saying – one thing isn't clear yet: I don't understand which part of the principle definition has been violated. Please read the principle explanation again and tell me which part of it was violated by the agency?*"
6. "*correct_claim, without_warrant, without_evidence*": "*Great! I just understood the principle that you mentioned which is correct. Please read the scenario and the principle's definition again and tell me which part of the definitions was violated and why?*"

A.5.2. Test Case 2

In Test Case 2, we used the same branches used in Test Case 2. The branches and the agent's responses are as follows:

1. "*incorrect_claim*": "*Mmmmm, interesting! Let's focus on the one which was clearly violated, the accuracy principle. Please read the principle and the scenario again and then tell me which part of the definitions was violated and why?*"
2. "*without_claim*": "*Mmmm, could you clearly specify which principle was violated? And explain why?*"
3. "*correct_claim, with_warrant, with_evidence*": "*Good job! Your answer is complete and correct.*"
4. "*correct_claim, with_warrant, without_evidence*": "*Great! I think I understood already a lot of what you're saying – one thing isn't clear yet: I don't understand how the hospital violated the principle. Please read the scenario again and tell me what exactly they did that led to the violation of the principle.*"
5. "*correct_claim, without_warrant, with_evidence*": "*I understood already a lot of what you're saying – one thing isn't clear yet: I don't understand which part of the principle definition has been violated. Please read the principle explanation again and tell me which part of it was violated by the hospital?*"
6. "*correct_claim, without_warrant, without_evidence*": "*Great! I just understood the principle that you mentioned which is correct. Please read the scenario and the principle's definition again and tell me which part of the definitions was violated and why?*"

A.5.3. Test Case 3

The branches are different here. The agent only checked the existence of the claim component. The branches and the agent's responses are as follows:

1. "*without_claim*": "*Mmmm, I don't get your opinion about it. Could you tell me clearly whether it is intelligent or not? And why?*"
2. "*with_claim, with_warrant, with_evidence*": "*Good job! Your answer is complete. You justified your claim using the definitions and some facts or evidence about monkeys.*"
3. "*with_claim, with_warrant, without_evidence*": "*Interesting! But the evidence part is missing! Could you explain why you think monkeys are intelligent based on the definition(s) that you mentioned?*"
4. "*with_claim, without_warrant, with_evidence*": "*Good to hear! But I couldn't understand to which one of the five definitions of intelligence you refer. Please use at least one of these definitions expilicitly when arguing why monkey are intelligent or not: thinking/acting humanly/rationally or being able to adapt behavior to a changing environment in order to achieve its goals.*"
5. "*with_claim, without_warrant, without_evidence*": "*Great! I couldn't understand to which of the five definitions of intelligence you refer to. Please use at least one of these definitions explicitly and tell me how it fits or doesn't fit to monkeys*"

A.6. Chapter 8: SBERT Models

In this section, the pre-trained transformer-based models used in the chapter are listed. We compared 16 different SBERT models to find the best one for evaluating the workflow. For more information, please check the Huggingface[1]. The models are as follows:

1. **all-mpnet-base-v2**:
   - *Base Model*: microsoft/mpnet-base
   - *Max Sequence Length*: 384
   - *Dimensions*: 768
   - *Normalized Embeddings*: true
   - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
   - *Size*: 420 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: 1B+ training pairs. For details, see the model card.

[1]https://huggingface.co/sentence-transformers

- *Model Card*: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

2. **all-distilroberta-v1**:
   - *Base Model*: distilroberta-base
   - *Max Sequence Length*: 512
   - *Dimensions*: 768
   - *Normalized Embeddings*: true
   - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
   - *Size*: 290 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: 1B+ training pairs. For details, see the model card.
   - *Model Card*: https://huggingface.co/sentence-transformers/all-distilroberta-v1

3. **all-MiniLM-L12-v2**:
   - *Base Model*: microsoft/MiniLM-L12-H384-uncased
   - *Max Sequence Length*: 256
   - *Dimensions*: 384
   - *Normalized Embeddings*: true
   - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
   - *Size*: 120 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: 1B+ training pairs. For details, see the model card.
   - *Model Card*: https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

