The X-AIssessment Assistant: Prototyping an Explainable GenAI-Powered Platform for Student Essay Assessment Through Design-Based Research

Tine Keulemans[1,2,*,†], Grzegorz Meller[2,3,4,†], Elke Vandermeerschen[2], Wim Van Den Noortgate[5], Katrien Verbert[3,4], Fien Depaepe[1,5], Annelies Raes[1] and Tinne De Laet[2]

[1]*Center of Instructional Psychology and Technology (CIP&T)*
[2]*Leuven Engineering & Science Education Center (LESEC)*
[3]*KU Leuven Department of Computer Science, B-3001 Leuven, Belgium*
[4]*Augment, an imec research group*
[5]*Itec, an imec research group*
KU Leuven, Dekenstraat 2, 3000 Leuven, Belgium

Abstract

Scoring students' essays is time-consuming and cognitively demanding for educators in higher education. Generative artificial intelligence (GenAI) offers promising opportunities for automated essay scoring (AES). Yet concerns like potential inaccuracies and reduced professional autonomy persist. As a result, there is a growing focus on how educators and AES systems can collaborate. Central to such collaboration is explainability: educators must understand the rationale behind AI-generated scores and feedback to evaluate and, when necessary, adapt them. However, research on human-centered explainability approaches in GenAI-based scoring systems is scarce. To address this gap, we developed *the X-AIssessment Assistant*, an explainable GenAI-powered scoring platform to support educators in assessing student essays in higher education, following a design-based research approach (DBR). This paper reports on (1) the design and development of a prototype based on two participatory design sessions and state of the art research in human-computer interaction and AES, and (2) one DBR-cycle involving a participatory design session with university educators. Participants' perceptions of the prototype were analyzed using the Theory of Planned Behavior. The findings show that educators' intentions to engage with the platform are shaped by their attitudes towards the system (e.g., perceived workload), subjective norms (e.g., assessment as a core pedagogical responsibility), and perceived behavioral control (e.g., the ability to override AI judgments). Based on these insights, we derive design guidelines for explainable GenAI-based essay assessment systems that foster human-AI collaboration in higher education.

Keywords

Explainable AI, Human-AI collaboration, Essay, Higher Education

1. Introduction

Scoring students' essays is a time-consuming and cognitively demanding task for educators, particularly in higher education, where limited resources and large student cohorts are often prevalent. To alleviate these burdens, one promising approach is the adoption of artificial intelligence (AI) to automate the scoring of essays, commonly referred to as automated essay scoring (AES). AES dates back to the 1960s

HAI-Agency: Workshop on Orchestrating Human and AI Agency for Proactive and Reflective Learning, co-located with the 27th International Conference on Artificial Intelligence in Education (AIED 2026), Seoul, Korea, June 27 - July 3, 2026.
*Corresponding author.
†These authors contributed equally.
tine.keulemans@kuleuven.be (T. Keulemans); grzegorz.meller@kuleuven.be (G. Meller); elke.vandermeerschen@kuleuven.be (E. Vandermeerschen); wim.vandennoortgate@kuleuven.be (W. V. D. Noortgate); katrien.verbert@kuleuven.be (K. Verbert); fien.depaepe@kuleuven.be (F. Depaepe); annelies.raes@kuleuven.be (A. Raes); tinne.delaet@kuleuven.be (T. D. Laet)
0009-0003-2804-0901 (T. Keulemans); 0000-0003-3626-6962 (G. Meller); 0009-0007-4475-3425 (E. Vandermeerschen); 0000-0003-4011-219X (W. V. D. Noortgate); 0000-0001-5440-1318 (F. Depaepe); 0000-0001-9237-9385 (A. Raes); 0000-0003-0624-3305 (T. D. Laet)

