Future Learning, Analytics & AI for Teaching

TU Graz
Graz University of Technology

Evaluation of AI Generated Educational Illustrations

A Practical Framework for BA Theses

Dr. Sandra Schön – https://orcid.org/0000-0003-0267-5215
Dr. Benedikt Brünner – https://orcid.org/0009-0005-8484-9160
Dr. Martin Ebner – https://orcid.org/0000-0001-5789-5296

Graz University of Technology, Austria

Version 1.1 – November 10, 2026

Contact: sandra.schoen@tugraz.at

Published under CC BY 4.0, if not stated differently (see cited Figures)

Citation suggestion: Schön, S., Brünner, B. & Ebner, M. (2026). Evaluation of AI generated illustrations for learning. A Practical Framework for BA Theses. November 11, 2026. Graz University of Technology. DOI: 10.3217/vszsa-1ve68

1. Introduction

Visual illustrations play a valuable role in educational contexts, ranging from school classrooms to higher education institutions. They serve not only as cognitive aids but also as essential tools for simplifying complex concepts and enhancing learner engagement (cf. Wang et al., 2025 for arguments and empirical insights). As Generative Artificial Intelligence (GenAI) rapidly evolves, a critical question emerges: To what extent are current GenAI models suitable for designing high-quality educational illustrations?

In a recent whitepaper (Schön et al., 2026), we established a general framework for the evaluation and integration of GenAI within university teaching. That document provided an overview of various methodologies for assessing the quality of AI-generated outputs across different domains. However, specific challenges arise when applying these broad principles to visual content, particularly regarding pedagogical accuracy, semantic coherence, and technical fidelity in learning materials.

To address this gap, we propose a tailored set of quality criteria specifically designed for bachelor's theses supervised by us. These criteria are intended to guide students in systematically evaluating the images they generate or analyse. By standardizing the assessment metrics, we aim to ensure that results from diverse research projects remain comparable, thereby contributing to more robust evidence base on the reliability and efficacy of AI-generated visuals for different educational settings.

2. Notes on piloting and conducting multi-layered evaluation when using AI for learning content creation (and more)

To conduct rigorous research on the quality of Generative AI (GenAI) outputs, researchers must first establish a clear methodology comprising defined test cases, evaluation rubrics, reviewer roles, and documentation protocols tailored to the specific purpose of the assessment.

For bachelor's thesis projects, we recommend that the student assume the role of the domain expert; therefore, they must possess sufficient knowledge in their chosen subject area to make informed judgments. Depending on the specific use case, the evaluation should encompass a comprehensive range of criteria, including factual correctness, pedagogical utility, accessibility, bias detection, robustness, safety, citation integrity, format compliance, usability, and alignment with task constraints. While AI-assisted review tools can support this process, they must never replace critical human judgment.

As illustrated in Figure 1, accurately assessing GenAI performance requires the systematic description and evaluation of several interconnected factors. It is important to note that the evaluation strategy may vary depending on the research domain, specific case study, or focus of inquiry (e.g., analysing gender bias).