4. **all-MiniLM-L6-v2**:
   - *Base Model*: nreimers/MiniLM-L6-H384-uncased
   - *Max Sequence Length*: 256
   - *Dimensions*: 384
   - *Normalized Embeddings*: true
   - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
   - *Size*: 80 MB
   - *Pooling*: Mean Pooling

- *Training Data*: 1B+ training pairs. For details, see the model card.
- *Model Card*: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

5. **paraphrase-multilingual-mpnet-base-v2**:
   - *Base Model*: Teacher: paraphrase-mpnet-base-v2; Student: xlm-roberta-base
   - *Max Sequence Length*: 128
   - *Dimensions*: 768
   - *Normalized Embeddings*: false
   - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
   - *Size*: 970 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: Multi-lingual model of paraphrase-mpnet-base-v2, extended to 50+ languages.
   - *Model Card*: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

6. **paraphrase-albert-small-v2**:
   - *Description*: All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs.
   - *Base Model*: nreimers/albert-small-v2
   - *Max Sequence Length*: 256
   - *Dimensions*: 768
   - *Normalized Embeddings*: false
   - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
   - *Size*: 43 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: AllNLI, sentence-compression, SimpleWiki, altlex, msmarco-triplets, quora_duplicates, coco_captions,flickr30k_captions, yahoo_answers_title_question, S2ORC_citation_pairs, stackexchange_duplicate_questions, wiki-atomic-edits
   - *Model Card*: https://huggingface.co/sentence-transformers/paraphrase-albert-small-v2

7. **paraphrase-multilingual-MiniLM-L12-v2**:
   - *Base Model*: Teacher: paraphrase-MiniLM-L12-v2; Student: microsoft/Multilingual-MiniLM-L12-H384
   - *Max Sequence Length*: 128
   - *Dimensions*: 384

- *Normalized Embeddings*: false
- *Suitable Score Functions*: cosine-similarity (util.cos_sim)
- *Size*: 420 MB
- *Pooling*: Mean Pooling
- *Training Data*: Multi-lingual model of paraphrase-multilingual-MiniLM-L12-v2, extended to 50+ languages.
- *Model Card*: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

8. **paraphrase-MiniLM-L3-v2**:
   - *Description*: All-round model tuned for many use-cases. Trained on a large and diverse dataset of over 1 billion training pairs.
   - *Base Model*: nreimers/MiniLM-L6-H384-uncased
   - *Max Sequence Length*: 128
   - *Dimensions*: 384
   - *Normalized Embeddings*: false
   - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
   - *Size*: 61 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: AllNLI, sentence-compression, SimpleWiki, altlex, msmarco-triplets, quora_duplicates, coco_captions,flickr30k_captions, yahoo_answers_title_quest S2ORC_citation_pairs, stackexchange_duplicate_questions, wiki-atomic-edits
   - *Model Card*: https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L3-v2

9. **distiluse-base-multilingual-cased-v1**:
   - *Base Model*: Teacher: mUSE; Student: distilbert-base-multilingual
   - *Max Sequence Length*: 128
   - *Dimensions*: 512
   - *Normalized Embeddings*: false
   - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
   - *Size*: 480 MB
   - *Pooling*: Mean Pooling
   - *Training Data*: Multi-Lingual model of Universal Sentence Encoder for 15 languages: Arabic, Chinese, Dutch, English, French, German, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish.

- *Model Card*: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v1

10. **distiluse-base-multilingual-cased-v2**:
    - *Base Model*: Teacher: mUSE; Student: distilbert-base-multilingual
    - *Max Sequence Length*: 128
    - *Dimensions*: 512
    - *Normalized Embeddings*: false
    - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
    - *Size*: 480 MB
    - *Pooling*: Mean Pooling
    - *Training Data*: Multi-Lingual model of Universal Sentence Encoder for 50 languages.
    - *Model Card*: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

11. **multi-qa-MiniLM-L6-cos-v1**:
    - *Base Model*: nreimers/MiniLM-L6-H384-uncased
    - *Max Sequence Length*: 512
    - *Dimensions*: 384
    - *Normalized Embeddings*: true
    - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
    - *Size*: 80 MB
    - *Pooling*: Mean Pooling
    - *Training Data*: 215M (question, answer) pairs from diverse sources.
    - *Model Card*: https://huggingface.co/sentence-transformers/multi-qa-MiniLM-L6-cos-v1

