Lisbora Likaj, Bsc

Ethical AI in Recruitment

Master's Thesis
to achieve the university degree of
Master of Science
Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor
Bernhard Wieser, Assoc.Prof. Mag.phil. Dr.phil.

Institute of Interactive Systems and Data Science
Head: Frank Kappe, Univ.-Prof. Dipl.-Ing. Dr.techn.

Graz, September 2024

Lisbora Likaj, Bsc

Ethische KI in der Personalbeschaffung

Masterarbeit
zur Erlangung des akademischen Grades eines
Diplom-Ingenieur
Masterstudium: Software Engineering and Management

eingereicht an der

Technische Universität Graz

Betreuer
Bernhard Wieser, Assoc.Prof. Mag.phil. Dr.phil.

Institute of Interactive Systems and Data Science
Vorstand: Frank Kappe, Univ.-Prof. Dipl.-Ing. Dr.techn.

Graz, September 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 master‘s 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

Acknowledgment

I would like to express my sincere gratitude to my thesis supervisor, Professor Bernhard Wieser, for his invaluable advice and guidance throughout this journey. I also extend my gratitude to AVL, particularly the e1 and HR departments, for their involvement and support during the experimental phase of this thesis. A special thanks to my colleagues at Creators Expedition for their support and encouragement, both in this thesis work and in my career development.

I am grateful to my parents, Dalina and Mustafa, and to my siblings, Lorena and Leo, for their unconditional love and support at every stage of my life.

Finally, I dedicate this master's degree and thesis to my partner, Besjon, whose support and love have been a constant source of strength not just during this thesis, but in everything I do.

Abstract

The rise of Artificial Intelligence (AI), particularly Large Language Models (LLMs), has transformed our interactions with technology and spread through various industries, including recruitment. This thesis explores the application of LLMs, specifically ChatGPT and Microsoft Copilot, in automating the resume screening and shortlisting process. As AI becomes more integrated into recruitment, it is essential to evaluate both its accuracy and ethical implications, particularly concerning fairness in terms of gender, age, and ethnicity. Through a practical experiment, this thesis bridges the gap between theoretical discussions of ethics and empirical data by holistically assessing LLMs' performance and fairness. Additionally, a novel retrieval-augmented generation (RAG)-based recruiting platform is evaluated to enhance decision-making in AI-driven recruitment. The experiment involves 102 resumes, screened and shortlisted by both LLMs and human recruiters and hiring managers for 10 job openings within a real company. Two main goals are pursued: first, to evaluate the accuracy of both LLMs using two methods—comparison with the recruiters' shortlist and with the hiring managers' final interview list; second, to assess fairness by analyzing the models against aggregated demographic data from the resume pool. Findings indicate that both LLMs but especially ChatGPT, aligns closely with hiring managers' final interview decisions in terms of accuracy. However, quantifying fairness across age, gender, and ethnicity proved more challenging, with an ongoing need to identify appropriate benchmarks for assessing fairness. A critical question remains in balancing accuracy with fairness in AI recruitment systems. This research contributes to the growing discourse on AI ethics in recruitment, highlighting the need for further studies to address fairness in AI hiring processes while preserving accuracy.

Kurzfassung

Der Aufstieg der Künstlichen Intelligenz (KI), insbesondere der Large Language Models (LLMs), hat unsere Interaktion mit der Technologie verändert und sich in verschiedenen Branchen, einschließlich der Personalbeschaffung, verbreitet. In dieser Arbeit wird die Anwendung von LLMs, insbesondere ChatGPT und Microsoft Copilot, zur Automatisierung der Lebenslaufprüfung und der Vorauswahl von Bewerbern untersucht. Mit der zunehmenden Integration von KI in die Personalbeschaffung ist es wichtig, sowohl ihre Genauigkeit als auch ihre ethischen Implikationen zu bewerten, insbesondere im Hinblick auf Fairness bezüglich Geschlecht, Alter und ethnischer Zugehörigkeit. Durch ein praktisches Experiment überbrückt diese Arbeit die Lücke zwischen theoretischen Diskussionen über Ethik und empirischen Daten, indem sie die Leistung und Fairness der LLMs ganzheitlich bewertet. Darüber hinaus wird eine neuartige, auf retrieval-augmented generation (RAG) basierende Rekrutierungsplattform evaluiert, um die Entscheidungsfindung bei der KI-gesteuerten Rekrutierung zu verbessern. Das Experiment umfasst 102 Lebensläufe, die sowohl von LLMs als auch von menschlichen Recruitern und Personalverantwortlichen für 10 offene Stellen in einem realen Unternehmen gesichtet und in die engere Wahl gezogen wurden. Zwei Hauptziele werden verfolgt: Erstens die Bewertung der Genauigkeit beider LLMs durch zwei Methoden – den Vergleich mit der Vorauswahlliste der Personalverantwortlichen und der endgültigen Interviewliste der einstellenden Manager; zweitens die Bewertung der Fairness durch die Analyse der Modelle anhand der aggregierten demografischen Daten aus dem Lebenslaufpool. Die Ergebnisse zeigen, dass beide LLMs, insbesondere ChatGPT, in Bezug auf die Genauigkeit den endgültigen Interviewentscheidungen der Personalverantwortlichen eng entsprechen. Die Quantifizierung der Fairness in Bezug auf Alter, Geschlecht und ethnische Zugehörigkeit erwies sich jedoch als schwieriger, sodass weiterhin geeignete Benchmarks für die Bewertung der Fairness ermittelt werden müssen. Eine kritische Frage bleibt das Gleichgewicht zwischen Genauigkeit und Fairness in KI-Rekrutierungssystemen. Diese Forschung trägt zum wachsenden Diskurs über KI-Ethische Fragen in der Personalbeschaffung bei und unterstreicht den Bedarf an weiteren Studien, die sich mit Fairness in KI-Einstellungsprozessen befassen und gleichzeitig die Genauigkeit wahren.

Contents

List of Figures

List of Tables

1. Introduction

The rise of Artificial Intelligence (AI), especially with Large Language Models (LLMs) such as ChatGPT and Microsoft Copilot, has revolutionized our interactions with technology. These advancements are not limited to just specific industries; they are increasingly becoming integral to our daily lives. From education to work and leisure, AI technologies are being utilized for a wide range of tasks, making them more accessible and pervasive than ever before. This growing accessibility brings AI closer to both individuals and companies, making it crucial to thoroughly examine its applications across various departments and domains, including recruitment.

The significance of exploring LLMs in recruitment lies in their potential to transform hiring practices. As AI becomes an established resource in various fields, it is essential to analyze, experiment with, and understand its implications for recruitment. Conversations about AI have shifted from speculative to practical, as this technology is now tangible and available for everyday use. As individuals increasingly use AI for personal tasks, it's becoming a notable trend for Human Resource (HR) departments to adopt these technologies for candidate screening and hiring decisions.

The increasing integration of AI into the labor market has raised important questions about its impact on HR management. According to Rigotti et al. (2023), AI systems are becoming more prevalent in HR, with many employers relying on them for tasks such as resume screening. Research by Gan et al. (2024) highlights how LLMs can significantly enhance recruitment processes by providing scalable and secure solutions for managing large volumes of resumes.

However, while AI and LLMs offer promising advancements, they also pose challenges, particularly concerning performance, quality, fairness, and bias. Li et al. (2023) discusses how LLMs can enhance job recommendation systems by enriching user resumes with external knowledge and processing capabilities. Yet, the ethical implications of these technologies must be carefully considered.

1.1. Research Questions

As AI and LLMs become increasingly prevalent in recruitment, it is crucial to evaluate their performance, quality, and ethical impact. This thesis explores the potential of LLMs, such as ChatGPT and Microsoft Copilot, to enhance the efficiency and accuracy of resume screening while ensuring fairness and minimizing bias. While AI can improve the speed and effectiveness of recruitment processes, it is equally important to assess the quality of outcomes and address ethical considerations. In this thesis, ethical considerations specifically refer to fairness criteria in terms of gender, age, and ethnicity. Rigotti et al. (2023) highlights the risks of AI perpetuating bias and discrimination, underscoring the necessity of examining these issues. This research focuses on understanding potential biases and fairness issues associated with LLMs in recruitment, emphasizing the need for a comprehensive evaluation of both performance and ethical implications.

To effectively assess AI-based recruitment in terms of performance, efficiency, and ethical quality criteria—such as fairness and bias—this thesis addresses the following research questions:

1.1.1. RQ1: What approach is best suited for the assessment of AI-based recruitment in relation to performance and efficiency criteria?

As AI emerges as a viable alternative for HR departments, it is crucial to assess its effectiveness in recruitment. This involves testing various approaches and comparing AI-based recruiting with traditional methods in terms of performance and quality of results. Performance criteria is evaluated through an experiment with a real-world company, comparing AI-driven recruitment with human recruiters.

1.1.2. RQ2:How can AI-based recruitment processes be assessed in relation to ethical criteria?

Beyond performance and quality of results, the ethical implications of AI-based recruitment are crucial. This research investigates how to evaluate the fairness and non-discriminatory nature of AI recruiting processes. The aim is to provide insights into ensuring that AI recruitment offers equal opportunities for all applicants, regardless of gender, age, or ethnicity.

1.2. Main Contributions

Despite extensive research on AI in recruitment, a significant gap remains in analyzing both the performance and ethical implications, particularly with regard to LLMs.

In this thesis, I adopt a comprehensive approach to exploring the use of LLMs in recruitment. I assess not only their accuracy and performance but also their quality and the ethical considerations related to fairness. This involves a practical experiment conducted in collaboration with the HR department of a real-world automotive company, allowing for a comparative analysis of traditional human recruiting versus AI-driven methods.

First, I offer an overview of AI, LLMs and the ethical concerns they raise, with a focus on models such as ChatGPT and Microsoft Copilot in the context of recruitment. I then evaluate these models through an experiment, assessing their performance and adherence to fairness criteria, including gender, age, and ethnicity. Unlike existing studies that often discuss ethical issues without empirical data due to the black-box nature of AI, my research provides experimental benchmarks using both recruiter and LLM-generated data.

Secondly, I have developed an innovative recruiting platform utilizing the retrieval-augmented generation (RAG) framework. To my knowledge, research applying the RAG framework in recruitment is scarce. This system automates aspects of the recruitment process, providing hiring managers with a clearer overview of automated resume screening and shortlisting, thereby supporting more informed decision-making.

In summary, my contributions are threefold: (1) providing a holistic evaluation of LLMs in recruitment, focusing on performance and ethical considerations; (2) bridging the gap between theoretical ethical discussions and empirical data through practical experimentation; and (3) developing and evaluating a novel RAG-based recruiting platform aimed at improving decision-making in AI-driven recruitment.

1.3. Thesis Outline

The second chapter of this thesis reviews related work, offering context and defining key concepts. It covers the definitions of AI and LLMs, existing research on the use of AI in recruitment, and the associated ethical challenges. The third chapter outlines the experimental design, materials, and methods employed to address the research questions. The fourth chapter presents and analyzes the experimental results. Chapter five focuses on the development and discussion of an automated recruiting system based on RAG and LLM technologies. The final chapter reflects on the findings, outlines the study's limitations, and provides recommendations for future research, with an emphasis on further exploring AI's performance and ethical implications in recruitment.

Disclaimer: Please note that throughout this thesis, the terms "AI" and "LLM" may be used interchangeably.

2. Background and Related Work

This section introduces key concepts and provides an overview of the topics discussed throughout this thesis. It begins with an introduction to LLMs, detailing the two LLM agents utilized in this thesis: ChatGPT and Microsoft Copilot. Following this, the section presents an overview of Retrieval-Augmented Generation (RAG), which is used and discussed in Chapter 5. Next, the section transitions to an overview of recruiting and LLMs by comparing traditional and automated recruitment processes, highlighting state-of-the-art solutions and exploring the potential of AI and LLMs in recruitment. Finally, the section addresses ethical considerations by examining potential ethical issues in AI-driven recruiting and explaining the fairness criteria considered in this thesis.

2.1. Large Language Models (LLMs)

Traditional machine learning approaches focused heavily on feature engineering, typically designed for single downstream tasks. The process involved training models for specific use cases, relying on manually selected features. Around 2015, deep learning began to gain significant attention due to its ability to automatically learn features during training through parameter updates and the introduction of backpropagation (LeCunn et al., 2015; Schmidhuber, 2014). Since then, AI has evolved rapidly.

Today, the buzz surrounding LLMs is impossible to ignore. These models are built to comprehend written input and produce text that closely resembles human language, making it difficult to differentiate from content written by people. They achieve this by being trained on massive amounts of structured data, with billions of parameters, allowing them to be utilized for multiple downstream tasks (J. Kaplan et al., 2020; Minaee et al., 2024). LLMs are capable of handling a wide range of language-related tasks, including machine translation, text summarization, text generation, and answering complex queries. According to Fletcher and Nielsen (2024), ChatGPT was one of the most widely used AI instruments in 2024, with usage rates nearly twice that of Google Gemini or Microsoft Copilot.

LLMs are essentially probabilistic models that mine a vast array of statistical patterns and capture the complex semantics of natural language text corpus. A crucial aspect of working with LLMs is the temperature setting, which ranges from 0 to 1. A temperature closer to 0 produces more creative responses, while a higher temperature allows the model to generate more deterministic and precise outputs. The LLM architecture is driven by Transformers (Vaswani et al., 2017) and word embeddings, which form the backbone of these advanced models.

2.1.1. Word Embedding

Word embeddings convert textual words into numerical vectors (Mikolov, Chen, et al., 2013), allowing machine learning models to interpret text corpora mathematically. This process involves splitting the sequence into tokens, where each token is mapped into a higher dimensional vector space. The embeddings capture the semantics of words based on their context and meaning. In simple terms, words with similar meanings or that are used in a similar context tend to have similar representations (Mikolov, Chen, et al., 2013). Various techniques exist for generating embeddings, with Word2Vec (Mikolov, Sutskever, et al., 2013) and GloVe (Pennington et al., 2014) being among the most well-known methods. Many natural language processing solutions are built upon these foundational vector models .

2.1.2. Transformers Architecture

Self-attention mechanism by Vaswani et al. (2017) revolutionized natural language processing by enabling the processing of textual streams more effectively. The attention is applied to input queries to identify relationships between tokens within a sequence, regardless of their position. This allows transformers to compute an importance score for each token relative to other tokens in the sequence, leading to a better understanding of context (Vaswani et al., 2017). This enables better insight into which tokens contribute most to the meaning of a sentence (Sampath, 2024). As a result, more efficient methods for handling contextualized textual data can be developed. This architecture is fundamental to the creation of models like GPT and BERT (Ahmad et al., 2022).

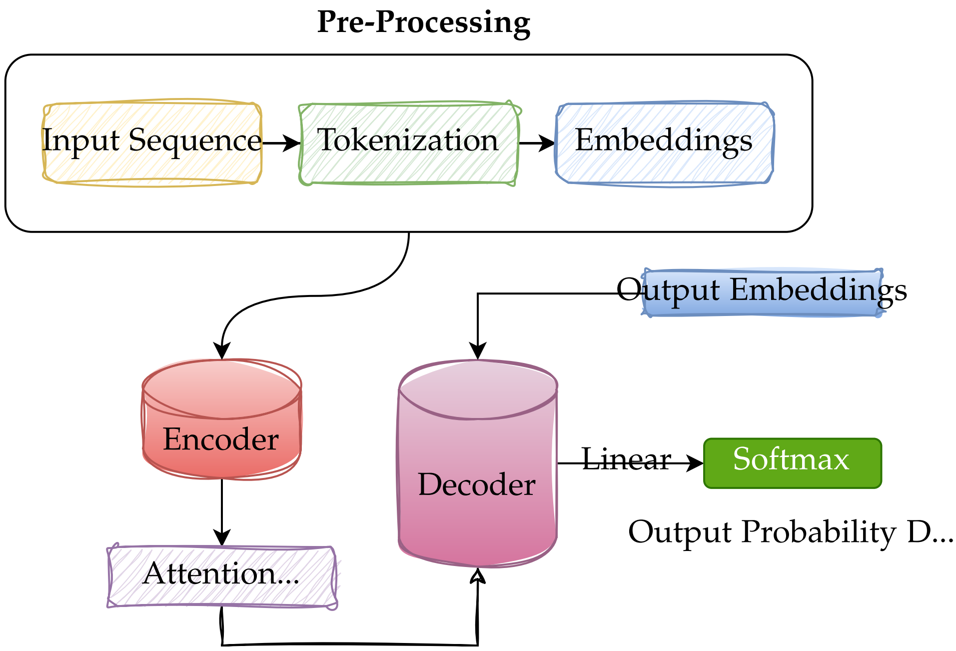

Figure 2.1.: High-Level Overview of Transformer Architecture in Common LLMs

2.1.3. Semantic Similarity Search

Word embeddings have significantly influenced the development of semantic search. It allows the identification of similar text based not just on keywords, but on context and meaning as well. Over the years, various methods have been introduced to enhance this approach (Chandrasekaran & Mago, 2021). One of the most widely used techniques is cosine similarity, which takes two vectors representing words as input and calculates the cosine of the angle between them to determine a similarity score. A score of 1 indicates a perfect match, while a score of 0 signifies no similarity.

$$cos(\theta) = \frac{A \cdot B}{||A||||B||} \tag{2.1}$$

where A and B represent vectors corresponding to two different words. The cosine value is computed by taking the dot product of these vectors and dividing it by the product of their magnitudes (norms). However, recent research suggests that cosine similarity should not be applied indiscriminately, and it highlights alternative methods that may be more appropriate depending on the context (Steck et al., 2024).

2.1.4. Models Landscape

Recent years have seen the emergence of various AI models, including open-source contributions like BERT (Devlin et al., 2019), LLaMA, and Mistral 7B (Jiang et al., 2023), as well as proprietary solutions such as OpenAI's GPT (OpenAI, 2022), Google's Gemini (Google, 2024), and Perplexity (Perplexity, 2024). According to (Cardillo, 2024), which ranks the 21 best models as of June 2024, GPT-4, introduced by OpenAI, is the top-performing model across many downstream tasks. Although the number of parameters in GPT-4 remains undisclosed, it continues to be a benchmark in AI performance. Alternatively, the open-source model Mistral 7B, with approximately 7.3 billion parameters, offers commendable performance despite its smaller size (Jiang et al., 2023).

This thesis focuses on two popular models: ChatGPT and Microsoft Copilot.

ChatGPT, introduced by OpenAI in 2022, is a conversational AI built to engage in dialogue, address follow-up questions, acknowledge errors, question faulty assumptions, and decline inappropriate requests (OpenAI, 2022). Widely used in daily life, ChatGPT was trained using Reinforcement Learning from Human Feedback (RLHF), involving supervised fine-tuning where human AI trainers played both sides of the conversation. A reward model was developed by ranking multiple AI-generated responses, allowing for iterative improvements. OpenAI has been transparent about ChatGPT's limitations, such as producing plausible but incorrect answers, being sensitive to input phrasing, and sometimes generating overly verbose responses. The model may also misinterpret user intent and occasionally respond inappropriately despite efforts to mitigate harmful content (OpenAI, 2022). However, research shows that ChatGPT is able to produce accurate answers, offer personalized real-time feedback, and improve task efficiency by enhancing access to information and streamlining complex learning tasks (Mohammadreza Farrokhnia & Wals, 2024).

ChatGPT Free offers access to the GPT-3.5 model and limited access to GPT-4, along with restricted features like advanced data analysis, file uploads, web browsing, and custom GPTs (Jenkins, 2024). Beyond the free version, ChatGPT Plus provides enhanced GPT access and DALL-E integration (OpenAI, 2024b), ChatGPT Team includes higher message limits and workspace management, and ChatGPT Enterprise offers unlimited access, extended inputs, and advanced administrative controls (Jenkins, 2024; OpenAI, 2022).

Microsoft Copilot, launched by Microsoft in 2023, integrates large language models with data from Microsoft Graph and Microsoft 365 apps to enhance productivity. It is embedded within apps like Word, Excel, PowerPoint, Outlook, and Teams, and includes a feature called Business Chat, which allows users to perform tasks via natural language prompts based on data from meetings, emails, and chats (Spataro, 2023). This thesis specifically focuses on the Business Chat function of Microsoft Copilot.