and has subsequently primarily been based on basic statistical approaches, machine learning, and finally deep learning [5]. The rise of GenAI, especially large language models (LLMs) has boosted interest in AES by the promise of offering greater accessibility and improved scoring accuracy compared to early AES approaches [1]. Despite their potential, (Gen)AI-based essay scoring systems raise several concerns that limit their responsible adoption in higher education. AI scores can be inaccurate and biased, which may have undesired consequences for student learning and academic careers. Moreover, fully automated scoring systems may reduce teachers' sense of agency [8]. Driven by these concerns, there is a growing focus on collaboration between AES and educators. Such collaboration can be conceptualized through the lens of human-AI collaboration, as outlined by Molenaar [16], who distinguishes six levels varying in how control and responsibility are distributed between humans and AI. A key prerequisite for human-AI collaboration is explainability: educators must get insight into the reasons behind AI scoring decisions, so they can evaluate and, when necessary, override the scores [18]. This is where the field of eXplainable AI (XAI) becomes relevant. XAI seeks to provide clear and coherent explanations about how an AI-based system decides or why it provides a particular output [6]. In the context of AES, Schlippe et al. [18] found that educators have a strong desire for explainability. Furthermore, explanations are gradually enforced by regulatory frameworks, such as the right to explanation in the GDPR (Art. 22) and transparency obligations in the AI Act (Art. 13). Since recently, there is growing recognition that no "one-size-fits-all" explanation can adequately serve all stakeholders or contexts [6]. This insight has driven a recent shift from computer-centered XAI research, focusing on the technical challenges of generating explanations, to human-centered XAI research, focusing on the perceptions and needs of the users who interact with the system [11]. However, **limited human-centered XAI research** has been conducted to inform the design of explanations in GenAI-based systems, particularly around essay scoring.

To address this gap, we designed an **explainable, GenAI-powered scoring platform** that aims to support the responsible use of GenAI in essay scoring via human-AI collaboration. Its name, *the X-AIssessment Assistant*, is intentionally chosen to reflect the assisting role of AI in the assessment process, in which the technology provides supportive information while the educator retains full control [16, 17]. A design-based research (DBR) approach was adopted, as it is particularly suited to address complex, practice-oriented problems while simultaneously generating theoretical and design knowledge [15]. Guided by the DBR principles, a theory-informed prototype of the *X-AIssessment Assistant* was developed and further refined through multiple small-scale feedback cycles. This paper reports on the design of the prototype (Section 2), as well as on one design cycle involving a participatory design session with university educators (Section 3). By gathering educators' perceptions and needs regarding the platform, the session aims to inform the next design of our *X-AIssessment Assistant* and other GenAI-based essay scoring systems. The research questions are as follows: RQ1: How can an explainable GenAI-powered platform be designed to support university educators in scoring student essays, while ensuring compliance with GDPR and AI Act regulations? RQ2: How do university educators perceive an explainable GenAI-powered assessment platform?

2. The Theory-informed Design of the X-AIssessment Assistant

The prototype of our *X-AIssessment Assistant*[1] is publicly accessible via X-AIssessment - Iteration 3.0. For this prototype, we used an example drawn from a first-year bachelor-level university course in Instructional Psychology and Technology in the Education Sciences program. This course aims to provide students with foundational knowledge of key concepts and approaches related to learning and teaching. As part of the summative assessment, students are required to submit a short digital essay in which they critically reflect on a pedagogical statement (e.g., "The integration of digital technology has a positive impact on the learning outcomes of students") by presenting arguments both in favor of and against the statement. The following sections describe the platform's key features: AI model selection and prompt engineering (Section 2.1) and interface design (Section 2.2).

[1] https://tinyurl.com/4vk9sdrz

2.1. AI Model Selection and Prompt Engineering

The goal of the platform is not to develop the best performing AES algorithm, but to develop guidelines for human-centric explainability in assessment platforms. Accordingly, we employ the *off-the-shelf* generative AI model GPT-4o combined with prompt engineering. We selected this non–fine-tuned closed-source model because it can be readily deployed without additional training and generally outperforms non–fine-tuned open-source alternatives (e.g., LLaMA, Mistral) in terms of internal consistency and alignment with human raters [19]. Building on the guidelines by AlGhamdi et al. [1], we developed a prompt incorporating several established principles, illustrated in Figure 1: Expert prompting (●), explicit task specification (●), the use of a scoring rubric (●), few-shot prompting (●), retrieval augmented generation (RAG) (●), structured output template (●), end-of-prompt instruction placement (●). The colors indicate how each prompt-engineering principle is reflected in the prompt design. A separate prompt was created for each rubric criterion, and the resulting outputs were integrated, to mitigate performance degradation caused by overly long prompts [3]. For feasibility testing, a small subset of essays was used to compare GenAI assessments with human assessments leading to prompt refinements. Prior to prompting, students' informed consent was asked, and all personal data were pseudonymized (Art. 6, GDPR).