12. **multi-qa-distilbert-cos-v1**:
    - *Base Model*: distilbert-base
    - *Max Sequence Length*: 512
    - *Dimensions*: 768
    - *Normalized Embeddings*: true
    - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
    - *Size*: 250 MB

- *Pooling*: Mean Pooling
- *Training Data*: 215M (question, answer) pairs from diverse sources.
- *Model Card*: https://huggingface.co/sentence-transformers/multi-qa-distilbert-cos-v1

13. **multi-qa-mpnet-base-cos-v1**:
    - *Base Model*: microsoft/mpnet-base
    - *Max Sequence Length*: 512
    - *Dimensions*: 768
    - *Normalized Embeddings*: false
    - *Suitable Score Functions*: cosine-similarity (util.cos_sim)
    - *Size*: 420 MB
    - *Pooling*: CLS Pooling
    - *Training Data*: 215M (question, answer) pairs from diverse sources.
    - *Model Card*: https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1
14. **msmarco-MiniLM-L6-cos-v5**:
    - *Base Model*: MSMARCO-Passage-Ranking
    - *Max Sequence Length*: 128
    - *Dimensions*: 384
    - *Normalized Embeddings*: true
    - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
    - *Size*: 280 MB
    - *Pooling*: Mean Pooling
    - *Training Data*: MS MARCO Passages dataset
    - *Model Card*: https://huggingface.co/sentence-transformers/msmarco-MiniLM-L6-cos-v5
15. **msmarco-MiniLM-L12-cos-v5**:
    - *Base Model*: MSMARCO-Passage-Ranking
    - *Max Sequence Length*: 128
    - *Dimensions*: 768
    - *Normalized Embeddings*: true
    - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance

- *Size*: 420 MB
- *Pooling*: Mean Pooling
- *Training Data*: MS MARCO Passages dataset
- *Model Card*: https://huggingface.co/sentence-transformers/msmarco-MiniLM-L12-cos-v5

16. **msmarco-distilbert-cos-v5**:
    - *Base Model*: MSMARCO-Passage-Ranking
    - *Max Sequence Length*: 128
    - *Dimensions*: 768
    - *Normalized Embeddings*: true
    - *Suitable Score Functions*: dot-product (util.dot_score), cosine-similarity (util.cos_sim), euclidean distance
    - *Size*: 540 MB
    - *Pooling*: Mean Pooling
    - *Training Data*: MS MARCO Passages dataset
    - *Model Card*: https://huggingface.co/sentence-transformers/msmarco-distilbert-cos-v5

A.7. Chapter 5: The Questions and Answers

Figure A.3, shows the complete dialogue's structure of the agent (DIGIBOT) discussed in Chapter 5. The agent asks about all the operators one by one and finally asks for the learner's feedback on the content.

Throughout the conversation, users are prompted to engage in self-assessment on two occasions. DIGIBOT incorporates this feature to encourage learners to reflect on their existing knowledge and also to become aware of the agent's expectations. Following the self-assessment phase, the agent poses a challenging question referred to as "the big question," as illustrated in Figure A.3. Answering this big question necessitates the utilization of most of the introduced operators. The question is presented as follows: "

I would like to ask you to write a search query for the following task:

- *You're looking for information about "Nikola Tesla" or "Thomas Edison".*
- *You need all the webpages whose titles contain exactly "Top X facts" and X should be a range from 3 to 10, not just a specific number.*
- *You also want to exclude all YouTube pages from the search results.*