While Microsoft Copilot offers a free version aimed at sparking creativity, its primary offerings are Copilot Pro, a paid subscription for enhanced features, and Copilot for Microsoft 365, which is designed for both individuals and teams. Additionally, various other versions of Copilot are available for specific use cases (Jenkins, 2024; Spataro, 2023). Microsoft has formed a significant partnership with OpenAI, investing in the organization and incorporating its technology into Microsoft products including copilot (OpenAI, 2019).

2.1.5. Text Generation Process

Large Language Models (LLMs) excel in logical text comprehension and reasoning, producing results comparable to those generated by humans (Malach, 2024).

Text generation in LLMs involves creating responses based on user queries. The architecture of LLMs typically follows an encoder-decoder model, where the encoder processes the input by understanding the underlying context of the tokens, and the decoder synthesizes textual content using the model's pre-trained knowledge. The decoder predicts the next token in the sequence by attending to previous tokens, a process known as auto-regressive next-word prediction (Malach, 2024; Nguyen, 2024). The prediction is made based on the most probable token in a softmax latent space distribution(Pearce et al., 2021).

Despite their powerful content generation capabilities, LLMs have limitations, such as generating *hallucinated* or unreliable information. This issue often arises when the model lacks factual references due to insufficient training data, leading it to infer responses based on the most probable tokens derived from learned patterns. Additionally, LLMs may struggle with complex queries, resulting in ambiguous or vague outputs. These limitations are significant concerns in their application across various industries (Kalyan, 2024). In the context of recruiting, particularly for resume screening and shortlisting, such limitations could lead to biased or inaccurate evaluations, affecting the fairness and effectiveness of the hiring process.

2.2. Retrieval Augmented Generator (RAG)

LLMs are capable of achieving state-of-the-art results in many downstream tasks, but they have limited access to external knowledge. Their effectiveness is constrained by the training data they consume before deployment. Retrieval-Augmented Generation (RAG) addresses this limitation by combining pre-trained parametric models (the Generator) with an external knowledge base (the Retriever), which acts as model memory(Gao et al., 2024; Lewis et al., 2021). This memory is built on dense vector representations of external data sources, allowing the model to produce more specific and factual results. Additionally, RAG reduces the need for extensive fine-tuning for specific tasks, leveraging the external knowledge base to enhance the model's performance.

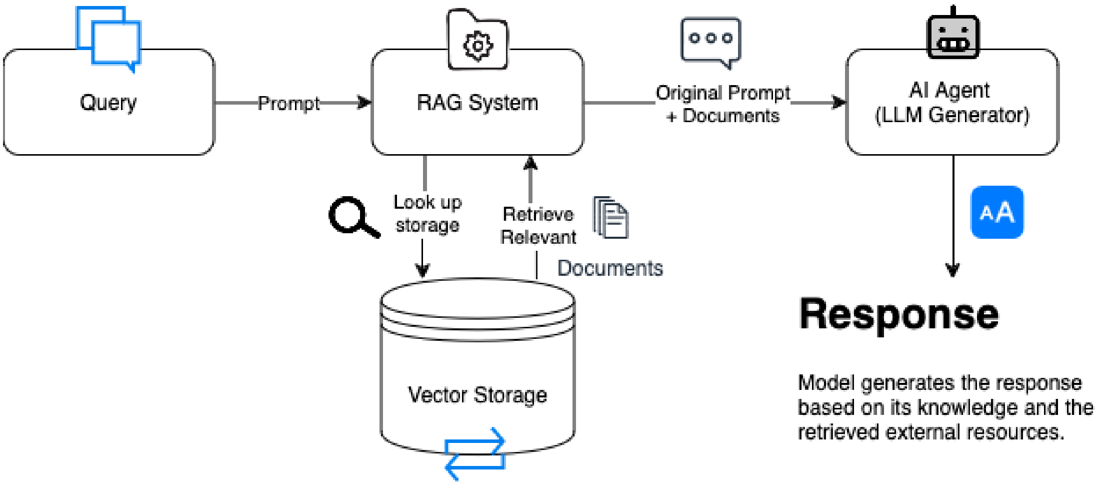

Figure 2.2.: Illustration of the Principle Behind RAG

2.2.1. Vector Databases and Retrieval

A retrieval component searches through a knowledge base or a large *vector* database to find documents that semantically match the user's query (Han et al., 2023). This process involves comparing the embedding of the query with embeddings stored in the database and returning the most similar documents. These documents are then combined with a pre-trained model as additional context, enhancing the quality of the generated response by grounding it in factual data.

Vector databases are specifically designed to efficiently store and query high-dimensional vectors, which are numerical representations of textual corpora (Biswas & Das, 2024). They are optimized for fast indexing and similarity matching between vectors, allowing for the quick identification of the most relevant documents for a given query. Popular implementations of vector databases include FAISS (Douze et al., 2024), Chroma DB (Chroma, 2024), and Annoy (Bernhardsson, 2018).These implementations vary in processing speed for storing, indexing, and retrieving data, while all offer scalability.

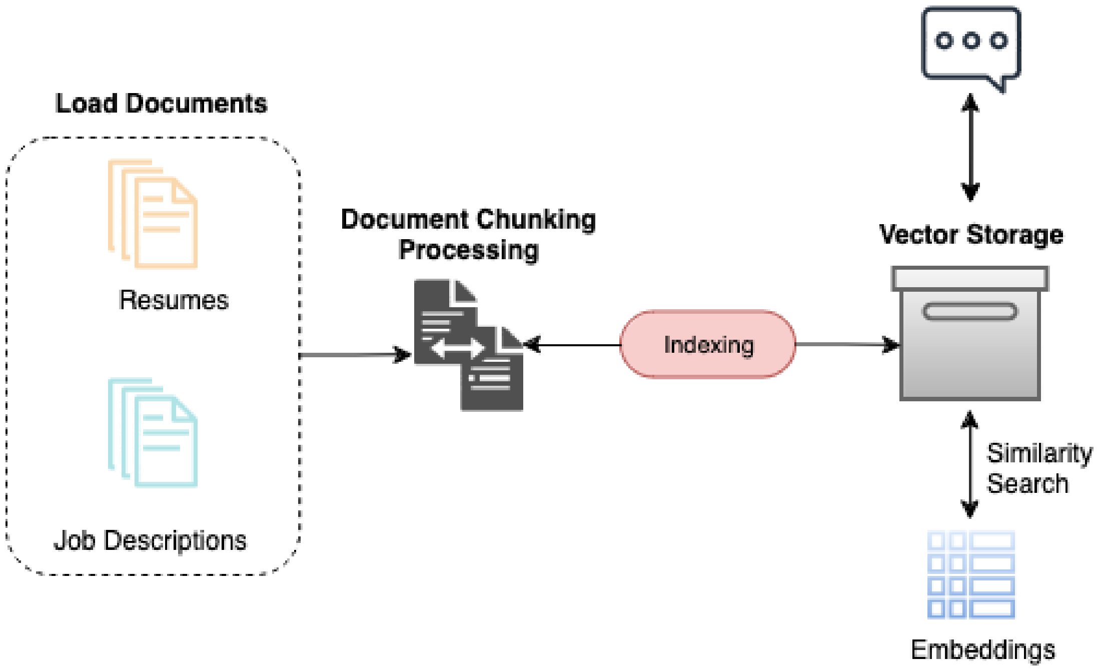

Figure 2.3.: Vector Storage and Retrieval Process in AI Recruiting

2.2.2. General Pipeline

The original RAG model (Lewis et al., 2021) is designed to be trained in an end-to-end fashion. However, due to the challenges and significant computational resources required for this approach, a more streamlined architecture is often employed. This simplified version uses pre-trained LLM agents for content generation and plug-and-play vector databases for data storage and retrieval. In Chapter 5, the RAG model is applied to illustrate how LLMs can impact the recruitment process.

2.3. Recruiting and LLMs

AI refers to a system's capability to accurately interpret external data, learn from it, and apply this knowledge to accomplish specific tasks and goals through adaptive processes (A. Kaplan & Haenlein, 2019). Global Industry Analysts (2022) show that the global AI market was valued at USD 95.9 billion in 2022 and is projected to expand significantly, reaching USD 276.3 billion by 2026 . This rapid growth underscores AI's increasing influence across various industries, including recruitment.

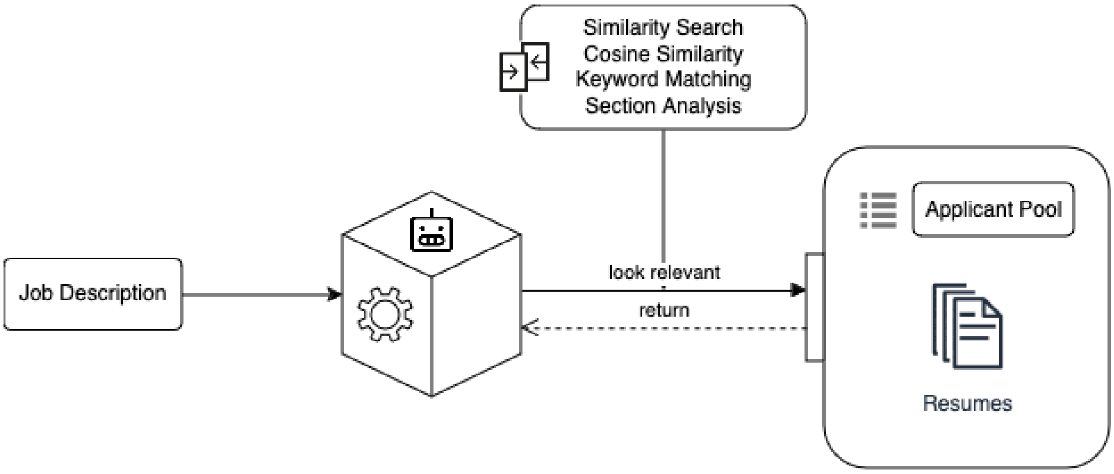

Figure 2.4.: Automated Recruiting Process: Candidate Matching Using Text Algorithms vs. LLM Reasoning

2.3.1. Traditional and Automated Recruiting Process

The recruitment process comprises several stages, including sourcing candidates, screening resumes, evaluating shortlists, conducting interviews, and finally, extending offers and onboarding. This thesis focuses specifically on the resume screening and shortlisting phase.

Resume screening is a crucial step in recruitment, often involving labor-intensive manual work, particularly in large organizations with a high number of job applications. Traditionally, recruiters sift through hundreds of resumes, shortlisting and interviewing candidates to find the most suitable hire. According to The Ladders (2018) recruiters typically spend only 6 to 7 seconds on an initial resume review, as revealed by an eye-tracking study analyzing how recruiters allocate their attention across different resume sections. Introducing AI into resume screening can significantly reduce the time and effort involved, making it a prime area for AI adoption within the recruitment process. Given that human recruiters continue to assess candidates in subsequent phases, AI-driven resume screening presents minimal risk while offering considerable efficiency gains.

In conventional resume screening, HR professionals invest extensive hours reviewing resumes, searching for keywords, and manually aligning candidate qualifications with job descriptions (M. Roy & Sawant, 2024). This method is not only time-consuming but also susceptible to human error, leading to potential inconsistencies in candidate evaluations (M. Roy & Sawant, 2024). Historically, recruitment has evolved from classified ads and agencies to digital platforms like LinkedIn, which have made the process more accessible and efficient. The advent of Applicant Tracking Systems (ATS) marked an early move toward automation, streamlining job postings, resume screening, and candidate tracking (Gotoro, 2024).

Research has investigated numerous methods to automate recruiting, particularly resume screening. One approach by P. K. Roy et al. (2020) involves machine learning (ML) classifiers to categorize resumes, then ranking candidates using content-based recommendation systems. Techniques like cosine similarity and k-Nearest Neighbors (k-NN) are employed to match resumes with job descriptions, enabling the selection of the most suitable candidates.

Further exploration into text similarity measures, such as Cosine, Sqrt-Cosine, and Improved Sqrt-Cosine (ISC) similarity, compares automated resume screening results with those of expert hiring managers. While these approaches are promising in terms of imitating recruiter decisions and saving time, there remains a critical gap in evaluating the quality of results, ethical concerns, and potential biases (Alsharef et al., 2023).

The growing use of cosine similarity for aligning resume data with job descriptions is noted by Ambareesh et al. (2024), but these methods often neglect the ethical dimensions of these kind recruitment systems. Similarly, BERT has been implemented for automated resume screening, demonstrating increased precision and speed in candidate selection, but this too often overlooks fairness and bias issues (Deshmukh & Raut, 2024).

Lastly, by integrating NLP techniques like named entity recognition and part-of-speech tagging with ML classifiers such as K-Nearest Neighbors and Support Vector Machines (Tayal et al., 2024), systems have been developed to improve candidate selection accuracy while reducing time and effort. Despite these advancements, challenges remain in addressing both the quality of results and the ethical implications.

2.3.2. Potential of AI

AI and machine learning (ML) have become indispensable instruments for recruiters seeking to streamline and enhance their hiring processes (Northreach, 2023). AI is now integrated into every phase of recruitment (M. Roy & Sawant, 2024):

- **Sourcing Potential Candidates**: AI software platforms, like Hero Hunt ("Official Website - HeroHunt," 2024), are used to identify and source potential applicants, driving data-driven decision-making and reducing repetitive tasks in recruitment (Heymans, 2024). AI chatbots, such as Olivia ("Official Website - Olivia," 2024), Mya (Levingston, 2024), and Talentsoft ("Official Website - Cegid," 2024), also provide HR-related support and employee engagement services. In fact, 92% of HR departments, particularly in the retail sector in the US, now direct new employees to use chatbots for information retrieval (Beckman, 2024).
- **Screening resumes**: AI-powered screening platforms are particularly valuable for large companies that manage vast numbers of applications.

These platforms can efficiently filter candidates, helping recruiters focus on the most qualified individuals. Software platforms like Zoho Recruit ("Official Website - Zoho Recruit," 2024) and Ideal ("Official Website - Ideal," 2024) are amongst top recommended AI-powered software to use for resume screening (Semetaite, 2023)
- **Evaluation of Shortlists**: AI assists in evaluating shortlisted candidates by analyzing their fit based on predefined criteria.
- **Interviewing Process:** AI-driven platforms can schedule interviews, provide initial assessments, and even conduct preliminary interviews using natural language processing capabilities. Platforms like HireVue ("Official Website - HireVue," 2024) offer not only AI-driven candidate screening, but also interview scheduling, and AI assessments.
- **Offer and Onboarding**: AI streamlines the offer and onboarding process by automating document handling and personalized communication. Software like Pymetrics ("Official Website - Pymetrics," 2024) use neuroscience-based assessments and AI to evaluate candidates' emotional and cognitive attributes, matching them to job roles where they are most likely to succeed.

2.3.3. Why Investigate the Role of LLMs in Recruiting?

Since the launch of ChatGPT by OpenAI in 2022, LLMs have become increasingly integrated into daily life, quickly reaching widespread adoption. Within just five days of its release, ChatGPT surpassed one million users, and in May 2024 alone, its website saw 573 million visits (Meer, 2024). As AI and LLMs solidify their role across various domains, including recruitment, it becomes essential to critically examine their implications. Key questions arise: Do LLM models merely accelerate the resume screening process, or do they also maintain the quality of outcomes? Are these models equitable across different demographic categories such as gender, age, and ethnicity? Addressing these questions is crucial.

Various ML and LLM approaches have been developed to automate resume screening and recommendation systems, offering advantages in efficiency and scalability. A 2024 report by McKinsey & Company (2024) highlighted the growing use of generative AI, including LLMs, across multiple business functions. Notably, human resources is the area where the highest percentage of respondents reported cost reductions due to generative AI (McKinsey & Company, 2024). LLMs have shown significant potential in achieving human-like intelligence, sparking a surge in research on LLM-based autonomous agents (Wang et al., 2024).

For example, O'Neal (2024) reported that GenAI has already been integrated into the activities of 15% of HR departments, optimizing performance monitoring procedures. HR teams are leveraging LLMs to automate complex tasks like personnel selection, enabling recruiters to swiftly identify top candidates and focus on more strategic aspects of their roles (O'Neal, 2024). While generative AI and LLMs can automate HR tasks such as job description creation, candidate assessment, and performance management, they also hold the potential to enhance decision-making and reduce biases in hiring and compensation. Although not without flaws, these technologies mark a significant advancement in making HR practices more efficient and data-driven (Bersin, 2023).

Research by Gan et al. (2024) introduced a novel LLM-based agent framework for resume screening, demonstrating a speed increase 11 times faster than traditional manual methods. This framework, utilizing a fine-tuned model, outperforms baseline models like GPT-3.5 in resume summarization and grading, highlighting its potential to significantly enhance recruitment processes.

Despite the emphasis on the implementation of AI and LLMs in various domains, including recruitment, the existing literature often overlooks the implications in terms of performance, result quality, and ethical concerns. When these implications are discussed, they are frequently addressed at a theoretical level. This thesis argues that the implementation and implications of LLMs cannot be separated; understanding the implications is essential before discussing implementation. Through an experiment conducted in a real-world company, this work aims to bridge this gap by analyzing the implications of LLMs in terms of performance and fairness (check Section 3.5 and Chapter 4 for more details).

An extensive research in the literature by Hunkenschroer and Luetge (2022) underscores the ethical implications of AI in recruitment, particularly during the initial resume screening phase. This study highlights the importance of addressing ethical concerns, such as fairness and bias, when implementing AI in recruitment. A recent resume audit study involving ChatGPT (specifically GPT-4) assessed bias in ranking resumes. The study compared a basic resume with an enhanced version featuring disability-related achievements, such as a leadership award, scholarship, panel presentation, and membership. The findings revealed that GPT-4 exhibited prejudice against the enhanced resumes, though this bias could be significantly mitigated by fine-tuning custom GPT models on principles of Diversity, Equity, and Inclusion (DEI) and disability justice (Glazko et al., 2024a). Similarly, in the context of resume screening and shortlisting, ChatGPT can be fine-tuned to provide more domain-specific insights, ensuring that the model takes into account the specific industry, company culture, and job requirements. This fine-tuning allows for more accurate and context-aware shortlisting, tailored to the needs of the organization. These findings underscore the necessity of further investigating LLMs before their implementation in recruitment processes.

2.4. Ethical Considerations

The application of AI in recruiting brings significant ethical challenges, particularly as the technology rapidly evolves beyond the current capacity of academic research to fully address. Many existing approaches to mitigating ethical risks in AI are general and not specifically tailored to the recruiting context (Hunkenschroer & Luetge, 2022). This gap allows firms and technology vendors to push the boundaries of acceptable practices, often without fully considering the ethical implications.

While AI screening tools are praised for their efficiency, especially in handling large volumes of applications, there are concerns that qualified applicants may be overlooked due to biases embedded within the AI systems (Persson, 2016). For instance, a study on GPT-4 demonstrated that it exhibited prejudice against resumes enhanced with disability-related achievements (Glazko et al., 2024b). However, this bias was shown to be reducible through targeted training on diversity, equity, and inclusion principles.

Research suggests that the validation of AI assessment tools should be approached with a focus on the unique attributes of AI, rather than merely comparing them to traditional hiring methods (Hunkenschroer & Luetge, 2022). Without this, there is a risk of perpetuating biases and unfair practices in recruitment processes .

Companies are increasingly attracted to AI tools across the entire hiring spectrum, often without a clear understanding of the potential ethical issues these technologies may introduce. This can result in outcomes that conflict with the company's goals for workforce diversity and fairness (MontrealEthics, 2020b). Additionally, many bias mitigation systems are designed to meet U.S. legal standards, making them less applicable or effective in European markets (Sánchez-Monedero et al., 2020).