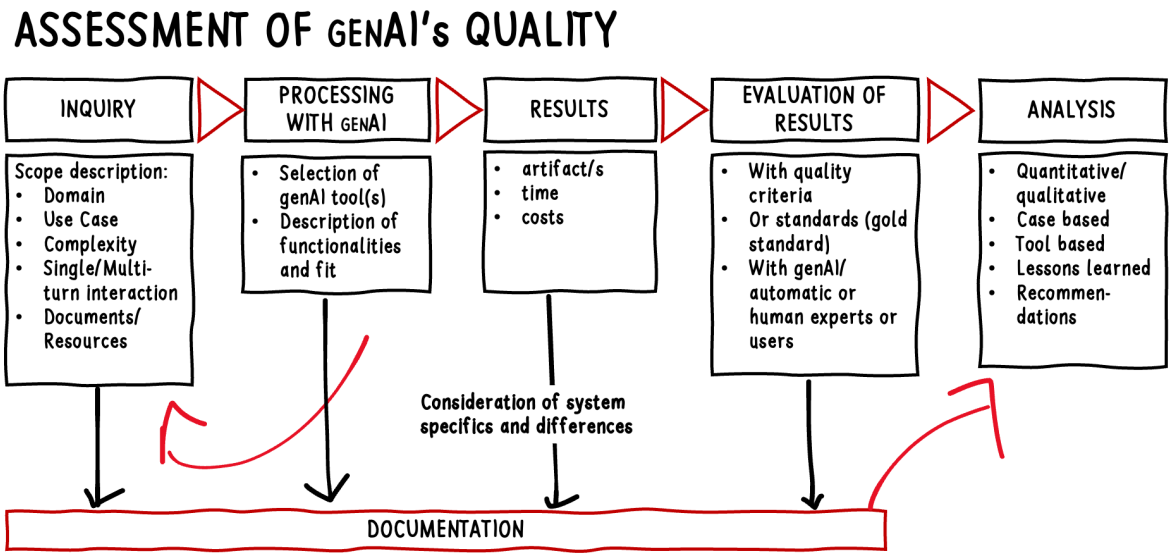


Figure 1: Assessment approaches for GenAI results and usage. Source: Schön et al. 2026, Fig.5

Following the prototype and evaluation phase, a controlled pilot and deployment is mandatory. A pilot precedes full-scale deployment to observe whether users understand the system, if the tool integrates meaningfully into learning activities, if learners exhibit over-reliance, if teachers can facilitate it effectively, and if unforeseen risks emerge.

3. Defining target audience and learning objectives

A meaningful evaluation of AI-generated educational illustrations presupposes a precise description of the intended target audience and the specific learning objectives. Only when it is clear who the learners are — e.g. primary school children, BA engineering students, or medical professionals — can criteria such as legibility, conceptual depth, and visual appropriateness be applied in a purposeful manner. Equally important is a clear articulation of what learners should understand or be able to do after engaging with the illustration, as this defines the standard against which conceptual representation and educational usability are measured. For instance, an illustration of the water cycle designed for Year 4 pupils in Austrian schools may prioritise clarity and minimal labelling, whereas the same topic prepared for university-level geography students would be expected to include technical terminology and process detail. Similarly, an image intended to support a single learning objective — such as identifying the stages of mitosis — requires a different evaluative focus than one meant to provide a comprehensive overview of cell biology. Without this contextual framing, evaluation criteria risk being applied inconsistently or too generically, undermining the comparability of results across different studies and settings.

4. Quality check of AI generated learning illustrations: Existing criteria from literature

Publication on computer generated imagery (Rehak, 2011) and AI generated learning and educational illustrations can be found with a search at e.g. Google Scholar or ERIC.

- "AI-generated" AND "educational illustrations"
- "Computer-generated imagery"
- "Generative AI" AND "visual learning materials"
- "Text-to-image" AND "instructional design"

There is already a systematic review on AI generated visual content in education available (Wang et al., 2025) which shows, that research is currently done most in the field of design and creative arts education and medicine (Wang et al., 2025 Fig. 4). This includes e.g. qualitive studies on how teachers might benefit and experience the tools and results (e.g. Mirici, 2026).

Several research papers on AI generated educational illustration used criteria to assess the prompting results. Some are described in the following:

Temsah et al. (2024) evaluated the AI generated Imagery of congenital heart diseases. They developed a detailed list of aspect that needed to be displayed and applied a collection of data as shown in Fig. 2:

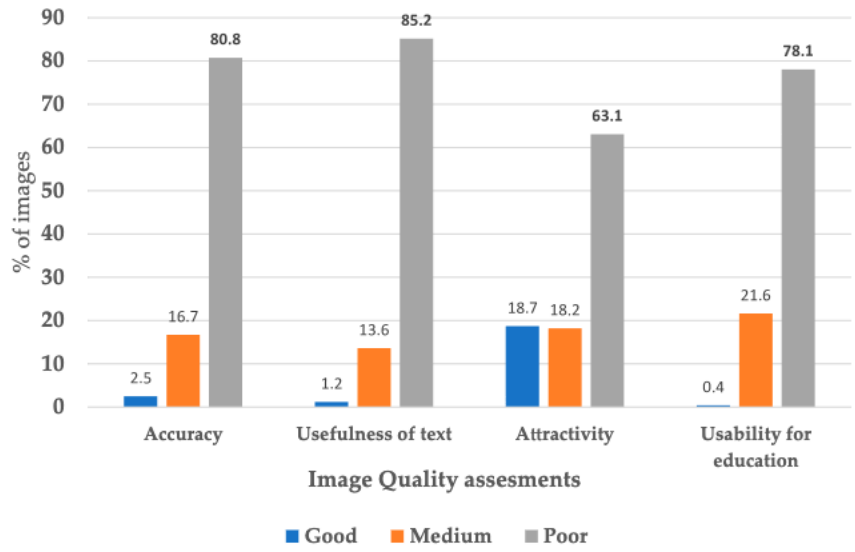


Figure 2: The overall rating of the AI-generated congenital cardiac anomalies images (N = 3630 ratings). Source: Temsa et al. (2024), Fig. 2, P. 54.
Please note: This figure is excluded from the CC BY 4.0 license

In details, this study used the following criteria and skales (see Temsa et al., Table 1, p. 54, highlights by authors):

- **Anatomic Accuracy** (accurate, midway, fabricated)
- **Illustrative text usefulness** (usefull, midway, not useful)
- **Usability for medical education** (usable as is, after modification, not useable)
- **Attractivity** (attractive, midway, not attractive)

Hnidenko (2025) evaluated images for schoolbooks according to the following criteria:

- "**Technical accuracy** – the extent to which the image matched the expected form and was free from artifacts;
- **Prompt correspondence** – whether the instructions regarding style, background, composition, etc., were followed;
- **Generation speed** – the time required to obtain an acceptable result." (highlights by authors p. 22)

Rakhski et al. (2025) wanted to know if AI can illustrate science. They used the following criteria (p. 29):

- **Conceptual capture** (acceptable / not acceptable)
- **Contextual relatedness** (acceptable / not acceptable)
- **Factual accuracy** (acceptable / not acceptable)
- **Aesthetic visual quality** (acceptable / not acceptable)

A bachelor thesis at TU Graz suggested the following criteria (see Maier, 2026, shortened and translated by authors):

- **Linguistic Accuracy:** No spelling/grammar errors in embedded text.
- **Factual Correctness:** Accurate content; correct labeling of components.
- **Clarity:** Logical layout; easy to follow; no confusion.
- **Informational Value:** Adequate detail; answers learner questions; educational utility.
- **Visual Presentation:** Appropriate use of icons/colors; high legibility of all elements.

It is to be noted, that many studies in the field of medical information uses the Global Quality Score (GQS, see Bernard, 2007[1]), such as Hamada et al. (2026) for genAI visualisation on impacted third molar extraction.

[1] It is the "Global Quality Score", not "Scale", see: Yeung (2025).

5. Suggestion of quality criteria for research work on educational illustrations

Building upon existing research and established evaluation frameworks, we have synthesized a set of core mandatory criteria that must be assessed for all generated results/artefacts to ensure a future comparability of data.

We suggest using the following criteria when evaluating quality of educational illustrated generated with AI:

#	Criterion (EN)	Definition (EN)	Kriterium (DE)	Definition (DE)
C	Conceptual Representation	Are the central concepts for the learning setting adequately depicted and visually represented?	Konzeptdarstellung	Sind die zentralen Konzepte für das Lernsetting angemessen dargestellt und visuell abgebildet?
V	Visual Correctness	Is the image free from AI artifacts, distortions, impossible geometry, or structural flaws?	Visuelle Korrektheit	Ist das Bild frei von KI-Artefakten, Verzerrungen, unmöglicher Geometrie oder Strukturfehlern?
L	Legibility & Labeling	Is text readable and labels correct/logical?	Lesbarkeit & Beschriftung	Ist Text lesbar und Beschriftung korrekt/logisch?
O	Overall Educational Usability	Based on the above: Is this image suitable for the intended educational purpose?	Gesamturteil Bildungseignung	Auf Basis der obigen Aspekte: Ist dieses Bild für den vorgesehenen Bildungszweck geeignet und nutzbar?