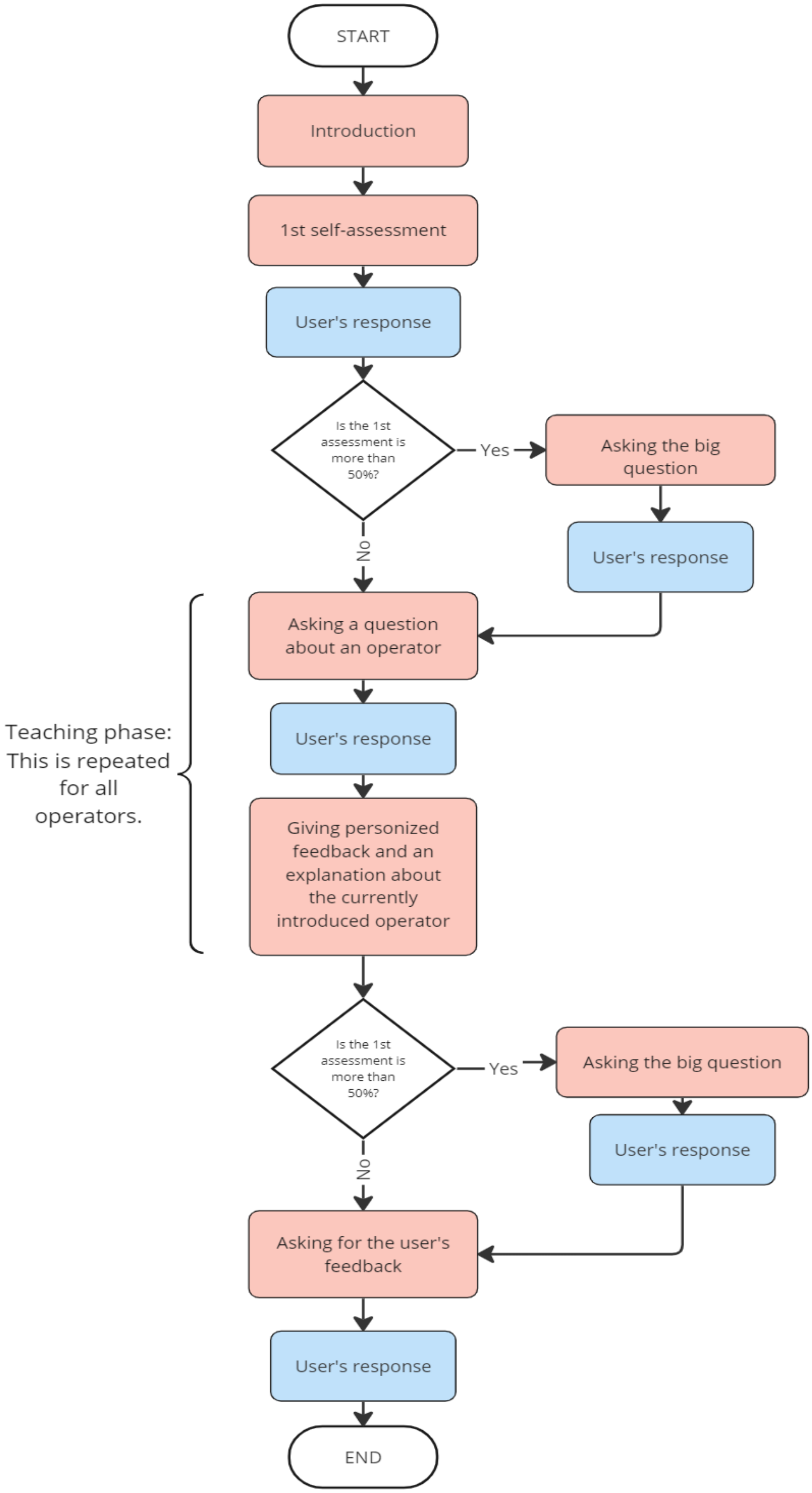


Figure A.3.: The dialogue's flow of the agent introduced in Chapter 5

DIGIVIDget and DIGIBOT introduce each operator through the presentation of questions (which can be in multiple-choice or free-text format) and also offer feedback on the learners' responses to these questions. Below, the list of questions and the feedback provided by DIGIBOT are shown. All the questions and feedback are translated into English.

Exact match operator (multiple-choice question): *Which query retrieves those webpages in which both words - Thomas and Edison - come next to each other?*

1. *"Thomas Edison"*
   - Feedback: *Awesome!!! That's great!*
     To force Google to perform exact-matching, the respective words need to be enclosed in quotation marks. This ensures also that Google does not use synonyms or different word forms.