Notable cases in the past have demonstrated the potential consequences when companies underestimate ethical considerations, even if unintentionally. For example, Amazon's AI-driven hiring tool, used in 2016, was found to penalize resumes that included the word "women's," such as in "women's chess club captain," revealing a gender bias embedded within its algorithm (Mujtaba & Mahapatra, 2019). Similarly, in 2018, Facebook faced a lawsuit for allowing job advertisers to target users based on age and gender, illustrating how AI can perpetuate discriminatory practices (BBC, 2019; MontrealEthics, 2020b). Additionally, a study showed that Facebook's targeted ads for supermarket cashier positions were shown predominantly to women, indicating that sourcing algorithms can lead to adverse impacts (Bogen, 2019). Another notable example is from a Carnegie Mellon study, which found that Google displayed advertisements for high-paying executive jobs significantly more often to men than to women, further highlighting gender bias in AI-driven ad targeting (Spice, 2015). These cases underscore the importance of thoroughly addressing ethical considerations in AI development and deployment, particularly in recruiting, where the impact on individuals' careers and lives is profound.

2.4.1. Fairness and Bias in AI Recruiting

AI systems used in recruiting must prioritize fairness and actively mitigate bias. Fairness is a multifaceted and complex concept (Selbst et al., 2019) with deep roots in various disciplines, explored extensively over a long period of time (Mulligan et al., 2019). For the purposes of this thesis, fairness is defined as the principle that AI systems and LLMs should make decisions impartially, without unjustly favoring any group or individual. This entails ensuring that the outcomes of LLMs are equitable across diverse demographic groups, particularly concerning age, gender, and nationality, specifically distinctions between EU and non-EU nationals.

Fairness Criteria in Focus

In this thesis, the fairness criteria of gender, age, and nationality are examined in detail:

- **Gender**: Research shows that human recruiters' gender biases can lead to women having a significantly lower chance of being interviewed for gender-neutral jobs compared to equally qualified men (Pisanelli, 2022). This study found out that introducing AI in the resume screening process shrinks such a gender gap by 43 percentage points. Introducing AI into the resume screening process has been found to reduce this gender gap significantly . Some innovative methods, like masking gender-specific terms in resumes, are being explored to further mitigate these biases (Gagandeep & Mathur, 2024).
- **Age**: AI systems must be examined and tested to ensure they do not discriminate based on age, an area where human biases often play a role in traditional recruiting.
- **Nationality**: AI systems must be thoroughly examined and tested to prevent biases related to nationality, ensuring that all candidates are treated fairly regardless of their background and nationality status.

***Note**: For the purposes of this experiment, the ethnicity fairness criterion has been simplified to distinguish between EU and non-EU nationals.*

Excluded Fairness Criteria

Although I recognize other relevant fairness criteria such as education, level of experience, career gaps, disabilities, sexual orientation, and religion, I have narrowed my focus to gender, age, and ethnicity due to time constraints. The experiment was conducted in the automotive industry, which has a notable gender disparity (Deloitte, 2020), making it particularly important to examine gender as a fairness criterion in AI recruitment. Additionally, as Barocas and Selbst (2016) highlight, AI systems pose significant risks of disparate impact, especially concerning racial and ethnic biases. Age discrimination in hiring is also a well-documented issue, with older applicants often being overlooked (Lahey, 2008). Given these factors, I chose to focus on age, gender, and ethnicity as key fairness criteria for analyzing the ethical implications of AI in recruitment.

2.4.2. The EU AI Act and Its Implications for AI in Recruiting

The introduction of the EU AI Act marks a significant milestone in the regulation of artificial intelligence (AI) within the European Union. Proposed by the European Commission in April 2021 and effective from August 1st 2024, the Act aims to create a comprehensive legal framework that governs the development, commercialization, and use of AI technologies across the EU (Future of Life Institute (FLI), 2024). This legislation is one of the first of its kind globally, setting a precedent for how AI should be responsibly integrated into various sectors, including recruitment.

The implications of the EU AI Act for AI in recruiting are particularly noteworthy due to the stringent regulations it imposes on high-risk AI systems and the emphasis on transparency and fairness. The EU AI Act classifies AI systems used in recruitment, candidate selection, and evaluation as high-risk, necessitating strict compliance with regulatory standards. Providers of these systems must implement comprehensive risk management frameworks, including data governance, technical documentation, and human oversight, to ensure accuracy, security, and the absence of bias.

HR departments and recruiters using high-risk AI systems must follow guidelines to maintain transparency and reliability, ensuring that these systems do not unintentionally perpetuate biases. The Act also emphasizes transparency, requiring that users are informed when they interact with AI, such as in chatbots or automated screening tools. Additionally, the Act mandates that AI systems be designed to avoid biases, promoting fair treatment of all candidates.

Certain AI practices are explicitly prohibited by the EU AI Act, including the use of manipulative techniques to influence hiring decisions and exploiting socio-economic vulnerabilities to distort candidate behavior. These measures are in place to protect candidates from unethical treatment during the recruitment process. For General Purpose AI (GPAI) models, which can be adapted for various uses including recruitment, the Act requires detailed technical documentation and adherence to compliance standards to ensure ethical and legal integrity when these models are applied in hiring contexts.

An important aspect of the EU AI Act's regulatory approach is its alignment with anticipatory AI governance principles. Research in this area highlights that anticipating the impacts of AI on downstream tasks is an ongoing field of study (Kieslich et al., 2024). Anticipatory governance approaches, which assess the societal impacts of emerging technologies, are crucial for understanding how AI affects stakeholders in terms of moral rights, potential discrimination, and privacy concerns (Diakopoulos & Johnson, 2021; Guston, 2014; Kieslich et al., 2023, 2024). By identifying potential hazards early in development, these approaches aim to mitigate risks and shape responsible and ethical AI use. The EU AI Act embodies these anticipatory principles by setting rigorous standards to address both positive and negative impacts of AI, ensuring that technological advancements do not compromise fairness and equity.

An additional legal consideration is raised by Sánchez-Monedero et al. (2020), who question whether, due to the new GDPR regulation, it is in fact illegal to use a solely automated hiring system in the EU, as the GDPR grants people the right to a "human in the loop" (Hunkenschroer & Luetge, 2022). This point highlights the potential tension between the EU AI Act and existing data protection laws, emphasizing the importance of human oversight in AI-driven recruitment processes.

The EU AI Act sets a high standard for the use of AI in recruitment, aiming to ensure that these systems are transparent, fair, and secure. By imposing significant obligations on both developers and users of high-risk AI systems, the Act seeks to mitigate bias, protect individual rights, and promote the ethical use of AI in hiring processes (Future of Life Institute (FLI), 2024). This regulatory framework is a crucial step towards responsible AI adoption in recruitment, ensuring that technological advancements do not come at the expense of fairness and accuracy.

2.5. Accuracy vs Fairness Trade-off

When evaluating the accuracy and fairness of LLM-based resume screening and shortlisting, it is essential to consider the trade-off between these two objectives. Achieving the highest levels of both accuracy and fairness simultaneously is often unrealistic; instead, a balance must be found, tailored to the specific use case.

In their book *The Ethical Algorithm*, Roth and Kearns (2019) introduce the concept of the "*Pareto frontier*" to explain this trade-off. The Pareto frontier represents the set of optimal models that balance accuracy and fairness. Moving along the frontier reveals that improving one objective, such as accuracy, often comes at the expense of the other, such as fairness. Any model not on the frontier is considered suboptimal, and no model can fully optimize both accuracy and fairness simultaneously. This trade-off requires thoughtful judgment, underscoring the necessity of human involvement in selecting the appropriate balance for each application. It's equally important to understand how hiring managers and job candidates behave in the context of automated resume screening as it is to understand the software itself (Selbst et al., 2019). Roth and Kearns (2019) further stress that fairness is not a one-size-fits-all solution; societal and ethical judgments must guide how these trade-offs are managed.

This issue is also addressed by Barocas and Selbst (2016), who analyze LinkedIn's TalentMatch. They argue that if LinkedIn recommends candidates based on employers' demonstrated interest in certain profiles, TalentMatch will likely reinforce any biases employers already hold. In this case, the concern shifts from accuracy to fairness. While TalentMatch may improve the quality of candidate recommendations based on recruiters' preferences, it risks creating a biased and unfair process that disadvantages specific groups not favored by those recruiters.

It is important to recognize that fairness is a property of social and legal systems, not of technical tools themselves (Selbst et al., 2019). Therefore, organizations using AI-driven recruiting platforms must clearly understand and clarify these trade-offs. Awareness of the implications is the first step toward managing them effectively.

To make the implications of accuracy and fairness more transparent, it is crucial to quantify both. In the context of resume screening and shortlisting, methods that can quantify the accuracy and fairness of LLMs are necessary to explain and evaluate this trade-off meaningfully.

3. Methodology and Materials

To address the two research questions of this thesis, an experiment was conducted within a real-world company. The goal of the experiment was to compare traditional resume screening by human recruiters with resume screening performed by LLMs. This comparison focused on evaluating the feasibility of using LLMs in recruitment based on performance metrics such as efficiency, performance, and quality of results. Additionally, the experiment aimed to assess the ethical implications of LLM-driven resume screening, particularly in terms of fairness across gender, age, and nationality. This chapter provides a detailed overview of the use case that underpins the experiment, including information about the participating company, its recruitment process—specifically resume screening and shortlisting—participants, the dataset used, and the experimental setup. The specific methodologies employed to answer each research question are also outlined at the end of this chapter

The results of the study are detailed in Chapter 4. Additionally, I developed a prototype designed to assist in AI-based resume screening and shortlisting. For more information, please refer to Chapter 5.

3.1. AVL Use Case

To carry out the experiment, I have partnered with AVL List GmbH (AVL), a leading global mobility technology company headquartered in Graz. AVL specializes in development, simulation, and testing for the automotive industry, as well as other sectors such as rail, marine, and energy. Through extensive in-house research, AVL provides concepts, technology solutions, methodologies, and development tools aimed at creating a greener, safer, and better world of mobility and beyond (AVL, 2024).

The objective of the collaboration with AVL is to explore the use of AI in recruiting and understand its implications. I have worked closely with AVL's Human Resource Department to design an experiment involving real recruiters and hiring managers from the company and LLM AI models. The methods for evaluating the ethical qualities of existing AI algorithms in HR recruitment are specifically tailored to meet AVL's HR criteria, ensuring the results are practically relevant and applicable within the organization.

3.2. Participants

The recruiting process involves two primary roles: **recruiters** and **hiring managers**. Recruiters handle the initial stages of the recruitment process, including the initial screening of candidates. They are responsible for shortlisting or manually rejecting applications. Once a shortlist is created, hiring managers review it and make decisions on whether to invite candidates for interviews, reject them, or place them on hold. A recruiter is a member of the recruitment team, while a hiring manager is typically a high-level employee within the department that is hiring.

To carry out the experiment, 3 recruiters and 10 hiring managers participated. Below, I provide a detailed description of the recruiting process, outlining the responsibilities of both the recruiters and the hiring managers. Note that this experiment specifically focuses on the resume screening and shortlisting part of the process.

3.2.1. Advertising a Job

Decision to Announce a New Job Opening
The initiation of a new job opening is determined by the department that has identified the need for additional personnel. The department reports its personnel requirements, which must be approved through an established approval flow. This ensures that the need for a new hire is justified and aligned with the organization's strategic goals.

Determination of Qualifications and Requirements
The qualifications and requirements for a specific job position are specified by the Hiring Manager. This is done in collaboration with the recruitment team during a coordination meeting. The Hiring Manager, being intimately familiar with the needs and responsibilities of the position, outlines the necessary qualifications. These specifications are then discussed and agreed upon with the recruitment team to ensure they are clearly defined and aligned with the organization's standards.

Creation of the Job Advertisement
The creation of the job advertisement is a collaborative effort between the department and the recruitment team. Initially, the department drafts the job advertisement. This draft is then reviewed in a coordination meeting involving both the department and the recruitment team. After this meeting, the recruitment team revises the job advertisement to ensure it meets the organization's standards and effectively attracts suitable candidates.

The job is then advertised on the AVL job portal and on various platforms such as LinkedIn.

3.2.2. Screening and Shortlisting

Applications for the specific job position are submitted through SAP SuccessFactors, where each applicant creates a profile and submits their resume and other required documents. SAP SuccessFactors HCM is a suite of cloud-based HCM software applications that supports core HR and payroll, talent management, HR analytics and workforce planning, and employee experience management (SAP, 2024). The recruiters review each application and, based on the job description and requirements, decide whether to reject or shortlist the applicants. They can provide comments about their decision for each application, but it is not mandatory.

3.2.3. Interviewing

The shortlisted candidates are forwarded to the respective hiring manager through SAP SuccessFactors. The hiring managers can view the shortlisted candidates, their resumes, and comments from the recruiters, and then decide to either reject, put on hold, or invite the shortlisted candidates for an interview. Again, they can write comments about their decision, but it is not mandatory.

3.2.4. Final Hiring Decision

Once a shortlisted candidate is invited for an interview and receives positive feedback from the hiring manager, an additional interview with the recruitment team is coordinated. Following this interview, the Hiring Manager and the recruitment team discuss the candidate's suitability for the role. Ultimately, the Hiring Manager makes the final hiring decision. However, the recruitment team has the authority to veto this decision if they identify any significant concerns or discrepancies.

3.3. Dataset

The dataset for our experiment comprises two types of data: job descriptions and resumes.

I retrieved 10 job descriptions from AVL's online job portal, where they list all open positions. These job descriptions span various roles, including Data Scientist, Electrical Engineer, IT Specialist, Program Manager, Project Plant Engineer, Sales Manager, SAP Consultant, Security Manager, Service Engineer, and System Lead. This selection covers both engineering and management areas.

There were various options to obtain resume data for this experiment. One option was to use historical data from the company, which could provide real-world relevance and authenticity. However, studies show that predictions based on historical data of a company for a customized tool can further deepen the underrepresentation of females, non-binary applicants, ethnic minorities, people with disabilities (Montreal Ethics, 2020a). Another option was to use open-source resumes available from public repositories, ensuring a wide variety of formats and styles. Nevertheless given that most employees in the automotive industry are male, with women representing only about 13% of the workforce (Deloitte, 2020), I aimed to create diverse resumes for each position using the free version of GPT-4 (OpenAI, 2024a) . For each job position, I generated 8 to 12 resumes. To avoid algorithmic bias and ensure a diverse dataset, I formulated the following prompt to generate the resumes:

```
Create 8-12 diverse resumes for the position of {Job Title}
at AVL List GmbH, one of the world's leading mobility
technology companies specializing in development,
simulation, and testing in the automotive industry, as
well as other sectors such as rail, marine, and energy.
The resumes should reflect a variety of candidates,
differing in age, gender, nationality, experiences,
and educational backgrounds.
Ensure that each resume includes:

 - A unique name, personal information such as age,
 gender, nationality, current location, and contact
 information
 - A summary or objective tailored to the job position
 - Detailed work experience relevant to the
 job position,  with diverse career paths
 and achievements
 - Varied educational backgrounds,
 including different  degrees and institutions
 - Any additional sections such as certifications,
 languages, volunteer work, or hobbies
 that highlight the candidate's unique background
```

To prepare the data for the experiment, I grouped the job descriptions and all respective resumes into a zip folder. This folder was then delivered to AVL for the first part of the experiment: Resume Screening with Real Recruiters.

3.3.1. Dataset Analysis

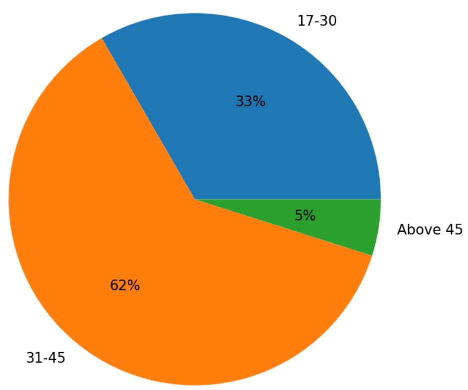

Figure 3.1.: Age Range Distribution of Generated Resumes

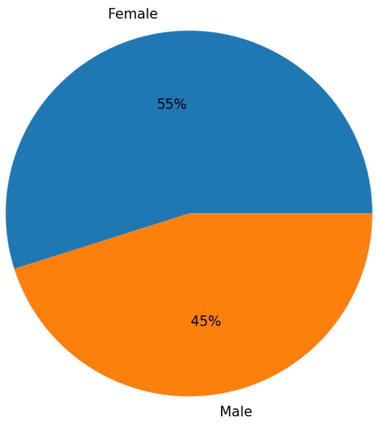

Figure 3.2.: Gender Distribution of Generated Resumes

A total of 102 diverse resumes were created. The applicants' ages range from 23 to 50 years, with an average age of 34. As illustrated in Figure 3.1, the majority of applicants (62%) fall within the age range of $31 - 45$ years, while only 5% are over 45, and 33% are between $17 - 30$ years old.

Age	Count	Status
17-30	34	33%
31-45	63	62%
Above 45	5	5%

Table 3.1.: Age Range Distribution of Generated Resumes: Number and Percentage Breakdown

Gender Distribution		
Male	Female	Other
46(45%)	56(55%)	0(0%)

Table 3.2.: Distribution of Gender in Generated Resumes: Number and Percentage Breakdown

Figure 3.2 shows that 55% of the applicants are female, and 45% are male. Figure 3.3 provides an overview of the applicants' citizenship, revealing that 61% are EU citizens, while 49% are non-EU citizens.

Additionally, Figure 3.4 depicts the highest degrees obtained by the applicants: 55% hold a bachelor's degree, 30% have a master's degree, 10% possess a Ph.D., and 5% have a high school diploma or a professional diploma. Another notable observation, which will be explored further in the next chapter, is the current geographic location of the applicants. As illustrated in Figure 3.3, 73% of the applicants are located in Austria, 10% in Germany, and approximately 17% in other countries around the world

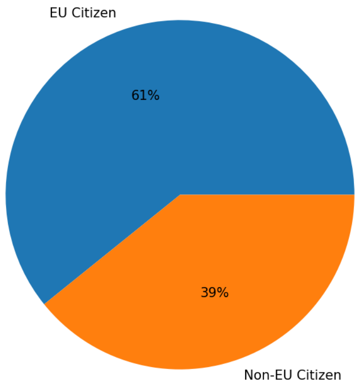

Figure 3.3.: Nationality Distribution n Generated Resumes

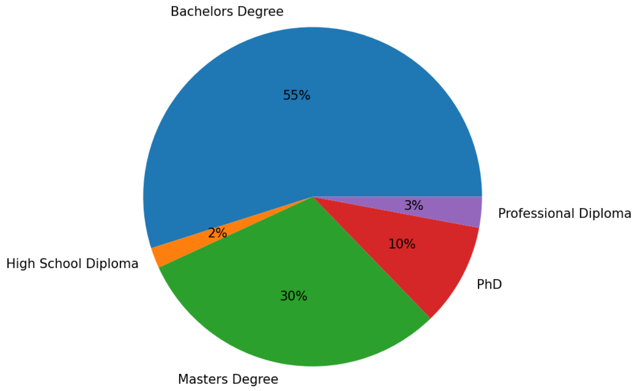

Figure 3.4.: Education Distribution in Generated Resumes

Location Distribution	
Country	**Stats**
Austria	76(76%)
Germany	10(10%)
Spain	3(3%)
United Kingdom	2(2%)
Morocco	2(2%)
Italy	2(2%)
Switzerland	1(1%)
United Emirates	1(1%)
Slovenia	1(1%)
Slovakia	1(1%)
Serbia	1(1%)
Japan	1(1%)
Greece	1(1%)

Table 3.3.: Location Distribution in Generated Resumes: Number and Percentage Breakdown

A significant proportion of the generated resumes are from individuals located in Austria, reflecting the fact that most job positions were based there. Notably, while the AI generated diverse locations globally, there is a marked concentration in Austria, aligning with the origin of the job descriptions.