We suggest the following rating scale for the evaluation:

Value	English Definition	Deutsche Definition
1 ❌	Poor/Unacceptable: Critical failure; image unusable without major reconstruction	Schwach/unakzeptabel: Kritischer Fehler; Bild nur durch Neugenerierung nutzbar
2 ⚠️	Moderate/Conditional: Minor issues present; fixable with editing/modification	Mittel/mit Einschränkung: Kleinere Mängel vorhanden; durch Nachbearbeitung korrigierbar
3 ✅	Excellent/Acceptable: No significant errors; ready for educational use	Ausgezeichnet/unbedenklich: Keine signifikanten Fehler; sofort einsatzberei

Additionally, the following criteria (or others) might be considered to additionally be used, in relation to your case, domain, and focus of work:

- Gender, race, ethnicity aspects (see e.g. Hosseini, 2024; Andrew & Montalbano, M, 2025)
- Prompt Adherence: Were style, composition, and background instructions followed as specified?
- Aesthetic Appeal: Is the visual design professional, engaging, and consistent with educational standards?
- Cultural Appropriateness: Are visual conventions (directionality, flow, colour symbolism, gestures) aligned with the cultural context of the target audience?
- Informational Depth: Does the image provide sufficient detail to answer learner questions or explain complex concepts effectively?
- Contextual Relevance: Do all visual elements contribute meaningfully to the educational topic? Are irrelevant background objects excluded?
- Educational Utility: Would this image significantly enhance learning outcomes compared to text-only descriptions?
- Complexity: Is the level of detail and visual complexity appropriate for the target audience and the intended learning objective? (Too little may oversimplify; too much may overwhelm or distract.)
- Absence of Redundancy: Are there distracting duplicate elements in the image? (That this might occur is one insight from Maier, 2026).

6. Good prompting

When evaluating AI-generated illustrations from a teaching perspective, researchers should invest effort in crafting well-structured prompts, as low-quality outputs may stem from inadequate prompting rather than limitations of the AI model itself. A good prompt clearly specifies the target audience (e.g., “Grade 4 pupils” or “BA level biology students”), which determines appropriate complexity, terminology, and visual style. Equally important is articulating the specific learning objective the illustration should support, such as “showing the three stages of photosynthesis for exam preparation”. The prompt should also indicate desired format characteristics, including composition (“infographic layout”), labelling requirements (“annotated labels in English”), and aesthetic direction (“clean line drawing, no photographic realism”). Including negative constraints can further refine outputs, for example “no unrealistic human figures” or “avoid cluttered backgrounds”. Ambiguous requests like “make it educational” should be avoided; instead, use concrete descriptors that convey exactly what pedagogical function the image needs to fulfil.

Researchers are encouraged to iterate on their prompts if initial results do not meet expectations, documenting all versions for transparency. This practice ensures fair evaluation of the AI tool's capabilities.

7. Documentation and publication of evaluation data

To ensure transparent and comparable evaluation of AI-generated educational illustrations, all relevant data shall be recorded in a structured table. This table serves as the central documentation of the evaluation process and will be published together with the illustrations under an open licence.

The structured format also enables comparative analysis across multiple bachelor's theses. Should data from different works be pooled for joint analysis or meta-evaluation, all contributing authors will be acknowledged appropriately according to the principles of good scientific practice.

7.1 Data storage

All illustrations generated and evaluated in the course of the bachelor's thesis, along with the structured evaluation table, must be stored by the student in a dedicated repository on TU Graz's GitLab service (gitlab.tugraz.at). We recommend one repository per thesis (e.g. following the pattern "edu-illustrations-ba-surname") to ensure unambiguous assignment and access management.

Supervisors must be granted access to the repository, so that the evaluation process can be reviewed and accompanied continuously. The repository must always reflect the structure described in section 7.2. Evaluations shall be documented in the repository promptly after generation; any changes to ratings or revised versions of illustrations must be traceable through commits.

During work on the thesis, the repository may be kept private; upon completion it shall be published under an open license (see 7.3). Complete and unaltered documentation of the prompts in the repository is essential for reproducibility.