2. *Thomas Edison*
   - Feedback: *Sorry, this answer is wrong.*
     This query returns all webpages that contain both words - Thomas and Edison - but they are not necessarily written next to each other. For example, this query could also find the following sentence: "Edison Chen, the lead actor, was awesome in his role as Thomas Jefferson, the president." The correct answer is "Thomas Edison".
     To force Google to perform exact-matching, the respective words need to be enclosed in quotation marks. This ensures also that Google does not use synonyms or different word forms.

3. *Both*
   - Feedback: *Sorry, this answer is wrong. The correct answer is "Thomas Edison".*
     If you had selected Thomas Edison, then documents with the following example sentence would be found, but Thomas and Edison are not next to each other: "Edison Chen, the lead actor, was awesome in his role as Thomas Jefferson, the president."
     To force Google to perform exact-matching, the respective words need to be enclosed in quotation marks. This ensures also that Google does not use synonyms or different word forms.

OR operator (free-text question[2]): *Let's practice the use of quotation marks (exact matching) and try out an additional operator. I'm curious if you can guess the additional operator we need for the following task:*
Please write a new query to find webpages that contain at least one of the full names "Nikola Tesla" or "Thomas Edison".

1. *Correct answer*

[2]Free-text questions are classified the learners' answers into different categories.

- Feedback: *Great, you hit the target!*
  The OR operator is the right choice. By using OR (in upper cases) we tell Google that the resulting webpages should contain at least one of the names "Nikola Tesla" or "Thomas Edison" or both.

2. *Lower-case "OR"*

   - Feedback: *Good try! The query is almost correct.*
     You used "or" in lower cases which is incorrect, because it is not interpreted as an operator by Google and Google would have found only those webpages that contain both names. The correct query should be like this "Thomas Edison" OR "Nikola Tesla".
     The OR operator is the right choice. By using OR (in upper cases) we tell Google that the resulting webpages should contain at least one of the names "Nikola Tesla" or "Thomas Edison" or both.

3. *"I don't know" or asking for help*

   - Feedback: *That's okay, I'll show you how to do it. The query should be like this "Thomas Edison" OR "Nikola Tesla".*
     The OR operator is the right choice. By using OR (in upper cases) we tell Google that the resulting webpages should contain at least one of the names "Nikola Tesla" or "Thomas Edison" or both.

4. *for all other cases*

   - Feedback: *Sorry but your answer is wrong. The query should be like this: "Thomas Edison" OR "Nikola Tesla".*
     The OR operator is the right choice. By using OR (in upper cases) we tell Google that the resulting webpages should contain at least one of the names "Nikola Tesla" or "Thomas Edison" or both.

Asterisk Operator (multiple-choice question): *Thomas Edison's full name is Thomas Alva Edison. This means that on some webpages the name might be written like this: Thomas Alva Edison or Thomas A. Edison. Please tell me, which query below covers both cases?*

1. *"Thomas Edison"*

   - Feedback: *Not this time! This query forces Google to find results with exactly the term Thomas Edison. We don't get any webpages which contain Thomas Alva Edison or Thomas A. Edison. The correct query is "Thomas * Edison" including the Asterisk operator.*
     The asterisk is a wildcard operator that tells Google to treat the asterisk as a placeholder for unknown terms.
     *For example, we can write "rock * roll" to cover all of the following cases: "rock and roll", "rock-n-roll", "rock 'n roll" and "rock & roll".*

2. *"Thomas * Edison"*

- Feedback: *That's it! You are great!*
  The asterisk is a wildcard operator that tells Google to treat the asterisk as a placeholder for unknown terms.
  *For example, we can write "rock * roll" to cover all of the following cases: "rock and roll", "rock-n-roll", "rock 'n roll" and "rock & roll".*

3. *Thomas Alva A. Edison*
   - Feedback: *I can see why one would choose this, but Google returns all webpages that contain all four keywords, Thomas, Alva, A. and Edison, or at least most of them. The correct answer is "Thomas * Edison" including the Asterisk operator.*
     The asterisk is a wildcard operator that tells Google to treat the asterisk as a placeholder for unknown terms.
     *For example, we can write "rock * roll" to cover all of the following cases: "rock and roll", "rock-n-roll", "rock 'n roll" and "rock & roll".*