3.4. Experimental setup

To gain insight into AVL's current recruiting process, I conducted an interview with the Global Hiring Manager. This interview covered all phases of the application process (for more details check section 3.2), from advertising the job, sourcing resumes to resume screening, shortlisting, interviewing, and hiring. To answer the two research questions of this thesis, the experiment focuses on the resume screening phase of AVL's recruitment process. As illustrated in Figure 3.5, every job application at AVL undergoes resume screening conducted by AVL recruiters. They make decisions to either reject the applications or shortlist them. If shortlisted, the candidates are then reviewed by the hiring manager, who decides whether to invite them for an interview, place them on hold, or reject their application. Based on the information and insights gathered and to better address the research questions of this thesis, I structured the experiment into two parts:

3.4.1. Resume Screening with Real Recruiters

For this part of the experiment, I submitted the job descriptions and corresponding resumes to AVL's HR department. All three recruiters manually entered the data for all 10 job positions into SAP SuccessFactors and then reviewed all applications as they would in their normal workflow. Each recruiter was assigned 3-4 job positions to review. For each job description, recruiters screened each resume and decided whether to shortlist or manually reject the candidates. In the regular process, the shortlist is then forwarded to the hiring manager, who decides whether to invite the applicant for an interview, reject them, or put them on hold.

A hiring manager was assigned to each individual job position, depending on the department that advertised the job. They received the shortlisted candidates, reviewed each resume along with comments from the respective recruiters, and then decided to either invite the candidates for an interview, reject them, or put them on hold.

At the end of the experiment, I received an Excel file containing each job position and its respective resumes. The file detailed the job ID, job title, candidate ID, candidate first name, candidate last name, recruiter's last name, application status label (manually rejected or shortlisted), hiring manager's last name, manager's feedback (invite to interview, candidate on hold, or reject candidate), and comments.

3.4.2. Resume Screening with LLMs

In this part of the experiment, I replicated the resume screening process using large language models (LLMs), specifically ChatGPT and Microsoft Co-Pilot.

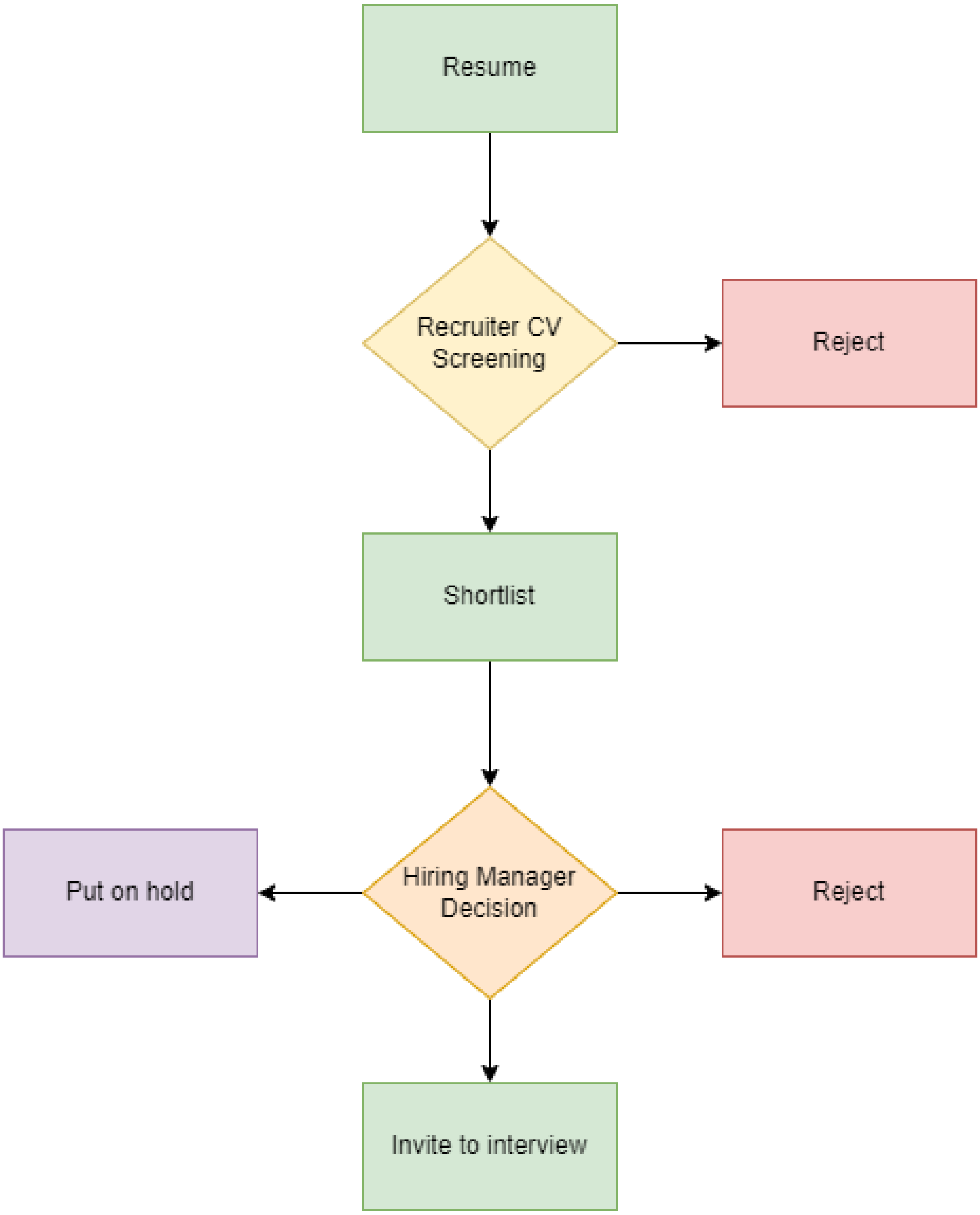

Figure 3.5.: AVL Recruiting Schema: Overview of the Process from Resume Submission to Hiring Manager Decision

(For clarity, I employed a consistent notation system throughout the documentation: prompts are presented in listings to differentiate them from the main text, general outputs are enclosed in dashed boxes, and resume evaluations are highlighted using sticky notes.) I used the following prompt for the LLMs:

```
Act as a seasoned recruiter in the automotive industry.
First, analyze the provided job description to determine
the most important criteria for the position. Based on
these criteria, screen the applicant resumes and
rate them by providing a decimal score from 1 to 10.
Create a shortlist of the best candidates.
For each candidate, state the reasoning behind
your decision and the rating,
both for those shortlisted and those not shortlisted.
```

This process exemplifies prompt engineering—a method of tailoring inputs to elicit more accurate and relevant outputs from the models. By refining prompts, I aimed to enhance the transparency and explainability of the AI algorithms, making it easier to interpret the LLMs' reasoning behind their decisions.

For instance, by specifying certain skills or qualifications in the prompt, I was able to guide the LLMs to focus on those aspects when screening candidates, thereby improving the relevance of the generated outputs. This allowed me to better understand the results produced by the LLMs and compare them against the criteria set by human recruiters. I enriched the Excel file from the real recruiting experiment by adding the LLM scores/ratings, final decisions, and comments/justifications for each job position and individual resume. This allowed for a direct comparison of the performance and decision-making processes between the LLMs and human recruiters (see section 3.5 for more details).

Find below an example of input provided for one of the job position: Electrical Engineer f/m/d for Test Systems - focus on Technical Sales, followed by the outputs produced by ChatGPT and Microsoft Copilot.

Example

Input:

```
Act as a seasoned recruiter in the automotive industry.
First, analyze the provided job description to determine the
most important criteria for the position. Based on these
criteria, screen the applicant resumes and rate them by
providing a decimal score from 1 to 10. Create a shortlist
of the best candidates. For each candidate, state the
reasoning behind your decision and the rating, both for
those shortlisted and those not shortlisted.
Job: {Job Description}
```

Resume1: Alex Jagsch **Personal Information Gender**: Male **Age**: 35 **Nationality**: Austrian Location: Graz, Austria **Summary**: Experienced electrical engineer with a strong background in technical sales. Proficient in conceptualizing ndustrial electrical systems and automation solutions. Skilled in detailed planning and cost calculation for measurement and control technology. Excellent communication and negotiation skills. **Education**: Master's degree in Electrical Engineering Graz University of Technology, Austria **Experience**: Senior Electrical Engineer Magna Engineering Solutions GmbH Led the conceptualization and detailed planning of industrial electrical systems for various projects. Managed quotation processing and collaborated with sales colleagues to meet customer requirements. April 2017 – Present (7 years) Graz, Austria Technical Sales Specialist DEF Automation Ltd. Collaborated with customers to understand their needs and propose tailored automation solutions. Negotiated contracts and managed supplier relationships for electrical installations. January 2014 – February 2017 (3 years) Vienna, Austria Languages: English: Fluent German: Native

Resume2: Alessia Ferrari **Personal Information Gender**: Female **Age**: 27 **Nationality**: Italian Location: Graz, Austria **Summary**: Entry-level electrical engineer with a background in electrical systems design and analysis. Proficient in AutoCAD and MATLAB. Strong problem-solving skills and attention to detail. Seeking an opportunity to grow and contribute to projects in a dynamic environment. Fluent in English and Italian. **Education**: Master's degree in Electrical Engineering University of Padua, Italy **Experience**: Electrical Systems Designer REAL DATA Engineering Solutions Assisted in the design and simulation of electrical systems for industrial applications. Conducted analysis and optimization of system performance. March 2021 – Present (3 years) **Languages**: English: Fluent Italian: Native

Resume3: ...

All resumes are added to the input for the given job position

Microsoft Copilot Output (generated using Microsoft (2024)):

Based on the job description provided, the most important criteria for the position of Electrical Engineer f/m/d for Test Systems with a focus on Technical Sales at AVL List GmbH are:

1. Educational Background: A completed education in electrical engineering, power engineering, or automation technology from a Secondary Technical College, University of Applied Sciences, or Technical University.
2. Practical Experience: At least 5 years of practical experience in electrical planning and/or quotation processing.
3. Technical Skills: Experience with EPLAN P8 and SAP is considered an advantage.
4. Soft Skills: Team spirit, diplomatic and self-confident demeanor.
5. Travel: Willingness to travel, though to a very limited extent.
6. Language Proficiency: Very good command of German and English, both written and spoken.

Rating of resumes

1. Alex Jagsch's Resume **Rating**: 9.5/10
Reasoning:

- **Educational Background:** Alex holds a Master's degree in Electrical Engineering from Graz University of Technology, which aligns with the job's educational requirements.
- **Practical Experience:** With 7 years of experience as a Senior Electrical Engineer at Magna Engineering Solutions GmbH and 3 years as a Technical Sales Specialist at DEF Automation Ltd., Alex exceeds the 5-year experience requirement.
- **Technical Skills:** His roles involved conceptualizing industrial electrical systems and automation solutions, detailed planning, and cost calculation for measurement and control technology, which are key responsibilities of the position.
- **Sales Collaboration:** His experience in managing quotation processing and collaborating with sales colleagues demonstrates his capability in this area.
- **Soft Skills:** The summary mentions excellent communication and negotiation skills, which are crucial for the interface function to customers and internal departments. Continues...

...continues...

- **Language Proficiency:** Alex is fluent in English and a native German speaker, meeting the language proficiency criteria perfectly. Continues next page...

Shortlist Justification: Alex Jagsch is shortlisted due to his strong alignment with the job's criteria, particularly his relevant educational background, extensive practical experience, proven technical and sales skills, and language proficiency. His profile stands out as highly suitable for the Electrical Engineer position focused on Technical Sales at AVL List GmbH.

2. Alesia Ferrari's Resume **Rating**: 6/10
Reasoning:

- Educational Background: Alesia has a Master's degree in Electrical Engineering from the University of Padua, Italy, which satisfies the educational requirement.
- Practical Experience: With 3 years of experience as an Electrical Systems Designer at REAL DATA Engineering Solutions, Alesia falls short of the 5-year experience criterion.
- Technical Skills: She is proficient in AutoCAD and MATLAB, which are valuable skills, but there is no mention of EPLAN P8 or SAP experience.
- Soft Skills: Her summary highlights strong problem-solving skills and attention to detail, but there is no explicit mention of team spirit or diplomatic demeanor.
- Language Proficiency: Fluent in English and Italian, but the job description specifies a very good command of German, which is not mentioned in her resume.

Continues...

Continues... Shortlist Justification: Alesia Ferrari is not shortlisted due to her lack of the required 5 years of practical experience, unspecified experience with EPLAN P8 and SAP, and no mention of German language skills, which are important for the role. Her profile suggests potential for growth and contribution in a dynamic environment, but it does not fully meet the criteria for the Electrical Engineer position focused on Technical Sales at AVL List GmbH.

ChatGpt Output (generated using OpenAI (2024a)):

Key Criteria:

1. Educational Background: Master's or equivalent in Electrical Engineering, Power Engineering, or Automation Technology.
2. Work Experience: At least 5 years in electrical planning and/or quotation processing.
3. Technical Skills: Experience with industrial electrical systems, automation solutions, technical building equipment, and familiarity with EPLAN P8 and SAP.
4. Project and Sales Skills: Concept planning, basic engineering, detailed planning, cost calculation, managing customer interfaces, and involvement with suppliers.
5. Language Proficiency: Very good command of German and English.
6. Personal Attributes: Team spirit, diplomatic, self-confident, and willingness to travel.

Rating of resumes

Resume 1: Alex Jagsch: Rating: 9.0/10

- Education: Master's in Electrical Engineering
- Experience:Senior Electrical Engineer at Magna Engineering Solutions (7 years) Technical Sales Specialist at DEF Automation Ltd. (3 years)
- Technical Skills: Concept planning, detailed planning, cost calculation, industrial electrical systems, automation solutions.
- Languages: English (Fluent), German (Native)
- Strengths: Strong technical and sales experience, leadership in conceptualizing electrical systems.
- Weaknesses: No mention of EPLAN P8 or SAP experience.

Resume 2: Alesia Ferrari: Rating: 5.5/10

- Education: Master's in Electrical Engineering
- Experience: Electrical Systems Designer at REAL DATA Engineering Solutions (3 years)
- Technical Skills: Electrical systems design, AutoCAD, MATLAB.
- Languages: English (Fluent), Italian (Native)
- Strengths: Strong technical foundation, attention to detail.
- Weaknesses: Entry-level, no significant sales experience, no EPLAN P8 or SAP experience, not fluent in German.

Revised Shortlist of Best Candidates: Alex Jagsch (9.0/10)

- Strong background in both technical and sales aspects.
- Solid leadership and extensive experience in electrical systems.
- Fluent in both English and German.

Candidates Not Shortlisted: Alesia Ferrari (5.5/10)

- Entry-level candidate with no significant sales experience.
- Lacks proficiency in German and relevant technical software experience.

3.4.3. Prompt engineering

When discussing LLMs, the role of prompt engineering is crucial in obtaining desirable responses and outcomes from the model. Prompts, which are instructions or inputs provided by the user, shape the model's generation process, enabling it to produce more coherent, complete, and engaging conversations ("Prompts - LangChain Docs," 2024). Effective Prompt Engineering Techniques (PETs) enhance interactions with LLMs by facilitating the creation of precise queries that result in accurate and relevant outputs (Thanasi-Boçe & Hoxha, 2024).

In developing this thesis, I experimented extensively with various prompts to achieve the desired responses from both ChatGPT and Microsoft Copilot, observing a significant impact of input variations on the resulting output. A study on AI in entrepreneurship education also demonstrated that PETs, such as COMIN and PREP, significantly enhance the capabilities of AI models like GPT-4, making them more effective in generating targeted and useful responses (Thanasi-Boçe & Hoxha, 2024). As detailed in section 3.4.2, I used clear and structured prompts to instruct both models to screen applicant resumes based on job advertisement criteria, rate the resumes, and provide decimal scores along with explanations for each decision to shortlist or not. This approach ensured that the LLMs' responses were both clear and explainable. To illustrate this, let me examine the initial prompt I used with Microsoft Copilot for the job position described in section 3.4.2, highlighting the differences more clearly.

Input:

```
Can you screen and rate this resume based on the job
description provided below and take a decision to shortlist
it or not for the next round?
Job: {Job Description here}.
Resume: Alex Jagsch Personal Information
Gender: Male Age: 35 Nationality: Austrian
Location: Graz, Austria
Summary: Experienced electrical engineer with a strong
```

```
background in technical sales. Proficient in conceptualizing
industrial electrical systems and automation solutions.
Skilled in detailed planning and cost calculation
for measurement and control technology.
Excellent communication and negotiation skills.
Education: Master's degree in Electrical Engineering
Graz University of Technology, Austria
Experience: Senior Electrical Engineer Magna Engineering
Solutions GmbH Led the conceptualization and detailed
planning of industrial electrical systems for various
projects. Managed quotation processing and collaborated
with sales colleagues to meet customer requirements.
April 2017 Present (7 years) Graz, Austria Technical Sales
Specialist DEF Automation Ltd. Collaborated with
customers to understand their needs and propose tailored
automation solutions. Negotiated contracts
and managed  supplier relationships for electrical
installations.  January 2014 February 2017 (3 years)
Vienna, Austria  Languages: English: Fluent German: Native
```

Microsoft Copilot Output (generated using Microsoft (2024)):

> Based on the provided resume and job description, Alex Jagsch's profile aligns well with the requirements for the Electrical Engineer position focused on Technical Sales at AVL List GmbH. His Master's degree in Electrical Engineering from Graz University of Technology satisfies the educational requirement. With over 7 years of experience in electrical planning and quotation processing, he exceeds the minimum requirement of 5 years of practical experience. While the job description mentions that experience with EPLAN P8 and SAP is an advantage, it is not listed as a requirement. Therefore, even if Alex Jagsch does not have this experience, it should not disqualify him from being shortlisted.
> His demonstrated team spirit and negotiation skills suggest he has the necessary team spirit and diplomatic presence. The willingness to travel, even if to a very limited extent, is also met given the nature of his previous roles which likely required some level of travel.
> Lastly, his fluency in English and German meets the language requirements of the position.
> In summary, Alex Jagsch's qualifications and experience strongly match the job description provided for the Electrical Engineer position at AVL List GmbH. He should be shortlisted for the next round of the selection process.

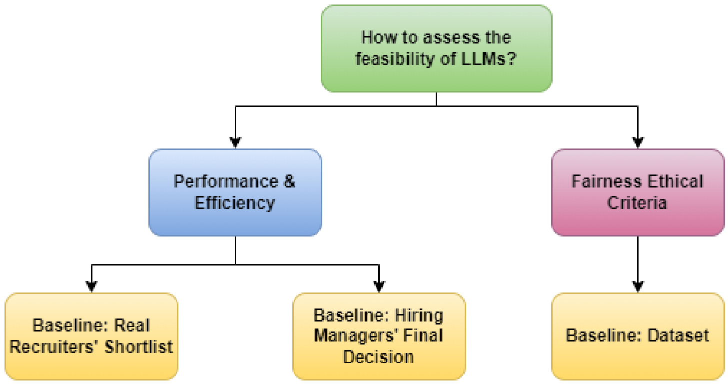

Figure 3.6.: Method used to assess the Feasibility of LLMs

As observed, the output generated by Microsoft Copilot using this prompt lacks the structured, clear, and organized presentation seen in the example from Section 3.4.2. It does not provide a resume rating or fully align the shortlist justification with the job description requirements, as achieved with my final prompt.

3.5. Methods for Assessing AI Model Feasibility

To answer the research questions and evaluate the feasibility of LLMs in terms of performance, speed, and fairness, I employed the following methods: validating LLM results against real recruiters' shortlists as a baseline, validating LLM results against hiring managers' final decisions, and assessing LLM results against the dataset's aggregated data on defined fairness criteria such as age, gender, and nationality. Each method is detailed in the subsections below, with an overview illustrated in Figure 3.6.

3.5.1. Real Recruiters' Shortlist as a Baseline

To assess the performance of the LLM results and address the first research question, the AI-generated shortlists were compared against those created by real recruiters for each job position. Using the real recruiters' shortlists as a baseline ensures that the LLMs are performing faster than human recruiters while maintaining the quality of the results. This approach helps validate that the LLMs can effectively expedite the resume screening process without compromising the quality of candidate selection. The results from both ChatGPT and Microsoft Copilot were assessed against the real recruiters' outcomes, and the similarity between them was observed.