7.2 Procedure

Each generated and evaluated illustration receives its own entry in the table. One set of evaluation criteria is applied per illustration; if the same illustration is evaluated multiple times (e.g. by different reviewers), a separate row is created for each assessment.

Table Structure

The following columns are mandatory:

Column	Description
Image ID	Unique identifier, e.g. LA-001, LA-002, etc. Allows unambiguous assignment to the image file.
Subject / Process	Brief description of the learning content or process depicted, e.g. "Cell division – Mitosis".
Tool (Version)	AI tool used, including version number, e.g. "Lumo Image Gen v2.0" or "DALL·E 3".
Date	Date of generation or evaluation in ISO 8601 format (YYYY-MM-DD).
Prompt	The complete, unaltered input prompt used to generate the illustration.

File Location / Image Reference	Path or link to the stored image, e.g. repository URL or filename in the shared folder.
Evaluated by	Name of the reviewer. An abbreviation may be used if anonymity is preferred.
C – Conceptual Representation	Rating (1–3).
V – Visual Correctness	Rating (1–3).
L – Legibility & Labeling	Rating (1–3).
O – Overall Educational Usability	Rating (1–3).
Additional Criteria *(optional)*	Further criteria as deemed relevant for the specific case (e.g. Prompt Adherence, Gender/Race aspects, Informational Depth).
Notes	Optional: brief remarks, e.g. justification of the rating or notable observations.

Rating Scale

The following rating scale applies to all four core criteria (C, V, L, O):

- 1 - Poor / Unacceptable – Critical failure; image unusable without major reconstruction
- 2 - Moderate / Conditional – Minor issues present; fixable with editing or modification
- 3 - Excellent / Acceptable – No significant errors; ready for educational use

Example Row

Image ID	Subject / Process	Tool (Version)	Date	Prompt	File Location	Evaluated by	C	V	L	O	Notes
LA-003	Photosynthesis	Lumo Image Gen v2.0	2026-07-06	"Photosynthesis process diagram, labelled in German, clean infographic style, 4K"	/illustrations/LA-003.png	Micha	3	2	3	3	Small artefact in upper left corner; otherwise suitable for use

Storage Format

- The table is saved as a CSV file to ensure easy processing and maximum compatibility.
- An additional copy may be kept in ODS (OpenDocument Spreadsheet) or XLSX format.

- Image files are stored in PNG format (for diagrams and schematics) or JPEG (for photorealistic illustrations) and named according to their respective Image ID, e.g. LA-003.png.
- The folder /golden-standard/ contains human-made reference illustrations used for comparison. Each image is stored in its own subfolder (e.g. GS-001) together with a licence file; only openly licensed images may be used (see 7.3).

Directory Structure (Example)

```
/educational_illustrations/
├── evaluation_table.csv
├── evaluation_table.ods
├── golden-standard/
├── golden-standard/
│   └── GS-001/
│       ├── GS-001.png
│       └── GS-001-LICENSE.md
├── illustrations/
│   ├── LA-001.png
│   ├── LA-002.png
│   └── LA-003.png
└── LICENSE.md
```

Notes

- Evaluations should be carried out promptly after generation to ensure consistent assessment.
- If an illustration is revised or regenerated, the new version receives a new Image ID so that previous evaluations remain traceable.
- The Prompt column is particularly important for reproducibility and should therefore be documented completely and unaltered.
- Additional criteria beyond the four core criteria (C, V, L, O) may be added depending on the research focus, domain, or specific use case (see Section 4 of the framework document for suggested additional criteria).

7.3 Licensing and Publication

All evaluation data and associated illustrations should be shared with us and / or published under an open licence that permits use, sharing, and adaptation. The following licensing is recommended:

- Illustrations: CC-0 (AI generated materials are public domain)
- Evaluation data (table): CC BY 4.0 (Creative Commons Attribution 4.0)
- Golden standard images: Reference images stored in the folder /golden-standard/ must be openly licensed (CC0, CC BY or CC BY-SA), e.g. images from

Wikimedia Commons. Source, author and licence of each image must be documented in the corresponding LICENSE file (e.g. GS-001-LICENSE.md). Images under more restrictive licences (e.g. NC or ND) or without a clear licence must not be used.