Range operator (multiple-choice question): *So far, we've found that this query "Thomas Edison" OR "Nikola Tesla" returns the webpages that contain at least one of the names. Now, we would like to find websites that contain the top 3 to 10 facts about Thomas Edison or Nikola Tesla. Important - the query should also return any number in between 3 and 10. Which of the queries below is correct?*

1. *"Thomas Edison" OR "Nikola Tesla" "Top 3-10 facts"*
   - Feedback: *Unfortunately, this is a wrong answer. The correct answer is "Thomas Edison" OR "Nikola Tesla" "Top 3..10 facts".*
     The second part of the query contains ".." which is a range operator. 3..10 denotes a range of numbers from 3 to 10. Then Google will return all webpages that match at least one of the names and "Top X facts" in which X ranges from 3 to 10.
     The quotation marks are important. If they are missing, this query also returns webpages containing either "Top" or "facts" and a number between 3 to 10, for example, "7 fun facts about Nikola Tesla".

2. *"Thomas Edison" OR "Nikola Tesla" "Top 3 TO 10 facts"*
   - Feedback: *Unfortunately, the answer is incorrect. The capitalized "TO" in the second part of the query isn't a Google operator. The correct answer is "Thomas Edison" OR "Nikola Tesla" "Top 3..10 facts".*
     The second part of the query contains ".." which is a range operator. 3..10 denotes a range of numbers from 3 to 10. Then Google will return all webpages that match at least one of the names and "Top X facts" in which X ranges from 3 to 10.
     The quotation marks are important. If they are missing, this query also returns webpages containing either "Top" or "facts" and a number between 3 to 10, for example, "7 fun facts about Nikola Tesla".

3. *"Thomas Edison" OR "Nikola Tesla" Top 3..10 facts*
    - Feedback: *Your selected answer is a good choice but not what we are looking for. The correct answer is "Thomas Edison" OR "Nikola Tesla" "Top 3..10 facts".*
      The second part of the query contains ".." which is a range operator. 3..10 denotes a range of numbers from 3 to 10. Then Google will return all webpages that match at least one of the names and "Top X facts" in which X ranges from 3 to 10.
      The quotation marks are important. If they are missing, this query also returns webpages containing either "Top" or "facts" and a number between 3 to 10, for example, "7 fun facts about Nikola Tesla".

4. *"Thomas Edison" OR "Nikola Tesla" "Top 3..10 facts"*
    - Feedback: *You got it right! This query is correct.*
      The second part of the query contains ".." which is a range operator. 3..10 denotes a range of numbers from 3 to 10. Then Google will return all webpages that match at least one of the names and "Top X facts" in which X ranges from 3 to 10.
      The quotation marks are important. If they are missing, this query also returns webpages containing either "Top" or "facts" and a number between 3 to 10, for example, "7 fun facts about Nikola Tesla".

The site operator (free-text question): *It's time to think about how we can search in a specific domain (e.g. .at) or on a specific site (e.g. youtube.com). So, how would you extend the query "Thomas Edison" OR "Nikola Tesla", to receive webpages only from Austria with the extension .at?*

1. *Correct answer*
    - Feedback: *Fantastic! By using site: we can narrow our results to a specific domain or website.*
      Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

2. *Not using a colon*
    - Feedback: *You're on the right track, but not there yet. The query is almost correct. You just need to use a colon : just right after the word site. Thus, the correct query is "Thomas Edison" OR "Nikola Tesla" site:.at*
      Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

3. *Adding extra space*

   - Feedback: *You're on the right track, but not there yet. The query is almost correct. You must not add an empty space before or after the colon. The correct query is "Thomas Edison" OR "Nikola Tesla" site:.at*
     Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

4. *Having space and colon problem*

   - Feedback: *You're on the right track, but not there yet. The query is almost correct. You just need to use a colon after site and also remove extra spaces. The correct query is "Thomas Edison" OR "Nikola Tesla" site:.at*
     Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

5. *Not using the site operator*

   - Feedback: *Unfortunately not. We need to use the site operator to narrow down the results to a specific domain or website. We need to add site:.at to the query. Thus, the correct query is "Thomas Edison" OR "Nikola Tesla" site:.at*
     Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

6. *Asking for help or a hint*

   - Feedback: *That's okay, I'll show you how to use the site operator to narrow down the results to a specific domain or website. Therefore, we need to add site:.at to the query. Thus, the correct query is "Thomas Edison" OR "Nikola Tesla" site:.at*
     Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.