3.5.2. Hiring Managers' Final List as a Baseline

To strengthen the comparison and more effectively address the first research question, I incorporated hiring managers into the experiment. As previously explained, the shortlist created by the real recruiters was forwarded to the hiring managers, who then decided whether to reject, put on hold, or invite the shortlisted candidates for an interview. The final list of candidates invited for an interview was used as a baseline to evaluate the results from the LLMs.

Using this final interview list as a benchmark enhances the assessment and allows for a deeper analysis of various aspects of the experiment. For example, it enables me to explore whether the top three candidates shortlisted by the LLMs align with the candidates ultimately invited for an interview by the hiring managers.

3.5.3. Addressing Fairness in AI Model Evaluation

Using real recruiters' shortlists and involving Hiring Managers is valuable for validating the speed and quality of AI models. However, this approach alone does not address ethical considerations such as fairness. The baseline used may exhibit biases based on factors like age, gender, or nationality.

To address ethical considerations regarding bias and fairness, the dataset was analyzed and aggregated into subcategories based on gender, age, and nationality in terms of EU and Non-EU nationalities. These subcategories served as a baseline for comparing the shortlisted candidates. This analysis aimed to assess whether the LLMs' shortlisting decisions align fairly across different demographic groups, providing insights into the model's performance beyond efficiency metrics.

4. Experiments and Results

Disclaimer: *The names of the applicants in this experiment part of my thesis are fictional and do not represent real individuals.*

4.1. Experiment and Results Overview

For this experiment, I utilize 10 job openings, each with 8 to 12 applicants, applying the shortlisting process consistently across all positions as described in Chapter 3.4. Due to space constraints, this section highlights one specific role: IT Specialist, to compare the shortlisting results generated by real recruiters, ChatGPT, and Microsoft Copilot. Additionally, I make a comparison with the decisions of the hiring manager, with an emphasis on understanding discrepancies in candidate ratings and the reasoning behind rejections and shortlisting decisions. Furthermore, I analyze the optional comments provided by the hiring manager during the experiment alongside the reasoning generated by the AI models. While the models consistently offer justifications for their decisions, the underlying logic remains unclear due to the black-box nature of LLMs. As mentioned in previous chapters, one recruiter and one hiring manager are assigned to evaluate the applications for each job posting in this experiment. For a comprehensive view of the results for all job positions, readers are encouraged to refer to the supplementary materials included with this thesis in Appendix A.

In Table 4.2, it is evident that the 'recruiter' shortlists candidates primarily based on experience, education, and how well their profiles align with the job description. However, the recruiter may lack expertise in the specific field, leading to a subjective decision-making process that often depends on how the candidate presents themselves in the resume and how a few key requirements match their experience or skill set. This subjectivity contrasts with the LLM models, which tend to align more closely with the hiring manager's decisions. The hiring manager, being a team supervisor, lead, or department head, typically has greater expertise in the field. This allows them to understand not only the formal job description but also the practical expectations of the role—details that may not always be captured in a job posting or advertisement. Consequently, job descriptions might be more generalized to attract a wider range of applicants, potentially resulting in a more extensive shortlist.

Table 4.1.: Example: Comparison of Shortlisting Outcomes: Recruiters, Hiring Managers, ChatGPT, and Copilot for the IT Specialist Position

ID	Name	Age	Gender	Nationality	Education	Location
22	Valentina Muller	31 – 45	Female	Austrian (EU)	Bachelors	Austria
23	Viktoria Schneider	17 – 30	Female	German (EU)	Masters	Germany
24	Pavel Sokolov	31 – 45	Male	Russian (Non EU)	Bachelors	Austria
25	Maria Lopez	17 – 30	Female	Spanish (EU)	Masters	Austria
26	Gulbeniz Ali	17 – 30	Female	Turkish (Non EU)	Bachelors	Austria
27	Christina Poltz	17 – 30	Female	German (EU)	Masters	Germany
28	Thomas Jagsch	31 – 45	Male	Austrian (EU)	Bachelors	Austria
29	Albina Gashi	17 – 30	Female	Albanian (Non EU)	High–School	Austria
30	Justin Davis	31 – 45	Male	British (Non EU)	Bachelors	Austria
31	Ahmed Ben Salah	17 – 30	Male	Moroccan (Non EU)	Bachelors	Morocco

On the other hand, LLM models are more deterministic in their approach, often rejecting candidates more definitively. For the 'IT Specialist' position, for example, GPT rejected nearly every candidate, assigning scores below 6 for most. Copilot, however, produced shortlists that were much closer to the hiring manager's selections, with all of its shortlisted candidates being invited for interviews. Interestingly, GPT assigned a score of 6.5 to some candidates who were later invited for interviews by the hiring manager. This suggests that GPT might be uncertain about how well an average candidate would perform based on the job description alone, whereas the hiring manager may consider broader factors, such as trainability and general skills, before making a final decision. While the alignment between LLMs and the hiring manager was mostly consistent across job descriptions in the experiment, there were a few exceptions.

Table 4.2.: Example: Comparison of Shortlisting Decisions: Recruiters, Hiring Managers (HM), ChatGPT, and Copilot (Model Scores in Brackets) for the IT Specialist Position.

ID	**Recruiter**	**Copilot (Rating)**	**GPT (Rating)**	**HM**
22	Shortlist	Shortlist (8,5)	Shortlist (9,5)	Interview
23	Reject	Reject (7,5)	Reject (4)	n/a
24	Shortlist	Reject (6,5)	Reject (4,5)	Reject
25	Shortlist	Shortlist (8)	Reject (6,5)	Interview
26	Shortlist	Reject (7)	Reject (5)	Reject
27	Shortlist	Shortlist (8,5)	Reject (6)	Interview
28	Reject	Shortlist (8)	Shortlist (7,5)	n/a
29	Reject	Reject (7.5)	Reject (5.5)	n/a
30	Shortlist	Reject (6,5)	Reject (6,5)	Reject, keep evidence
31	Reject	Shortlist (8,5)	Reject (4,5)	n/a

Tables 4.3 and 4.4 provide examples of the feedback from Copilot, GPT, and the hiring managers. In some cases, all parties rejected the same candidate, while in others, the recruiter rejected a candidate whom the models considered suitable. There were also instances where a recruiter shortlisted a candidate and the hiring manager invited them for an interview, but GPT rejected the same candidate. In this scenario, GPT identified a lack of required experience, which the hiring manager acknowledged but still decided to proceed with an interview. This decision likely stemmed from the manager's desire to explore how the candidate's broader experience might contribute to the role.

To provide a comprehensive overview of the results across all 10 job positions, I address the outcomes related to each research question in the section below.

Table 4.3.: Comparison of Hiring Manager Comments vs. LLM-Generated Comments: Analysis of Shortlisting Decisions for IT Specialist Position (Scores Indicating Shortlisting Status: 1 for Shortlisted, 0 for Not Shortlisted; Shortlisting status from Recruiter/-Copilot/Gpt/Hiring manager)

ID (decision)	**Comments**	
22 (1/1/1/1)	Copilot	Valentina Muller is a strong candidatefor the IT Infrastructure Specialist position, with relevant experience and skills that match the job description. She is shortlisted for further consideration.
	ChatGPT	Valentina has extensive relevant experience, strong technical skills, and excellent communication abilities. She meets almost all requirements perfectly.
	HM	According to her CV, the candidate has 5 years of relevant professional experience in the administration of Jira Service Management and Confluence as well as scripting. In any case, invite her to a personal interview to find out more details (server :: cloud, assets, etc.)
23 (0/0/0/0)	Copilot	Viktoria Schneider has a strong IT background, but her lack of specific experience with Jira, Confluence, and the required scripting and data processing skills means she falls short of the ideal candidate profile for this particular role. She is not shortlisted for this position.
	ChatGPT	Lacks specific experience with Jira and Confluence.
	HM	n/a – *HM does not screen this resume as it is manually rejected by the recruiter in a previous step.*

Table 4.4.: Supplementary Comments from Copilot, GPT, and Hiring Managers: Continuation of Table 4.3

ID (decision)		**Comments**
24 (1/0/0/0)	Copilot	Pavel Sokolov has a relevant IT background and language proficiency, but the lack of specific experience with the tools and platforms critical for the IT Infrastructure Specialist role at AVL List GmbH means he is not an ideal fit for this position. He is not shortlisted for this role.
	ChatGPT	Pavel skills do not align with the specific requirements for Jira and Confluence administration.
	HM	According to his CV, the candidate has no experience with Jira, Jira Service Management or Confluence (neither in administration nor as a user) and would need to be completely trained in the subject. There are no resources for this.
25 (1/1/0/1)	Copilot	Maria Lopez is a strong candidate for the IT Infrastructure Specialist position, with a solid educational background and relevant experience. Her recent work on Jira Service Management and Confluence is particularly promising. She is shortlisted for further consideration.
	ChatGPT	Maria has some relevant experience, but it's limited. Her technical skills and language proficiency are strong.
	HM	The candidate has an interesting profileregarding cloud services and scripting. Unfortunately, she probably has little to no experience in the administration of Jira, Jira Service Management and Confluence. **This should be clarified in an interview.**
28(0/1/1/0)	Copilot	Thomas Jagsch is a strong candidate for the IT Infrastructure Specialist position, with a solid educational background, relevant professional experience, and a good mix of technical and soft skills. He is shortlisted for further consideration.
	ChatGPT	Thomas has some experience with Jira and Confluence, but not extensive. Strong technical and analytical skills.
	HM	Prio B Junior

4.2. Observations

4.2.1. Research Question 1

As detailed in Section 3.5, to address RQ1 (1.1.1), I first use the real recruiters' shortlist as a baseline to evaluate the performance and quality of the shortlists generated by ChatGPT and Microsoft Copilot. To further validate the findings and test out another method, I extend the comparison by using the Hiring Managers' final interview list as an additional baseline.

The results from the first phase of the experiment are presented in Table 4.5. The experiment involves 102 CVs distributed across 10 job positions, with 69 candidates shortlisted by the real recruiters, 38 by ChatGPT, and 73 by Microsoft Copilot. Notably, the sizes of the real recruiters' and Microsoft Copilot's shortlists are quite similar, with Copilot shortlisting just 4 more candidates than the recruiters, while ChatGPT shortlisted 31 fewer. It's important to note that the individual number of applicants for each job position ranges from 8 to 12, which is why for this analysis I consider the aggregated number of 102 total applicants across the 10 positions.

ChatGPT, unlike Microsoft Copilot and real recruiters, tends to produce a more selective shortlist by choosing fewer candidates. While this might reduce the hiring manager's workload by providing a smaller pool to review, it could also potentially limit the range of candidates who get further consideration. This balance between efficiency and opportunity is an important aspect and is further investigated in the second part of this experiment.

Table 4.5.: Comparison of Shortlisted Candidates by Real Recruiters, ChatGPT, and Microsoft Copilot

Recruiter	**ChatGPT**	**Copilot**	**Total Applicants**
69	38	73	102
Matching Shortlisted Candidates			
Copilot **VS**. Recruiter			56/69
ChatGPT **VS**. Recruiter			36/69
Shortlisted by **Both** ChatGPT **and** Copilot			36/102

- 81.15% of the Copilot shortlist is identical to recruiter shortlist.
- 52.17% of the ChatGPT shortlist is identical to recruiter shortlist are invited in the interview by a hiring manager.
- 35.29% out of total number of candidates are shortlisted by both ChatGPT and Copilot.

n assessing LLM-based resume screening with regard to performance and efficiency, comparing the overlap of shortlists generated by different LLMs with those of real recruiters is an approach that offers valuable insights. As shown in the table, Microsoft Copilot's shortlist overlaps 81.15% with the real recruiters' list, whereas ChatGPT's overlaps only 52.17%. **This indicates that, when using the real recruiters' shortlist as a benchmark, Microsoft Copilot performs better than ChatGPT in terms of accuracy.**

Interestingly, despite a 35-candidate difference between the total number of candidates shortlisted by ChatGPT and Microsoft Copilot, 36 candidates are identical across both LLMs' shortlists. This means that 36 out of ChatGPT's 38 shortlisted candidates are also shortlisted by both the real recruiters and Microsoft Copilot.

To explore a second method, I use the hiring managers' final interview list as a baseline to evaluate the performance and efficiency of the LLM-generated shortlists. As described in Section 3.2, during the experiment the real recruiters' shortlisted candidates are forwarded to the hiring managers. The list of candidates invited for interviews from the hiring managers serves as a benchmark for this second phase of the experiment. The results from this phase are detailed in Table 4.6.

Table 4.6.: Comparison of Shortlisted Candidates by Real Recruiters, ChatGPT, and Microsoft Copilot Against Hiring Manager Interview Selections

Recruiter	**ChatGPT**	**Copilot**	**HM**
69	38	73	39

Matching Interviewing Candidates	
Hiring Manager **VS**. Recruiter	39/69
Hiring Manager **VS**. Copilot	33/73
Hiring Manager **VS**. ChatGPT	25/38

- 56.52% of the applicants shortlisted by a recruiter are invited in the interview by a hiring manager.
- 45.21% of the applicants shortlisted by Copilot are invited in the interview by a hiring manager.
- 65.79% of the applicants shortlisted by ChatGPT are invited in the interview by a hiring manager.

These results indicate that ChatGPT's shortlist aligns more closely with the hiring managers' final interview selections, suggesting a higher level of precision in candidate selection. While Microsoft Copilot shortlisted a larger number of candidates—73 out of 102 applicants—only 45.21% of these are ultimately invited for interviews, reflecting a broader but less targeted approach. Similarly, real recruiters shortlisted 69 candidates, with 56.52% progressing to interviews, indicating a relatively liberal selection process that may include more candidates who are not ultimately advanced.

In contrast, ChatGPT shortlisted fewer candidates (38 out of 102) but achieved the highest interview invitation rate at 65.79%, demonstrating a more selective and efficient selection process. **This suggests that ChatGPT may be more effective in identifying candidates who meet the hiring managers' criteria, potentially streamlining the recruitment process by reducing the number of unnecessary evaluations.**

Considering the results from both evaluation methods, I conclude that using the hiring managers' final interview list as a benchmark is the most appropriate approach for assessing the efficiency and performance of ChatGPT and Microsoft Copilot. Based on this method, ChatGPT outperformed Microsoft Copilot in the specific experiment conducted.

4.2.2. Research Question 2

As outlined in Section 3.5, to address RQ2 (1.1.2), I evaluate the shortlists generated by the LLMs against the dataset of 102 resumes, focusing on fairness criteria such as gender, age, and nationality. This dataset includes 46 (45%) male and 56 (55%) female candidates; 34 (33%) aged $17 - 31$, 63 (62%) aged $31 - 45$, and 5 (5%) over 45; and 62 (60.78%) EU citizens and 40 (39.22%) non-EU citizens.

Gender Representation in Shortlists:

As shown in Table 4.7, 42.03% of the applicants shortlisted by real recruiters are male, while 57.97% are female. ChatGPT shortlisted 39.47% male and 60.53% female candidates, while Microsoft Copilot shortlisted 38.36% male and 61.64% female candidates.

Table 4.7.: Gender Distribution of Shortlisted Candidates by Real Recruiters, ChatGPT, Copilot, and Hiring Managers

Gender Distribution		
	Male	**Female**
Total (Dataset)	46(45%)	56(55%)
Recruiters	29(42.03%)	40(57.97%)
ChatGPT	15(39.47%)	23(60.53%)
Copilot	28(38.36%)	45(61.64%)
Hiring Managers	15(38.46%)	24(61.54%)

Both LLMs (ChatGPT and Microsoft Copilot) and real recruiters show a tendency to favor female candidates over male candidates. The shortlists generated by ChatGPT and Microsoft Copilot include a slightly higher proportion of female candidates (60.53% and 61.64%, respectively) compared to the shortlists made by real recruiters (57.97%). This trend is also reflected in the hiring managers' interview invitations, where 61.54% of the selected candidates are female.

These findings indicate that both AI models and human recruiters tend to select female candidates at a rate higher than their representation in the original dataset, which comprised 55% female candidates. While this may suggest a shift towards greater gender fairness, it also raises important questions about the underlying causes of this trend. The higher selection rate could signal a positive bias toward gender representation, but this outcome should be approached with caution. It may reflect underlying biases in the models' training or the specific characteristics of the dataset, rather than a genuine move toward fairness in recruitment practices.

In conclusion, while the LLMs align with or even exceed real recruiters in selecting female candidates, this result warrants further investigation. It's crucial to ensure that the models are not inadvertently introducing or amplifying biases, as this could impact the fairness and diversity of the recruitment process.

Age Representation in Shortlists:

As shown in Table 3.2, real recruiters shortlisted 44.03% of applicants aged $17 - 31$, 57.97% aged $31 - 45$, and 7.25% aged over 45. ChatGPT shortlisted 78.95% of applicants aged $17 - 31$, 7.89% aged $31 - 45$, and 13.16% aged over 45. In contrast, Microsoft Copilot shortlisted 67.12% of applicants aged $17 - 31$, 26.03% aged $31 - 45$, and 6.85% aged over 45.

Table 4.8.: Age Range Distribution of Shortlisted Candidates by Real Recruiters, ChatGPT, Copilot, and Hiring Managers

Age Range Distribution			
	17–31	31–45	**Above** 45
Total (Dataset)	34(33%)	63(62%)	5(5%)
Recruiters	50(42.03%)	14(57.97%)	5(7.25%)
ChatGPT	30(78.95%)	3(7.89%)	5(13.16%)
Copilot	49(67.12%)	19(26.03%)	5(6.85%)
Hiring Managers	9(23.07%)	27(69.23%)	3(7.7%)

ChatGPT:

- **Bias Towards Younger Applicants**: ChatGPT shows a significant preference for younger applicants, with 78.95% of its shortlisted candidates aged $17 - 31$, despite this age group making up only 33% of the original dataset. This suggests that ChatGPT may disproportionately favor younger candidates.

- **Underrepresentation of Mid-Career Applicants:** Only 7.89% of ChatGPT's shortlisted candidates are aged $31 - 45$, a stark contrast to their 62% representation in the original dataset. This underrepresentation might indicate a potential bias against mid-career professionals.
- **Slight Overrepresentation of Older Candidates:** ChatGPT shortlisted 13.16% of candidates aged over 45, which is higher than their 5% representation in the dataset.

Microsoft Copilot:

- **Preference for Younger Candidates:** Microsoft Copilot also shows a preference for younger applicants, though less pronounced than ChatGPT's, with 67.12% of its shortlisted candidates aged 17-31.
- **Better Representation of Mid-Career Applicants**: Copilot shortlisted 26.03% of candidates aged 31-45, which, while still below their proportion in the dataset, is a more balanced approach compared to ChatGPT.
- **Proportional Representation of Older Candidates:** The percentage of candidates over 45 shortlisted by Copilot (6.85%) closely mirrors their representation in the dataset.

Real Recruiters:

- **Balanced Approach:** Real recruiters showed a more balanced approach in their shortlisting process, with 44.03% of candidates aged 17-31, 57.97% aged 31-45, and 7.25% aged over 45. These proportions are more reflective of the original dataset distribution, suggesting that human recruiters are less prone to age bias.

Interview Invitations by Hiring Managers:

- **Mid-Career Preference:** Hiring managers invited 69.23% of applicants aged 31-45 for interviews, reflecting a strong preference for mid-career candidates, which aligns more closely with the original dataset's proportions.
- **Underrepresentation of Younger Candidates:** Only 23.07% of candidates aged 17-31 were invited for interviews, indicating a potential mismatch between the shortlists generated by the LLMs (particularly ChatGPT) and the hiring managers' preferences.
- **Fair Representation of Older Candidates:** The percentage of older candidates invited for interviews (7.7%) is consistent with their representation in both the shortlists and the original dataset.

The analysis indicates that while real recruiters generally maintain a balanced approach to age representation, the LLMs, especially ChatGPT, exhibit a marked preference for younger candidates. Microsoft Copilot also shows a bias towards younger applicants but offers a more balanced shortlist in comparison to ChatGPT. This trend could disadvantage older candidates, particularly those aged 31-45, who form the majority in the original dataset and are more frequently invited to interviews by hiring managers.