The chosen licence is declared in the LICENSE.md file within the publication directory. In addition, licence information and attribution requirements should be clearly stated in the table or in an accompanying README file.

8. Changes and open questions

Please note that we might update this document building up on the ongoing research work and insights. Please contact us for any missing information or open issue.

Disclaimer on AI usage

Based on Academic Integrity and Transparency in AI-assisted Research and Specification Framework (Bozkurt, 2024), the authors of this work acknowledge that it was drafted with the assistance of Lumo and DeepL (Versions as of June 2026), complementing the human editorial process. The human authors critically assessed and validated the content to maintain academic rigor. The authors also assessed and addressed potential biases inherent in the AI-generated content. The final version of the paper is the sole responsibility of the human authors.

References

Andrew, K., & Montalbano, M. J. (2025). Through a glass darkly: Perceptions of ethnoracial identity in artificial intelligence generated medical vignettes and images. Medical Science Educator, 1-16. https://doi.org/10.1007/s40670-025-02332-9

Bernard, A., Langille, M., Hughes, S., Rose, C., Leddin, D., & Van Zanten, S. V. (2007). A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. American Journal of Gastroenterology, 102(9), 2070–2077. https://doi.org/10.1111/j.1572-0241.2007.01325.x

Hamada, M., Nishiyama, K., Ichiyama, S., Yamamoto, S., Motoki, A., Fujita, Y., ... & Uzawa, N. (2026). Evaluation of text-and image-based generative artificial intelligence for simulating impacted third molar extraction. Journal of Stomatology Oral and Maxillofacial Surgery, 127(6), 102863. https://doi.org/10.1016/j.jormas.2026.102863

Hosseini, D. D. (2024). Generative AI: A problematic illustration of the intersections of racialized gender, race, ethnicity. OSF Prepr. https://doi.org/10.31219/osf.io/987ra

Hnidenko, S. (2025). The use of artificial intelligence for image generation in next-generation school textbooks. Viae Educationis, 4(1), 18-25. https://bibliotekanauki.pl/articles/63786316.pdf

Maier, A. (2026). Vergleich von ChatGPT Image Generator mit Nano Banana und Flux AI für OER-Inhalte [Bachelorarbeit, Technische Universität Graz].

Mirici, I. (2026). Image-Generating AI In ELT Material Development: High School Teachers' Perceptions and Support Needs. International Online Journal of Education & Teaching, 13(1), 165. https://iojet.org/index.php/IOJET/article/view/2264

Rakhshi, R., Khosravi, T., Rahimzadeh, A., Shahraki, F., & Oladnabi, M. (2025). Can AI illustrate science? A comparative benchmarking study of text-to-image artificial intelligence models for scientific communication. Journal of Clinical and Basic Research, 9(2), 28-33. https://jcbr.goums.ac.ir/files/site1/user_files_eca66e/oladnabidozin-A-10-127-8-8143efb.pdf

Rehak, B. (2011). "Computer-Generated Imagery". Cinema And Media Studies. DOI: 10.1093/obo/9780199791286-0068 https://works.swarthmore.edu/fac-film-media/22

Schön, S., Brünner, B., Ebner, M., & Leitner, P. (2026). Evaluating GenAI Innovation in Higher Education. A Whitepaper (Version 1.0 - May 22, 2026). Graz University of Technology. https://doi.org/10.3217/1tfkj-ntj85

Temsah, MH., Alhuzaimi, A.N., Almansour, M. et al. Art or Artifact: Evaluating the Accuracy, Appeal, and Educational Value of AI-Generated Imagery in DALL·E 3 for Illustrating Congenital Heart Diseases. J Med Syst 48, 54 (2024). https://doi.org/10.1007/s10916-024-02072-0

Wang, D., Wei, S., Chen, X., Chai, C. S., Chen, G., & Hu, L. (2025). AI-Generated Visual Content in Education: A Systematic Review of the Past Decade. https://doi.org/10.2139/ssrn.5935496

Yeung, A. W. K. (2025). Evaluation of content quality of online health information by global quality score: a case study of researchers misnaming it and citing secondary sources. *Publications*, *13*(2), 23. https://doi.org/10.3390/publications13020023