7. *Other cases*

- Feedback: *Unfortunately, your formulated query is not correct, but that's okay. I'll show you how to use the site operator to narrow down the results to a specific domain or website. Therefore, we need to add site:.at to the query. Thus, the correct query is "Thomas Edison" OR "Nikola Tesla" site:.at Regarding the site operator, you need to remember: 1) It's used to narrow down the results to a specific domain or a website. 2) You need to use the term site: (the word site followed by a colon ":" without any space) 3) The website or the domain is written after the colon (without any spaces): site:.at or site:youtube.com.*

Filetype operator (multiple-choice question): Now suppose we want to search for a presentation (.ppt) about Nikola Tesla that is available on an educational website (domain: .edu). Which of the queries below is the correct one?

1. *"Nikola Tesla" site=.edu file=ppt*
   - Feedback: *Unfortunately, your answer isn't correct. First, file= isn't a valid operator in Google. You need to use filetype:ppt instead. And for the site: operator, you also need to use a colon : instead of the equation symbol =. Thus, the correct query is: "Nikola Tesla" site:.edu filetype:ppt. The filetype operator tells Google to search for a specific filetype. In this example, we are looking for a PowerPoint presentation, which has the extension .ppt. Therefore, we use "filetype:ppt". Similar to the site operator, we need to use a colon after the word filetype:, but we must not use an extra space before and after the colon.*

2. *"Nikola Tesla" site:.edu file:ppt*
   - Feedback: *Unfortunately, your answer isn't correct. file: isn't a valid operator in Google. You need to use filetype:ppt instead. However the site: operator is correct. Thus, the correct query is: "Nikola Tesla" site:.edu filetype:ppt The filetype operator tells Google to search for a specific filetype. In this example, we are looking for a PowerPoint presentation, which has the extension .ppt. Therefore, we use "filetype:ppt". Similar to the site operator, we need to use a colon after the word filetype:, but we must not use an extra space before and after the colon.*

3. *"Nikola Tesla" site=.edu filetype=ppt*
   - Feedback: *This query is almost correct . To get what we're looking for, the "=" character must be replaced by a colon : in both site and filetype. Thus, the correct query is: "Nikola Tesla" site:.edu filetype:ppt The filetype operator tells Google to search for a specific filetype. In this example, we are looking for a PowerPoint presentation, which has the extension .ppt. Therefore, we use "filetype:ppt". Similar to the site operator, we need to use a colon after the word filetype:, but we must not use an extra space before and after the colon.*

4. *"Nikola Tesla" site:.edu filetype:ppt*

   - Feedback: *Awesome, your answer is right! Google returns all webpages of the domain .edu that contain Nikola Tesla and a presentation file .ppt.*
     The filetype operator tells Google to search for a specific filetype. In this example, we are looking for a PowerPoint presentation, which has the extension .ppt. Therefore, we use "filetype:ppt". Similar to the site operator, we need to use a colon after the word filetype:, but we must not use an extra space before and after the colon.

Intitle operator (multiple-choice question): With the site and filetype operators we can narrow down the results. To get more precise results, we can look for our keywords in specified parts of the results found for example in their titles or even in their URLs. Which query will return webpages whose titles contain "Top 3..10 facts about Nikola Tesla".

1. *title:"Top 3..10 facts about Nikola Tesla"*

   - Feedback: *Unfortunately your selected answer is not correct as title: is not a valid Google command. If you click on the query, you will see the result in which Google will say "Your search - title:"Top 3..10 facts about Nikola Tesla" - did not match any documents." Thus the correct answer is: intitle:"Top 3..10 facts about Nikola Tesla"*
     The intitle: is an operator that tells Google to look for a keyword only in the titles of the webpages. If we are looking for a word in the websites' titles, we need to use intitle: and the keyword, e.g. "intitle:Tesla" (again with : and without any space).