The mismatch between the age distribution in the original dataset and the LLMs' shortlists underscores the importance of examining fairness in AI-driven recruitment processes. Although tools like ChatGPT and Microsoft Copilot can enhance the efficiency of hiring, their tendency to favor younger candidates could inadvertently introduce age-related biases, leading to less diverse and potentially inequitable outcomes. Therefore, while LLMs are valuable in streamlining recruitment, ensuring fairness in age representation is crucial to prevent the introduction of unintended biases.

Nationality Representation in Shortlists:

As shown in Table 4.9, real recruiters shortlisted 59.42% of applicants as EU nationals and 40.58% as non-EU nationals. ChatGPT shortlisted 65.79% EU nationals and 34.21% non-EU nationals, while Microsoft Copilot shortlisted 63.01% EU nationals and 36.99% non-EU nationals.

Table 4.9.: Nationality Distribution of Shortlisted Candidates by Real Recruiters, ChatGPT, Copilot, and Hiring Managers

Nationality Distribution		
	EU Citizen	**Non EU Citizen**
Total (Dataset)	62(60.78%)	40(39.22%)
Recruiters	41(59.42%)	28(40.58%)
ChatGPT	25(65.79%)	13(34.21%)
Copilot	46(63.01%)	27(36.99%)
Hiring Managers	25(64.1%)	14(35.9%)

Real Recruiters:

- **Proportional Representation:** Real recruiters' shortlists are quite balanced, with 59.42% of the shortlisted candidates being EU nationals and 40.58% non-EU nationals. This distribution closely aligns with the original dataset, which consists of 60.78% EU nationals and 39.22% non-EU nationals, indicating that human recruiters are fairly proportional in their selection process.

ChatGPT:

- **Slight Preference for EU Nationals:** ChatGPT shows a slight preference for EU nationals, with 65.79% of its shortlisted candidates being EU nationals compared to 34.21% non-EU nationals. This suggests a small bias in favor of EU nationals, as the proportion of non-EU candidates is lower than in the original dataset.

Microsoft Copilot:

- **Moderate Preference for EU Nationals**: Microsoft Copilot also favors EU nationals, with 63.01% of its shortlisted candidates being EU nationals and 36.99% non-EU nationals. While this is closer to the original dataset distribution than ChatGPT's shortlist, it still shows a preference for EU nationals.

Interview Invitations by Hiring Managers:

- **Alignment with Shortlists:** The hiring managers' interview invitations are reflective of the shortlists provided by real recruiters, with 64.1% of interviewees being EU nationals and 35.9% non-EU nationals. This suggests that hiring managers' decisions align with the proportions seen in the shortlists, particularly those generated by the real recruiters.

The analysis reveals that both ChatGPT and Microsoft Copilot show a modest bias toward selecting EU nationals over non-EU nationals when compared to the original dataset. This trend is less pronounced in Microsoft Copilot's shortlist, which is closer to the original dataset's proportions. Real recruiters appear to maintain a more balanced approach, closely reflecting the nationality distribution in the dataset. These findings suggest that while LLMs like ChatGPT and Microsoft Copilot can streamline recruitment processes, they may introduce subtle biases in nationality representation.

In conclusion, evaluating the performance of LLMs on the dataset to assess fairness across gender, age, and ethnicity has shown that LLMs may introduce biases in candidate selection. This highlights the need for further investigation. For more detailed discussions on methods to quantify fairness in LLMs, please refer to Section 6.1 in Chapter 6.

5. Use Case: Building an LLM-based Automated Recruiting System

This use case presents a practical implementation of a simple Retrieval-Augmented Generation (RAG) approach, designed to complement my thesis. The primary objective was to develop a rapid prototype that tests an automated recruiting process using RAG. The system was intentionally designed to be user-friendly, enabling users to easily grasp its functionality and understand how AI can assist in making informed decisions when selecting candidates. It is important to note that this system is not intended to serve as scientific proof that RAG is superior in automated resume screening but rather as a supplementary tool to explore the system's capabilities and potential implementation.

This work serves as an initial exploration of how AI can support HR processes, particularly in resume screening. It sets the stage for future enhancements and studies on the practical implications of AI in recruiting. Additionally, I conclude this chapter by emphasizing the importance of developing effective strategies to mitigate 'hallucination' issues while improving model performance. For additional references and user interface images, please refer to the appendix A.

5.1. A top down approach: Requirement Engineering

When designing the system, I focused on several key requirements. First, the system should provide users with a comprehensive overview of the data, including access to resumes and job openings stored in the database. Users should also be able to view data distribution metrics, such as gender, age, and nationality. The system is designed to shortlist candidates for specific job descriptions using the RAG approach, ensuring that the reasoning behind each decision is transparent and easily understood.The hybrid integration of conversational agents and RAG technology enhances efficiency and relevance by overcoming the limitations of each technology through precise, dynamic interactions and advanced memory capabilities (Thanasi-Boçe & Hoxha, 2024). Additionally, the system allows users to interact in a question-and-answer format (i.e. chatbot), enabling them to query resumes and receive job-matching results, or input job descriptions to receive best candidate recommendations based on semantic matching and the data available in the vector database.

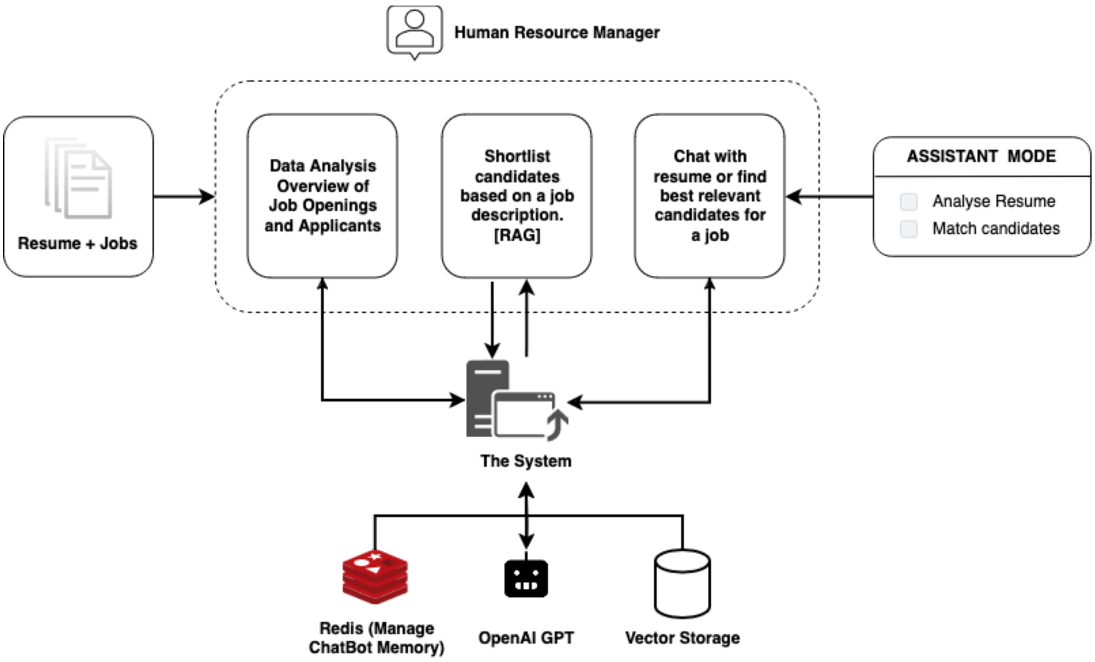

Figure 5.1.: Overview of the Use Case: Key Components and Affected Systems. Redis is utilized for storing persistent data during user interactions with the AI, allowing LLMs to retain session context and improve relevance. A new session ID is generated with each reload to maintain context.

When users first open the application, they are presented with an overview of the system's data, including graphical representations of data distribution. The application's front-end communicates with the back-end to fetch and display the relevant data analysis results.

Additionally, users can see which candidates best match specific job descriptions, allowing them to effectively manage the AI-generated shortlists. When a shortlist is requested, the system retrieves the most relevant resumes from the vector database based on the job description. The user can then query the system to rank these retrieved documents, with the LLM providing precise responses based on the query and the data.

Upon opening the chat feature, users are prompted with configuration options, allowing them to choose between two modes: "Analyse Resume" or "Find Relevant Candidate."

- In "Analyse Resume" mode, users can upload a resume or create one manually. They can then interact with the system to analyze the resume's data, and further query a job description to determine if the candidate is a good match.
- In "Find Relevant Candidate" mode, users are prompted to upload a job description or select one from existing descriptions in the system. They can then inquire about potential candidates from the candidate pool who match that job description.

5.2. Pipeline Architecture

The overview of the RAG system is depicted in Figures 2.2 and 2.3, with the architecture further explained in Chapter 2.2. The process begins with indexing parsed resumes and job descriptions into vector storage. During this phase, the documents are pre-processed and segmented into smaller text chunks to help the retriever efficiently identify relevant sections. This prevents overloading of documents with irrelevant information. These chunks are then converted into embedding vectors and stored in the vector database.

For this process, I used *recursive chunking*, which iteratively splits input text using pre-configured separators until the desired chunk size is achieved (Adapted from (Nguyen, 2024)). While more advanced methods such as 'content-based chunking' can divide resumes into meaningful sections, producing more relevant chunks, these approaches are computationally expensive and complex to implement.

In the retrieval phase, similarity search techniques are employed to extract the most relevant chunks from the vector storage, as described in Section 2.3. The retrieved chunks are then passed to the LLM through prompt engineering, providing the model with enriched context to generate more accurate responses. The detailed workings of the shortlisting process are illustrated in Figure 5.2.

5.2.1. Embedding Models

For this task, I used OpenAI's `text-embedding-3-large` model, one of the most powerful embeddings models available[1]. The model can process up to 8191 tokens per input, making it highly efficient for handling large text inputs. Additionally, open-source alternatives such as Hugging Face models offer competitive performance and can be considered for similar tasks (Wolf et al., 2020).

5.2.2. Prompts

Prompt templates were sourced from Langchain[2]. The complete implementation for this task is available on my GitHub profile.

5.2.3. Vector Storage

For vector storage, I utilized Chroma DB (Chroma, 2024), an open-source vector database known for its speed, ease of use, and scalability (Pan et al., 2023). Additionally, to maintain chat context and session management, I integrated

[1] OpenAI Embeddings
[2] Langchain Prompts

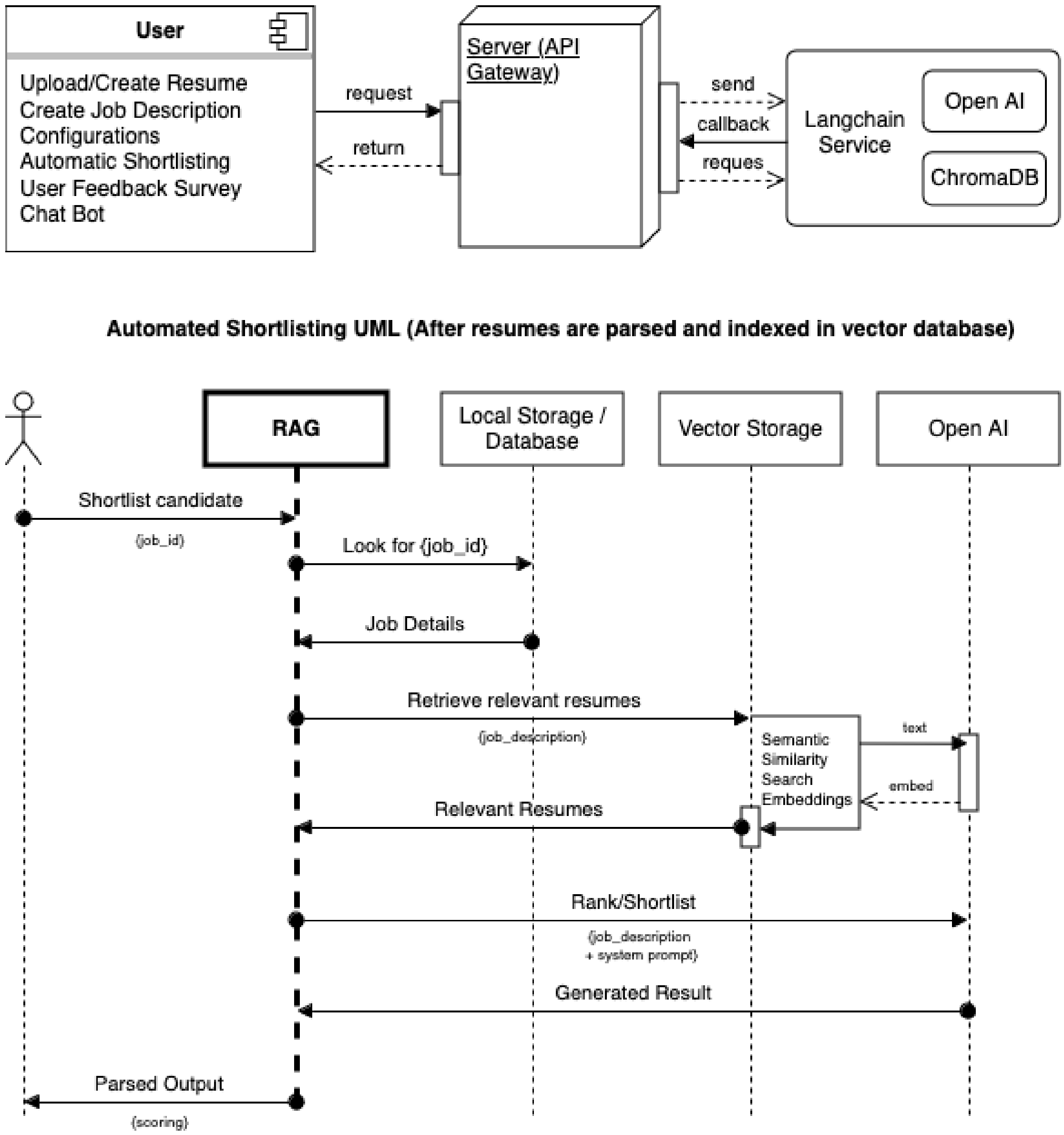

Figure 5.2.: Sample Diagram of the Shortlisting Process: The user initiates the process by requesting a shortlist for a specific job description. The system performs a semantic similarity search in the vector database to identify relevant candidates. These candidates and the job description are then processed by an LLM agent for reasoning and decision-making. Shortlisting is one of the two main components, with the other being a chat component that operates similarly and optionally integrates RAG Fusion (Section 2.2).

Redis[3] as a persistent storage solution, enabling the AI model to retain previous conversation details. This setup allows the model to produce more coherent and contextually relevant responses in multi-turn interactions.

[3]Redis.io

5.3. Evaluate RAG Pipeline

To evaluate the RAG pipeline, several methods have been proposed. In this thesis, I considered employing RAGAS evaluation policies (Es et al., 2023) to automate the evaluation process, as illustrated in Nguyen (2024). This method compares retrieved documents against a pre-generated ground truth dataset for a given task. In the context of this study, the process would involve retrieving resumes that match a specific job description and comparing them to a ground truth resume—essentially, a highly relevant resume for that job. This benchmark could serve as a reliable standard to evaluate the model's effectiveness.

While this aspect of the experiment is not the primary focus of this thesis, I still provide a brief outline of the procedure in this section. Alternatively, user feedback can be incorporated into the evaluation process. I introduced a *"survey mode"* post-AI shortlisting, allowing hiring managers to review the results, provide feedback, and suggest alternate rankings. This method helps assess user satisfaction and provides valuable insights for refining the system.

5.3.1. RAGAS

Introduced by Es et al. (2023), Retrieval Augmented Generation Assessment (RAGAs) is a toolkit that facilitates the evaluation of RAG systems without requiring human annotation. It offers automated metrics to assess the system's performance by comparing retrieved documents with a ground truth dataset. RAGAs enables the measurement of semantic similarity between the system's retrieved resumes and the pre-determined ground truth resumes, providing a robust framework for assessing the accuracy and relevance of the retrieval process (Es et al., 2023).

5.3.2. Participatory User Feedback

Users can provide feedback on the agent's performance after each shortlist generation via a survey. They can specify if the answers are incorrect, lack critical information, or exhibit bias. Additionally, users are invited to comment on how the AI's performance compares to their own judgment. This feedback approach is rooted in participatory design, which emphasizes iterative user input to enhance the system's effectiveness. This method offers a deeper understanding of how well the AI manages recruitment tasks.

In contrast, Peng et al. (2023) presents LLM-Augmenter, which enhances the generation process by integrating external knowledge, similar to RAG. However, LLM-Augmenter further refines LLM prompts by using automatic feedback through plug-and-play utility scoring functions to improve response quality.

5.4. One step further: RAG-Fusion

While RAG models generally perform well in real-world scenarios by recommending candidates similar to those chosen by human hiring managers, their effectiveness heavily relies on the quality of the retrieval component. The retrieval step enhances the large language model's reasoning capabilities and is essential for improving the generated responses. RAG Fusion represents an advancement in the RAG pipeline (Rackauckas, 2024), refining the retrieval process to better address ambiguous and complex queries.

In this system, when a user query is received, the model generates **K** sub-queries, which are then processed by the retriever. This process retrieves more relevant documents from vector storage. For each sub-query, the top-**K** most similar documents are selected. These documents are subsequently mixed, reranked, and filtered. This method enables the system to obtain more accurate external knowledge, thereby enhancing the responses generated by the language model.

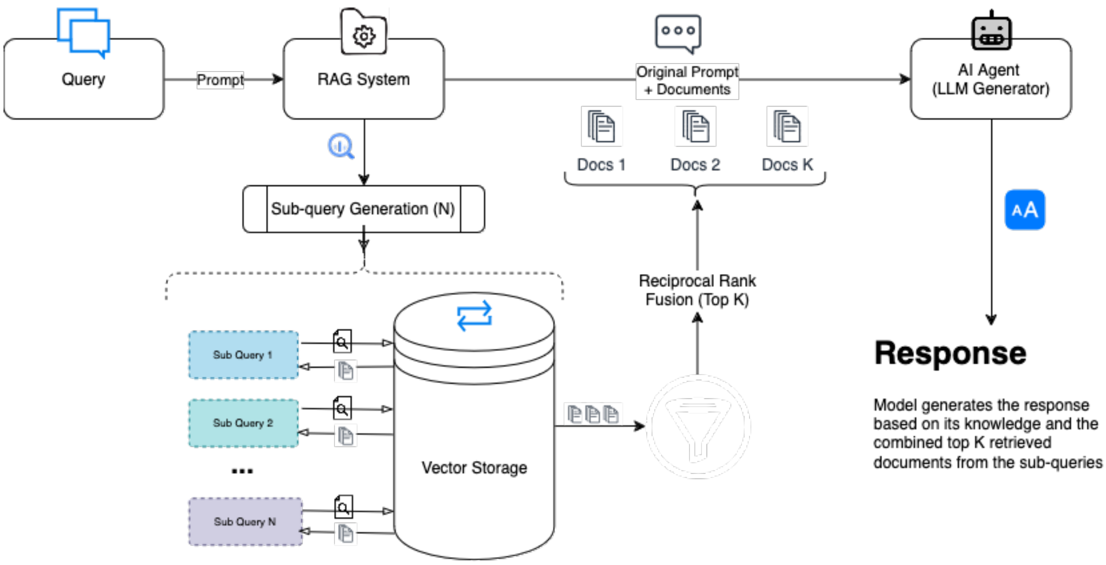

Figure 5.3.: RAG Fusion Setup: An enhanced version of the original RAG, which includes generating sub-queries for the main query. This approach allows the system to retrieve a broader context, filter documents using mean reciprocal rank, and return the top-K matches.

Alternative methods to improve the RAG framework include integrating and training the RAG components into a single end-to-end system, allowing simultaneous backpropagation through both the language model and the retriever (ContextualAI, 2024).