2. *intitle:"Top 3..10 facts about Nikola Tesla"*

   - Feedback: *Perfect! You make it look easy!*
     The intitle: is an operator that tells Google to look for a keyword only in the titles of the webpages. If we are looking for a word in the websites' titles, we need to use intitle: and the keyword, e.g. "intitle:Tesla" (again with : and without any space).

3. *inheader:"Top 3..10 facts about Nikola Tesla"*

   - Feedback: *Unfortunately your selected answer is not correct as inheader: is not a valid Google command. If you click on the query, you will see the result in which Google will say "Your search - inheader:"Top 3..10 facts about Nikola Tesla" - did not match any documents." Thus the correct answer is: intitle:"Top 3..10 facts about Nikola Tesla"*
     The intitle: is an operator that tells Google to look for a keyword only in the titles of the webpages. If we are looking for a word in the websites' titles, we need to use intitle: and the keyword, e.g. "intitle:Tesla" (again with : and without any space).

Exclude operator (free-text question): Now the question is, what do we need to add to the query to remove the YouTube videos?

1. *Correct answer*
   - Feedback: *I'm impressed! You already know how to exclude pages :) The exclude operator, - (hyphen), is used to exclude a word or reverse the effect of an operator. In general, by putting a hyphen (-) before each word or operator, we reverse its effect. For example, the search query, Tesla -car, will return webpages that contain "Tesla" and but NOT the word "car". In our preceding example, we used it to reverse the effect of the site: operator.*
     Now let's see how the exclude operator affects the results. You can see the difference by clicking on the queries:
     1) intitle:"Top 3..10 facts about Nikola Tesla"
     2) intitle:"Top 3..10 facts about Nikola Tesla" -site:youtube.com (YouTube's videos are excluded.)
2. *Incorrect answer*
   - Feedback: *Unfortunately our answer is wrong because you have not applied the exclude operator correctly. The exclude operator, - (hyphen), is used to exclude a word or reverse the effect of an operator. In general, by putting a hyphen (-) before each word or operator, we reverse its effect. For example, the search query, Tesla -car, will return webpages that contain "Tesla" and but NOT the word "car". In our preceding example, we used it to reverse the effect of the site: operator.*
     Now let's see how the exclude operator affects the results. You can see the difference by clicking on the queries:
     1) intitle:"Top 3..10 facts about Nikola Tesla"
     2) intitle:"Top 3..10 facts about Nikola Tesla" -site:youtube.com (YouTube's videos are excluded.)
3. *Asking for a hint*
   - Feedback: *It's okay that you don't know how to exclude youtube from the results. I'll show it to you. The exclude operator, - (hyphen), is used to exclude a word or reverse the effect of an operator. In general, by putting a hyphen (-) before each word or operator, we reverse its effect. For example, the search query, Tesla -car, will return webpages that contain "Tesla" and but NOT the word "car". In our preceding example, we used it to reverse the effect of the site: operator.*
     Now let's see how the exclude operator affects the results. You can see the difference by clicking on the queries:
     1) intitle:"Top 3..10 facts about Nikola Tesla"
     2) intitle:"Top 3..10 facts about Nikola Tesla" -site:youtube.com (YouTube's videos are excluded.)
4. *All other cases*

- Feedback: *The exclude operator, - (hyphen), is used to exclude a word or reverse the effect of an operator. In this example, we used it to reverse the effect of the site: operator. In general, by putting a hyphen (-) before each word or operator, we reverse its effect. For example, the search query, Tesla -car, will return webpages that contain the word "Tesla" but NOT the word "car".*
  Now let's see how the exclude operator affects the results. You can see the difference by clicking on the queries:
  1) intitle:"Top 3..10 facts about Nikola Tesla"
  2) intitle:"Top 3..10 facts about Nikola Tesla" -site:youtube.com (YouTube's videos are excluded.)