5.4.1. Comparison

Recent studies suggest that RAG-Fusion yields more accurate results (Nguyen, 2024; Rackauckas, 2024). However, the process of generating sub-queries is computationally expensive and resource-intensive, leading to increased time and costs, with only marginal improvements in overall outcomes. In my case, I did not observe a significant difference between RAG-Fusion and the original model, likely due to the limited number of documents in the retrieval database. The retriever's reported relevant resumes were generally accurate, but as the volume of stored data increases, so will the complexity. In such scenarios, RAG-Fusion is expected to outperform the original model. This expectation is further supported by Sawarkar et al. (2024), who propose "Blended RAG" as a method to enhance the retrieval model using dense vector indices.

5.5. Observations and Analysis

Overall, the outcomes are quite promising. The AI effectively understands complex, multi-faceted resumes and job descriptions, providing well-reasoned decisions. However, issues such as missing facts or fabricated answers could result in errors in real-world settings. The system performs well by encoding resumes and recommending relevant candidates based on user queries. The chat feature also offers additional options to test the model's reasoning when parsing resumes.

Limitations and Challenges: The system was initially limited to 102 resumes and 10 job descriptions, which constrained the analysis of the RAG system's ethical values. One major concern is "hallucination," where the AI generates imaginary data, potentially leading to incorrect candidate shortlisting or mismatches based on keywords in resumes. Additionally, the system may miss relevant information due to the chunking and retrieval phases, and incorporating a similarity measure beyond cosine similarity could improve the process.

Reproducibility and Model: The system is open-sourced on GitHub and may undergo significant changes as it continues to evolve. My focus has been primarily on models from OpenAI, but I am also exploring the integration of other models such as LLama and Mistral. The default model temperature is set to 0.7, but this setting can be configured by the user. The application is not deployed online due to hosting costs, and there are currently limited security measures to protect the application from potential online attacks. More information can be found in Appendix A.

6. Discussion

6.1. Baselines for AI in Global Recruitment

Performance Baseline: As detailed in Section 3.5, I evaluated the performance of both ChatGPT and Microsoft Copilot using two distinct baselines: the recruiters' shortlist and the hiring managers' final interview list. I found that the hiring managers' final interview list is the most suitable benchmark for comparison. Both ChatGPT and Microsoft Copilot assign a rating from 0 to 10 to each applicant based on how well their qualifications match the job requirements. For accuracy assessment, I quantified each applicant's performance as either 0 (not included in the hiring managers' final list) or 1 (included in the final list).

This approach is a valid measure of performance, as the decision of hiring managers to shortlist or invite a candidate for an interview indicates that the candidate is likely well-suited for the role. However, this method is ultimately based on the preferences of the recruiters and hiring managers, and incorporating fairness considerations into this evaluation may influence the results.

Fairness Baseline: Due to data limitations, the fairness baseline in this experiment was restricted to the 102 resumes used. I analyzed the results at an aggregated level rather than individual records. While considering various scenarios and baselines, such as using historical data from the company, I found that historical data may not always be a reliable fairness baseline, especially if it reflects past biases against certain groups.

When it comes to local recruitment, alternative baselines can be considered to ensure fairness. For instance, one approach could involve analyzing the number of students graduating annually from local computer science programs at universities. By comparing the proportion of women and men graduating from these programs with the proportion of each gender employed in entry-level positions such as computer scientists or software engineers, it may be possible to establish a more context-specific benchmark for gender representation in the local job market. However, while this method might provide useful insights into local recruitment practices, it may not fully capture the complexities of global recruitment, where diversity and fairness issues can vary significantly across different regions and industries.

Further research is needed to develop methods for evaluating AI models' fairness in terms of gender, age, and ethnicity on a global scale. This is crucial for ensuring equitable practices in AI-driven recruitment processes.

Quantifying fairness for each record is essential for assessing the trade-off between accuracy and fairness in both ChatGPT and Microsoft Copilot. This will enable organizations to make informed decisions when integrating these LLMs into their recruitment processes. I recommend consulting *Chapter 2: "Algorithmic Fairness - From Parity to Pareto"* in Kearns and Roth's *"Ethical Algorithm"* book (Roth & Kearns, 2019) for further insights on this topic.

6.2. Importance of Job Advertisement

While analyzing the ratings and comments from ChatGPT and Microsoft Copilot, several patterns emerged. One notable observation, as shown in 6.1, is the misalignment between the two models. For instance, candidate ID 57 was shortlisted by Microsoft Copilot with the justification that, despite only basic German language skills, the candidate had solid business development experience. In contrast, ChatGPT rejected the same candidate, citing their limited German skills as a significant drawback, despite their extensive experience. A similar situation occurred with candidate ID 62. Both job descriptions for the Sales Manager and SAP Consultant roles clearly state that good or excellent German language skills are required. My assumption is that while this requirement is mentioned in the job ad, the models may become confused due to the lack of clear distinction between 'required' and 'preferred' qualifications in the job description.

This inconsistency becomes even more evident in 6.2, where ChatGPT provides varying decisions for four different applicants applying for the same Electrical Engineer role, which requires strong German language skills. Despite this, candidates ID 2, 5, and 6 were shortlisted based on their overall strong qualifications, even though they lacked the required German language skills. However, candidate ID 4 was rejected, with one of the reasons being their lack of German language proficiency. This highlights ChatGPT's inconsistency in balancing experience with language skills, suggesting that the model makes a trade-off between these factors.

Depending on how you interpret this issue, it may not necessarily be a negative outcome. The LLMs appear to give candidates with exceptional experience a chance, even if their German language skills are lacking, possibly assuming that the candidate could learn the language. Ultimately, the hiring manager receives the shortlist and makes the final decision on whether to invite applicants without the required German language skills to an interview. If proficiency in German is mandatory, the hiring manager can simply reject those candidates at this stage. As Selbst et al. (2019) suggest, we may be less concerned with false positives than with false negatives, as there is additional filtering during the interview process. False negatives eliminate candidates entirely, while false positives may only result in a bit more effort for the employer during interviews.

While this trade-off may not be inherently problematic, it is important for organizations using LLMs for resume screening to be aware of patterns that arise during testing. This allows them to decide whether they want the model to be more flexible, treating all job requirements as 'preferred,' or more precise in strictly distinguishing between 'preferred' and 'required' qualifications.

The significance of job advertisement wording extends beyond improving model accuracy—it is also crucial for ensuring fairness in recruitment. As Perez (2019) highlights in her book *"Invisible Women: Exposing Data Bias in a World Designed for Men"*, the language used in job descriptions plays a vital role in shaping who applies. This isn't solely about enhancing model performance; it raises ethical concerns about fairness. In male-dominated industries, such as the automotive sector, job advertisements can either encourage or deter female applicants based on how they are worded. Research by Gaucher et al. (2011) demonstrates that replacing gendered terms like "dominant," "competitive," or "leader" with more neutral language can substantially increase the number of female candidates. This issue deserves further exploration and presents a valuable opportunity for future research.

6.3. Transparency in private vs public sector

In analyzing the results, it became evident that LLMs and AI have the potential to enhance transparency in hiring practices for organizations that utilize them.

While the private sector is progressively adopting transparent recruitment practices due to legal obligations and public expectations, this transparency is not as rigorously mandated as in the public sector; it is more a matter of encouragement. Conversely, the public sector is required to maintain transparency in its hiring processes, including the reasons for applicant rejections and acceptances. This transparency is crucial to ensure fairness, accountability, and equal opportunity in recruitment. For instance, in the U.S. federal hiring process, regulations mandate that all stages of recruitment be conducted in a fair and open manner, adhering to merit-based principles. This includes transparency regarding decisions, such as the reasons for accepting or rejecting applicants, ensuring selections are based solely on candidates' abilities and qualifications (Help, 2024) (U.S. Merit Systems Protection Board, 2024). In Europe, public sector organizations are similarly required to adhere to transparency in their hiring processes under various EU regulations. Specifically, the EU Directive on Transparent and Predictable Working Conditions mandates that employers, including public bodies, clearly communicate job application procedures, reasons for applicant rejections, and criteria for selection (PwC, 2022).

AI has the potential to further enhance transparency by providing detailed feedback and data-driven insights to job applicants about their strengths and areas for development (Dattner et al., 2019). For example, when candidates are

Table 6.1.: Comparison of Shortlisting and Rejection Decisions in ChatGPT vs. Microsoft Copilot

ID (decision)	**Model**	**Comments**
57 - Dimitris Papadopoulos - Sales Manager	Copilot, 8.5 (shortlist)	Dimitris Papadopoulos is recommended for shortlisting due to his solid sales and business development experience, as well as his familiarity with the Austrian market. The lack of a technical degree and only basic German skills are areas to explore further during the interview process.
	ChatGPT, 8 (reject)	Dimitris has extensive sales experience, but his background in business administration and limited German proficiency may pose challenges in this technical and language-specific role.
62 - Kareem Khan - SAP Consultant	Copilot, 9 (shortlist)	Kareem Khan is recommended for shortlisting due to his relevant educational background, SAP expertise, and strong technical skills. His minor language limitationintermediate German proficiency is the only minor concernthat can be addressed with language training.
	ChatGPT, 7.5 (reject)	Relevant experience and technical skills, with a minor language limitation.

ranked by an AI algorithm, it is crucial to explain the basis for the ranking to ensure a fair matching procedure (Textkernel, 2023). Transparency in this context helps clarify the rationale behind the selection of shortlisted candidates and ensures that irrelevant factors, such as gender or ethnicity, do not influence the ranking (Rotaru & Kok, 2022).

However, this topic warrants further investigation, as it could significantly impact the hiring practices in both the private and public sectors. Enhanced transparency not only requires a clearer explanation of hiring decisions but also involves providing constructive feedback to rejected candidates.

For instance, consider the following rejection examples from the experiment conducted in this thesis for the applicant with candidate ID 4, who applied for the position of Electrical Engineer. ChatGPT rejected the applicant with a rating of 5.5 out of 10, citing: ***Entry-level candidate with no significant sales experience; Lacks proficiency in German and relevant technical software experience.***

Table 6.2.: Comparison of Shortlisting and Rejection Decision in Chatgpt for the same Job Position

ID (decision)	**GPT (Rating)**	**Comments**
2 - Emma Chen - Electrical Engineer	8.5 (shortlist)	Good balance of technical and sales experience; Multilingual capabilities (English, Mandarin, Japanese); Slightly less work experience but still a strong contender.
4 - Alessia Ferrari - Electrical Engineer	5.5 (reject)	Entry-level candidate with no significant sales experience; Lacks proficiency in German and relevant technical software experience.
5 - Sonja Ruppi - Electrical Engineer	9.5 (shortlist)	Extensive experience in both technical and sales roles; Leadership in managing projects and strategic sales plans; Ph.D. in Electrical Engineering and PMP certification; Some weakness in German proficiency but overall a strong candidate.
6 - Andra Benali - Electrical Engineer	8.5 (shortlist)	Extensive experience in technical and sales roles; Fluent in English and French, extensive project management experience; Not fluent in German but has strong overall qualifications.

Microsoft rejected the applicant with a rating of 6.0 out of 10, citing: ***Alesia Ferrari is not shortlisted due to her lack of the required*** 5 ***years of practical experience, unspecified experience with EPLAN P8 and SAP, and no mention of German language skills, which are important for the role. Her profile suggests potential for growth and contribution in a dynamic environment, but it does not fully meet the criteria for the Electrical Engineer position.***

Receiving such detailed justifications could greatly enhance the transparency and clarity of the hiring process, as opposed to a generic rejection message like: ***Thank you once again for your application for the position of XYZ and your interest in working with us. We have examined your application carefully. Unfortunately, we regret to inform you that we have chosen candidates for the further selection process whose qualifications we believe are a slightly better***

fit for the advertised position.

Nevertheless, even though LLMs follow instructions and provide ratings and justifications based on key criteria required in job advertisements, it is important to acknowledge that the black-box nature of AI systems currently limits our ability to fully assess the extent to which LLMs enhance explainability and transparency in recruitment processes. The underlying complexities of AI decision-making processes can obscure how and why certain decisions are made, necessitating ongoing research to fully understand their impact on recruitment practices.

6.4. The Importance of Human Oversight

Building on the research conducted in this thesis, it is essential to emphasize the importance of maintaining human oversight in AI and LLM-driven processes.

While the black-box nature of AI can seem intimidating, it is essential to acknowledge rather than fear it. LLMs operate on probabilistic neural networks, which means they can yield varying results from the same input data due to the inherent randomness in the model's structure or during the training process (O'Neal, 2024). This variability presents an open technical challenge: developing algorithms and AI applications that produce explainable and consistent results (Hunkenschroer & Luetge, 2022).

AI's transformative potential lies in its ability to automate tasks traditionally performed by humans, such as "decision making" (Diakopoulos, 2019). However, the quality of these decisions is limited by the scope of automation. In recruitment, for example, there is no definitive answer to who the "best" candidate is; the decision is often subjective, influenced by the recruiter's perception and the candidate's presentation on their resume. AI operates within explicit rules and processes vast amounts of data to generate decisions based on learned representations (Fridman, 2024). As noted by Diakopoulos (2019), algorithms should not be viewed as objective decision-makers but as tools that analyze data according to a few encoded rules.

Can AI perform the tasks of a human recruiter? To some extent, yes. AI excels at analyzing content, matching candidate resumes with job descriptions, and predicting potential job performance. However, it is best used in roles such as resume screening, where AI can assist by narrowing down candidates for further review by hiring managers. AI should be viewed as an assistant rather than a replacement for recruiters and hiring managers. Companies should leverage research and experimentation to determine the most effective ways to integrate AI into their recruitment processes. It is clear that while AI can assist in decision-making, the ultimate responsibility for employment decisions rests with human agents (Lin et al., 2021).

Furthermore, there is uncertainty about what LLM APIs might be doing in the background. These models could potentially access the internet and scrape external resources to enhance their outputs. While this may impact the accuracy of the results for better or worse, it certainly complicates the evaluation of their accuracy and fairness.

6.5. Limitations and Future Work

There are several limitations that should be considered when interpreting the findings of this thesis.

Firstly, the experiment was conducted with only 10 job advertisements and 102 resumes. The methods proposed for evaluating the accuracy and fairness of LLMs in resume screening and shortlisting might yield different results if applied to a larger sample. Expanding the sample size could provide more robust and accurate insights into how well these models perform. It would be worthwhile to replicate this experiment with a larger dataset to examine whether the results remain consistent.

Secondly, in order to ensure diversity in terms of gender, age, ethnicity, experience, and educational backgrounds, I chose to generate the resumes using the free version of GPT-4. This introduces two potential limitations: First, the generated resumes followed a uniform structure, which could affect the LLMs' performance during the screening and shortlisting phases. Since both ChatGPT and Microsoft Copilot were given highly standardized input, it is unclear how they would handle more varied and complex resume structures. Second, since the resumes were created using the same model — ChatGPT — that was also used for screening and evaluation, there is a potential for bias in the selection process. The model may be inherently more familiar with the structure and content of its own outputs, which could influence the fairness and objectivity of the shortlisting process. It's important to note that all the resumes used in the experiment were generated by ChatGPT, with no other externally generated resumes included. This presents a significant limitation and highlights the need for further exploration in future research.

A minor limitation worth investigating in future research is the deeper involvement of job applicants, recruiters, and hiring managers in the LLMs' screening and shortlisting process. As Selbst et al. (2019) highlights, understanding the behaviors and perspectives of hiring managers and job candidates when using automated resume screening is just as important as understanding the role of the software itself. Some research already exists on applicants' reactions to AI interviews. For instance, a study by Lin et al. (2021) found that participants felt AI lacked certain human qualities necessary for recruiting, such as intuition, and that it made judgments based primarily on keywords, overlooking qualities that are difficult to quantify.

To add further value to this thesis, it would be interesting to gather feedback from recruiters and hiring managers regarding the screening and shortlisting results from ChatGPT and Microsoft Copilot. Similarly, exploring how job applicants feel about having their resumes screened and shortlisted by an LLM could provide valuable insights.

Future research could also focus on additional fairness criteria, such as disabilities and sexual orientations, as well as other ethical concerns like personal data privacy. For example, the implications of AI systems accessing applicants' social media profiles or potential discrimination related to career transitions or gaps warrant further investigation. A study conducted in Sweden revealed that at least half of the interviewed recruiters had reviewed applicants' social media profiles during the hiring process (Persson, 2016). While social media content is legally considered public data, the ethicality of using such data for recruitment purposes—especially when users may not have consented to data analysis—remains questionable (Dattner et al., 2019).

Overall, the findings of this thesis should be interpreted with caution, acknowledging the outlined limitations, and further research should be pursued.

7. Conclusion

This thesis examined the use LLMs, specifically ChatGPT and Microsoft Copilot, in automating the resume screening and shortlisting stages of the recruitment process. The primary research questions addressed were: (1) What is the most effective approach for assessing AI-based recruitment in terms of performance and efficiency? and (2) How can AI-driven recruitment processes be evaluated for fairness, particularly concerning gender, age, and ethnicity?

To explore these questions, an experiment was conducted in collaboration with a real-world automotive company and their HR department. The study involved 102 resumes for 10 distinct job positions, which were evaluated by both the company's recruiters and the LLMs—ChatGPT and Microsoft Copilot. The LLMs assessed the resumes on a scale of 1 to 10 and made decisions on rejection or shortlisting based on the provided job advertisements and requirements. The recruiters' shortlisted resumes were then reviewed by hiring managers for final decisions on rejections or interview invitations.

To assess the performance and accuracy of the LLMs, two methods were employed: (1) comparing the LLMs' shortlists with those of the actual recruiters, and (2) comparing the LLMs' shortlists with the final interview lists of the hiring managers. The analysis revealed that Microsoft Copilot's shortlist was more aligned with the recruiters' shortlist, whereas ChatGPT's shortlist was more consistent with the hiring managers' final decisions. Given that hiring managers make the ultimate selection decisions, using their final interview list as a benchmark for LLM performance appears to be the most suitable approach. This suggests that, in this specific context, ChatGPT demonstrated better accuracy compared to Microsoft Copilot.

Regarding the evaluation of fairness, the aggregated resume pool was used as a benchmark. Findings indicated that both LLMs may introduce biases, underscoring the need for further research into methods for quantifying fairness and balancing it with accuracy. The current fairness assessment method involved analyzing aggregated data rather than individual records, highlighting the importance of developing a robust approach to measure and quantify fairness in global recruitment processes.

Additionally, this thesis highlights the impact of job advertisement wording on the accuracy and fairness of AI-driven recruitment. It also explores the potential for LLMs to enhance transparency in both the private and public sectors. The study underscores the crucial need for human oversight in AI recruitment systems to ensure ethical and effective implementation.

Despite limitations such as the relatively small sample size and the use of ChatGPT to generate part of the dataset—while ChatGPT was also one of the LLMs used for evaluating resumes—this thesis lays the groundwork for future research into the implications of LLMs in recruitment, particularly concerning accuracy and fairness. In response to the needs of recruiters and hiring managers, a Retrieval-Augmented Generation (RAG) prototype was developed to improve interaction with AI in the resume screening and shortlisting process.

Companies aiming to integrate AI and LLMs into their recruitment processes must conduct thorough research and experimentation to fully understand the implications for accuracy and fairness. By ensuring these factors are quantifiable, organizations will be better positioned to make informed decisions and achieve an optimal balance between accuracy and fairness in their recruitment practices.

Bibliography

Ahmad, W. U., Ribeiro, M. T., Sadeqi Azer, E., Feng, Y., Rajani, N. F., & Galstyan, A. (2022). Fairness in natural language processing: A progress report. *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, 2574–2583. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.275.pdf (cit. on p. 6).

Alsharef, A., Sonia, Nassour, H., & Sharma, J. (2023). Exploring the efficiency of text-similarity measures in automated resume screening for recruitment. *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, 36–42 (cit. on p. 13).

Ambareesh, S., Thakur, N. K., Bhattarai, U., Yadav, S. K., Thakur, J. N., & Mahato, A. K. (2024). Resume shortlisting using nlp. *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, 1–5. https://doi.org/10.1109/ICDECS59733.2023.10502580 (cit. on p. 13).

AVL. (2024). About avl. https://www.avl.com/en/about-avl (cit. on p. 21).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671. https://doi.org/10.2139/ssrn.2477899 (cit. on pp. 18, 20).

BBC. (2019). Facebook settles job discrimination case. https://www.bbc.com/news/technology-47639776 (cit. on p. 16).

Beckman, J. (2024). 120+ chatbot statistics for 2024 (already mainstream). https://techreport.com/statistics/software-web/chatbot-statistics/ (cit. on p. 13).

Bernhardsson, E. (2018). Annoy: Approximate nearest neighbors in c++/python [Python package version 1.13.0]. https://pypi.org/project/annoy/ (cit. on p. 10).

Bersin, J. (2023). The role of generative ai and large language models in hr. https://joshbersin.com/2023/03/the-role-of-generative-ai-and-large-language-models-in-hr/ (cit. on p. 15).

Biswas, C., & Das, S. (2024). Aria-qa: Ai-agent based requirements inspection and analysis through question answering [Department of Computer Science and Engineering, Amity University, Kolkata, India, and Department of Environmental Science, Informatics, and Statistics, Ca' Foscari University, Venezia, Italy]. *Research Square*. https://assets-eu.researchsquare.com/files/rs-4399368/v1_covered_ed89d388-3829-4b7d-88f4-f809ada58912.pdf?c=1716311817 (cit. on p. 10).

Bogen, M. (2019). All the ways hiring algorithms can introduce bias. https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias (cit. on p. 16).

Cardillo, A. (2024). List of the best 21 large language models (llms) (july 2024) [Accessed: 2024-08-04]. (Cit. on p. 8).

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *ACM Computing Surveys*, *54*(2), 1–37. https://doi.org/10.1145/3440755 (cit. on p. 7).

Chroma, T. (2024). Chroma db. https://www.trychroma.com/ (cit. on pp. 10, 53).

ContextualAI, T. (2024). Introducing rag 2.0. https://contextual.ai/introducing-rag2/ (cit. on p. 56).

Dattner, B., Chamorro-Premuzic, T., Buchband, R., & Schettler, L. (2019). The legal and ethical implications of using ai in hiring. https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring (cit. on pp. 61, 66).

Deloitte. (2020). 2020 women in automotive industry study. https://www.deloitte.com/uk/en/Industries/industrial-construction/research/2020-women-in-automotive-industry-survey.html (cit. on pp. 18, 24).

Deshmukh, A., & Raut, A. (2024). Applying bert-based nlp for automated resume screening and candidate ranking. *Annals of Data Science*. https://doi.org/10.1007/s40745-024-00524-5 (cit. on p. 13).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805 (cit. on p. 8).

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press. https://books.google.de/books?id=mlWQDwAAQBAJ (cit. on p. 64).

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811 (cit. on p. 19).

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2024). The faiss library. https://arxiv.org/abs/2401.08281 (cit. on p. 10).

Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. https://arxiv.org/abs/2309.15217 (cit. on p. 55).

Fletcher, R., & Nielsen, K. (2024). What does the public in six countries think of generative ai in news? https://reutersinstitute.politics.ox.ac.uk/what-does-public-six-countries-think-generative-ai-news (cit. on p. 5).

Fridman, L. (2024). *Yann lecun: Meta ai, open source, limits of llms, agi and the future of ai lex fridman podcast 416*. https://www.youtube.com/watch?v=5t1vTLU7s40 (cit. on p. 64).

Future of Life Institute (FLI). (2024). Eu artificial intelligence act [Formally signed on 13 June 2024, expected publication in the EU's Official Journal in July 2024. Enters into force 20 days after publication.]. https://artificialintelligenceact.eu/ai-act-explorer/ (cit. on pp. 18, 19).

Gagandeep, J., Kaur, & Mathur, S. (2024). Evaluating and mitigating gender bias in machine learning based resume filtering. *Multimedia Tools and Applications*, *83*, 26599–26619. https://doi.org/10.1007/s11042-023-16552-x (cit. on p. 17).

Gan, C., Zhang, Q., & Mori, T. (2024). Application of llm agents in recruitment: A novel framework for resume screening [arXiv:2401.08315v1 [cs.CL]]. *arXiv preprint arXiv:2401.08315*. https://arxiv.org/abs/2401.08315v1 (cit. on pp. 1, 15).

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. https://arxiv.org/abs/2312.10997 (cit. on p. 9).

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, *101*(1), 109. https://doi.org/10.1037/a0022530 (cit. on p. 61).

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024a). Identifying and improving disability bias in gpt-based resume screening, 687–700. https://doi.org/10.1145/3630106.3658933 (cit. on p. 15).

Glazko, K., Mohammed, Y., Kosa, B., Potluri, V., & Mankoff, J. (2024b). Identifying and improving disability bias in gpt-based resume screening. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 687–700. https://doi.org/10.1145/3630106.3658933 (cit. on p. 16).

Global Industry Analysts, I. (2022). With market size valued at $276.3 billion by 2026, it's a healthy outlook for the global artificial intelligence (ai) market. https://www.prnewswire.com/news-releases/with-market-size-valued-at-276-3-billion-by-2026--its-a-healthy-outlook-for-the-global-artificial-intelligence-ai-market-301557712.html (cit. on p. 11).

Google. (2024). Official website - gemini. powered by google. https://gemini.google.com/ (cit. on p. 8).

Gotoro. (2024). The evolution of recruitment: An expedition through history. https://gotoro.io/evolution-of-recruitment/ (cit. on p. 12).

Guston, D. (2014). Understanding 'anticipatory governance'. *Social Studies of Science*, *44*, 218–242. https://doi.org/10.1177/0306312713508669 (cit. on p. 19).

Han, Y., Liu, C., & Wang, P. (2023). A comprehensive survey on vector database: Storage and retrieval technique, challenge. https://arxiv.org/abs/2310.11703 (cit. on p. 10).

Help, U. (2024). Transparency in federal hiring processes. https://www.usajobs.gov/Help/ (cit. on p. 61).

Heymans, Y. (2024). Ai sourcing data driven recruitment. https://www.herohunt.ai/blog/ai-sourcing-data-driven-recruitment-tools (cit. on p. 13).

Hunkenschroer, A., & Luetge, C. (2022). Ethics of ai-enabled recruiting and selection: A review and research agenda. *Journal of Business Ethics*, *178*, 977–1007. https://doi.org/10.1007/s10551-022-05049-6 (cit. on pp. 15, 16, 19, 64).

Ibrahim, A. (2024). Embarking on the ai adventure part 4: Building a simplified rag system. https://medium.com/javascript-in-plain-english/embarking-on-the-ai-adventure-part-4-building-a-simplified-rag-system-9aabe6fdf995 (cit. on p. 81).

Jenkins, B. (2024). Microsoft copilot vs. chatgpt: What's the difference? https://redriver.com/artificial-intelligence/microsoft-copilot-vs-chatgpt-whats-the-difference (cit. on pp. 8, 9).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7b. https://arxiv.org/abs/2310.06825 (cit. on p. 8).

Kalyan, K. S. (2024). A survey of gpt-3 family large language models including chatgpt and gpt-4. *Natural Language Processing Journal*, *6*, 100048. https://doi.org/https://doi.org/10.1016/j.nlp.2023.100048 (cit. on p. 9).

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/https://doi.org/10.1016/j.bushor.2018.08.004 (cit. on p. 11).

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. https://arxiv.org/abs/2001.08361 (cit. on p. 5).

Kieslich, K., Diakopoulos, N., & Helberger, N. (2024). Anticipating impacts: Using largescale scenario writing to explore diverse implications of generative ai in the news environment. https://arxiv.org/abs/2310.06361 (cit. on p. 19).

Kieslich, K., Lünich, M., & Došenović, P. (2023). Ever heard of ethical ai? investigating the salience of ethical ai issues among the german population. *International Journal of Human–Computer Interaction*, *40*(11), 2986–2999. https://doi.org/10.1080/10447318.2023.2178612 (cit. on p. 19).

Lahey, J. N. (2008). Age, women, and hiring: An experimental study of age discrimination. *Journal of Human Resources*, *43*(1), 30–56. https://doi.org/10.3368/jhr.43.1.30 (cit. on p. 18).

Langchain. (2024). Official website - applications that can reason. powered by langchain. https://python.langchain.com/v0.1/docs/use_cases/chatbots/ (cit. on p. 81).

LeCunn, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. http://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf (cit. on p. 5).

Levingston, E. (2024). What is mya chatbot? features, uses, and how to learn more. https://vanchat.io/post/mya-chatbot-features-uses-learn-more (cit. on p. 13).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks. https://arxiv.org/abs/2005.11401 (cit. on pp. 9, 11).

Li, J., Zhang, X., Wang, Y., Chen, S., Zhao, X., Tang, J., & Yin, D. (2023). Enhancing job recommendation through llm-based generative adversarial networks. *arXiv preprint arXiv:2307.10097*. https://arxiv.org/abs/2307.10097 (cit. on p. 1).

Lin, Y., Hung, T., & Huang, L. (2021). Engineering equity: How ai can help reduce the harm of implicit bias. *Philosophy & Technology*, *34*(Suppl 1), 65–90. https://doi.org/10.1007/s13347-020-00406-7 (cit. on pp. 64, 65).

Malach, E. (2024). Auto-regressive next-token predictors are universal learners. https://arxiv.org/abs/2309.06979 (cit. on p. 9).

McKinsey & Company. (2024). The state of generative ai in human resources. https://www.mckinsey.com/business-functions/people-and-organizational-performance/our-insights/the-state-of-generative-ai-in-human-resources (cit. on p. 14).

Mdwoicke. (2024). Ragfusion node / javascript implementation. https://gist.github.com/mdwoicke/da86e5cbf45239f6afdb3d378fa5ceaa (cit. on p. 81).

Meer, D. V. (2024). Number of chatgpt users and key stats (august 2024). https://www.namepepper.com/chatgpt-users (cit. on p. 14).

Microsoft. (2024). Microsoft copilot. https://copilot.microsoft.com/ (cit. on pp. 31, 35).

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:5959482 (cit. on p. 6).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:16447573 (cit. on p. 6).

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. https://arxiv.org/abs/2402.06196 (cit. on p. 5).
Mohammadreza Farrokhnia, O. N., Seyyed Kazem Banihashem, & Wals, A. (2024). A swot analysis of chatgpt: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846 (cit. on p. 8).
Montreal Ethics. (2020a). Why was your job application rejected: Bias in recruitment algorithms? (part 2). https://montrealethics.ai/why-was-your-job-application-rejected-bias-in-recruitment-algorithms-part-2/ (cit. on p. 24).
MontrealEthics. (2020b). Why was your job application rejected: Bias in recruitment algorithms? (part 1). https://montrealethics.ai/why-was-your-job-application-rejected-bias-in-recruitment-algorithms-part-1/ (cit. on p. 16).
Mujtaba, D. F., & Mahapatra, N. R. (2019). Ethical considerations in ai-based recruitment. *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 1–7. https://doi.org/10.1109/ISTAS48451.2019.8937920 (cit. on p. 16).
Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW), 36. https://doi.org/10.1145/3359221 (cit. on p. 17).
Nguyen, H. (2024). *Enhancing resume screening efficiency and quality with retrieval augmented generation*. https://github.com/Hungreeee/Resume-Screening-RAG-Pipeline/tree/main (cit. on pp. 9, 53, 55, 57, 81).
Northreach. (2023). Harnessing the power of ai, chatgpt & llms in recruitment. https://www.linkedin.com/pulse/harnessing-power-ai-chatgpt-llms-recruitment-northreach-uk/ (cit. on p. 13).
Official website - cegid. (2024). https://www.cegid.com/global/talentsoft-joins-cegid/ (cit. on p. 13).
Official website - herohunt. (2024). https://www.herohunt.ai/ (cit. on p. 13).
Official website - hirevue. (2024). https://www.pymetrics.ai/ (cit. on p. 14).
Official website - ideal. (2024). https://ideal.com/ (cit. on p. 14).
Official website - olivia. (2024). https://olivia.chatbotportal.com/home (cit. on p. 13).
Official website - pymetrics. (2024). https://www.pymetrics.ai/ (cit. on p. 14).
Official website - zoho recruit. (2024). https://www.zoho.com/recruit/resume-management.html (cit. on p. 14).
O'Neal, H. (2024). How large language models (llms) are reshaping hr management. https://www.metadialog.com/blog/large-language-models-for-hr-how-to-use-it-in-human-resource/ (cit. on pp. 14, 15, 64).

OpenAI. (2019). Microsoft and openai. https://openai.com/blog/microsoft/ (cit. on p. 9).
OpenAI. (2022). Introducing chatgpt. https://openai.com/index/chatgpt/ (cit. on p. 8).
OpenAI. (2024a). Chatgpt [large language model]. https://chat.openai.com/chat (cit. on pp. 24, 33).
OpenAI. (2024b). Official website - dall-e. powered by openai. https://openai.com/index/dall-e-2/ (cit. on p. 8).
Pan, J. J., Wang, J., & Li, G. (2023). Survey of vector database management systems. https://arxiv.org/abs/2310.14021 (cit. on p. 53).
Pearce, T., Brintrup, A., & Zhu, J. (2021). Understanding softmax confidence and uncertainty. https://arxiv.org/abs/2106.04972 (cit. on p. 9).
Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). Check your facts and try again: Improving large language models with external knowledge and automated feedback. https://arxiv.org/abs/2302.12813 (cit. on p. 55).
Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162 (cit. on p. 6).
Perez, C. C. (2019). *Invisible women: Exposing data bias in a world designed for men* [The Sunday Times number one bestseller, nominated for the Royal Society Insight Investment Science Book Prize 2019]. Chatto & Windus. (Cit. on p. 61).
Perplexity. (2024). Official website - perplexity ai. https://www.perplexity.ai/ (cit. on p. 8).
Persson, A. (2016, August). Implicit bias in predictive data profiling within recruitments. https://doi.org/10.1007/978-3-319-55783-0_15 (cit. on pp. 16, 66).
Pisanelli, E. (2022). Your resume is your gatekeeper: Automated resume screening as a strategy to reduce gender gaps in hiring. *Economics Letters*, *221*, 110892. https://doi.org/https://doi.org/10.1016/j.econlet.2022.110892 (cit. on p. 17).
Prompts - langchain docs. (2024). https://docs.langchain.com/docs/components/prompts (cit. on p. 34).
PwC. (2022). Eu directive on transparent and predictable working conditions. https://www.pwc.nl/en/topics/eu-directive-transparent-predictable-working-conditions.html (cit. on p. 61).
Rackauckas, Z. (2024). Ragfusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, *13*(1), 37–47. https://doi.org/10.5121/ijnlc.2024.13103 (cit. on pp. 56, 57).

Raudaschl, A. ((2023). Rag-fusion: The next frontier of search technology. https://github.com/Raudaschl/rag-fusion (cit. on p. 81).

Rigotti, C., Puttick, A., Fosch-Villaronga, E., & Kurpicz-Briki, M. (2023). Mitigating diversity biases of ai in the labor market. *European Workshop on Algorithmic Fairness*, 1–8. https://www.biasproject.eu/wp-content/uploads/2023/03/paper-47.pdf (cit. on pp. 1, 2).

Rotaru, M., & Kok, K. (2022). Seven limitations of large language models (llms) in recruitment technology. https://www.textkernel.com/learn-support/blog/seven-limitations-of-llms-in-hr-tech/ (cit. on p. 62).

Roth, A., & Kearns, M. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press. https://global.oup.com/academic/product/the-ethical-algorithm-9780190948207 (cit. on pp. 19, 20, 60).

Roy, M., & Sawant, R. (2024). A guide to llm agents for recruitments: 4 use cases & examples. https://www.ionio.ai/blog/a-guide-to-llm-agents-for-recruitments-4-use-cases-examples (cit. on pp. 12, 13).

Roy, P. K., Chowdhary, S. S., & Bhatia, R. (2020). A machine learning approach for automation of resume recommendation system [International Conference on Computational Intelligence and Data Science]. *Procedia Computer Science*, *167*, 2318–2327. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.284 (cit. on p. 13).

Sampath, P. (2024). The animated transformer. https://prvnsmpth.github.io/animated-transformer/ (cit. on p. 6).

Sánchez-Monedero, J., Dencik, L., & Edwards, L. (2020). What does it mean to 'solve' the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 458–468. https://doi.org/10.1145/3351095.3372849 (cit. on pp. 16, 19).

SAP. (2024). Human capital management. https://www.sap.com/products/hcm.html (cit. on p. 23).

Sawarkar, K., Mangal, A., & Solanki, S. R. (2024). Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. https://arxiv.org/abs/2404.07220 (cit. on p. 57).

Schmidhuber, J. (2014). Deep learning in neural networks: An overview. https://people.idsia.ch/~juergen/DeepLearning2July2014.pdf (cit. on p. 5).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598 (cit. on pp. 17, 20, 60, 65).

Semetaite, J. (2023). 9 resume screening tools to boost your hiring. https://toggl.com/blog/resume-screening-tools (cit. on p. 14).

Spataro, J. (2023). Introducing microsoft 365 copilot – your copilot for work. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ (cit. on pp. 8, 9).

Spice, B. (2015). Fewer women than men are shown online ads related to high-paying jobs. https://csd.cmu.edu/news/fewer-women-than-men-are-shown-online-ads-related-to-highpaying-jobs (cit. on p. 16).

Steck, H., Ekanadham, C., & Kallus, N. (2024). Is cosine-similarity of embeddings really about similarity? *Companion Proceedings of the ACM on Web Conference 2024*. https://doi.org/10.1145/3589335.3651526 (cit. on p. 7).

Tayal, D., Sharma, T., Singhal, S., & Thakur, A. (2024). Resume screening using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*, 602–606. https://doi.org/10.32628/CSEIT2410275 (cit. on p. 13).

Textkernel. (2023). Seven limitations of LLMs in HR tech. https://www.textkernel.com/learn-support/blog/seven-limitations-of-llms-in-hr-tech/ (cit. on p. 62).

Thanasi-Boçe, M., & Hoxha, J. (2024). From ideas to ventures: Building entrepreneurship knowledge with llm, prompt engineering, and conversational agents. *Education and Information Technologies*. https://doi.org/10.1007/s10639-024-12775-z (cit. on pp. 34, 51).

The Ladders. (2018). Eye-tracking study: How long do recruiters spend on your resume? https://www.theladders.com/career-advice/eye-tracking-study-how-long-do-recruiters-spend-on-your-resume (cit. on p. 12).

U.S. Merit Systems Protection Board. (2024). Fair and open hiring practices in the u.s. public sector. https://www.mspb.gov/ (cit. on p. 61).

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:13756489 (cit. on p. 6).

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z.-Y., Tang, J., Chen, X., Lin, Y., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 1–42. https://doi.org/10.1007/s11704-024-40231-1 (cit. on p. 14).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., . . . Rush, A. M. (2020). Huggingface's transformers: State-of-the-art natural language processing. https://arxiv.org/abs/1910.03771 (cit. on p. 53).

Appendix

Appendix A.

Appendix: Dataset and Code Repository

The dataset used in this experiment, which includes job advertisements and resumes, can be accessed through the following link: Link to Dataset: Job Advertisements and Resumes This dataset contains a total of 102 resumes and 10 job advertisements, as described in Chapter 3.4. The data is structured for use in evaluating the shortlisting processes of real recruiters, ChatGPT, and Microsoft Copilot.

The code for the app used in the experiment, presented in Chapter 5, can be found in the following GitHub repository. This repository includes all necessary files for re-implementation and further experimentation: Link to GitHub Repository: App for Resume Shortlisting Use Case.

A tutorial demonstrating how to use the app and its features is available at the following link: RAG Automated Recruiting Tutorial.

The use-case is inspired by (Mdwoicke, 2024; Nguyen, 2024; Raudaschl, 2023) and (Ibrahim, 2024). It leverages Langchain AI wrapper tools (Langchain, 2024) and incorporates frontend components primarily derived from the open-source library: Minimal UI.

Digrams.net is used to generate all diagrams, icons and images presented in this thesis.