Role: "You are an educational evaluator specializing in the assessment of academic writing within the domain of Instructional Psychology and Technology. Your expertise lies in..."
Task: Your task is to assess a student's written essay exclusively on the essential criterion. To do so, you will be provided with the rubric guidelines, contextual examples with their corresponding scores, and the relevant course materials, as outlined below.
Context:
Scoring Rubric: Rubric
Examples: Examples
Course materials: {context_text}
Based on context and instructions, evaluate the following essays: {essay_text}
Output:

Criterium	Score	Explanation based on subcriteria
Argument 1 Essential	0 or 0.5	

Instruction: Indicate on which parts of the course text you based your assessment

Figure 1: Sample prompt for rubric criterion "Essential"

2.2. Interface with Explanations

The current interface design of *the X-AIssessment Assistant*, shown in Figure 2, is guided by state-of-the-art literature insights on human-centric explainability in AES, and two previous DBR-cycles.

State of the Art Literature Insights. Although the body of research on human-centric explainability in AES is limited, existing research suggests that natural language explanations, which justify scores in human-readable terms aligned with assessment criteria, are more promising than traditional explanation approaches such as feature-importance highlighting, which merely indicate the words that most influenced a model's decision [4][13]. Moreover, prior work suggests that educators find limited value in numeric confidence indicators that convey how certain the model is about its scoring decision when these are presented without any additional information [4] [12].

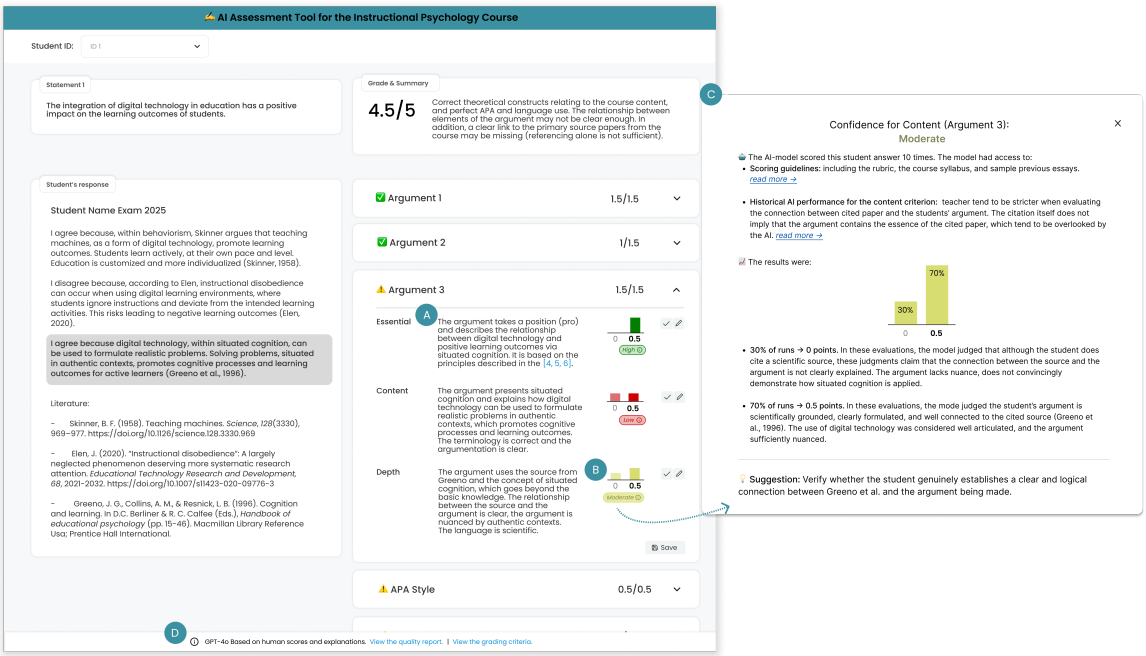

Fig. 2. Interface of the X-AIssessment Assistant

Previous DBR-cycles. Before the participatory design session with educators (see Section 3), two other participatory design sessions were conducted: one in May 2025 during an XAI in education workshop at the EATEL summer school involving eight experts in technology-enhanced learning (TEL), and one in November 2025 involving seven HCI experts (researchers within the Augment group at KU Leuven). Following an introduction to the prototype, participants were asked to identify beneficial and problematic aspects of the system individually using post-its on a digital whiteboard application (Miro). These reflections subsequently served as the basis for a group discussion, facilitating a more balanced exchange of perspectives. The analysis of the sessions led to four key design requirements (DRs), with DR1 and DR2 derived from the TEL and DR3 and DR4 from the HCI experts. *DR1 - Facilitate constructive alignment, via the ability to upload course-specific learning materials.* TEL experts expressed concern that model-generated assessments may rely on general knowledge encoded in LLMs, rather than on the specific course materials that were taught. Based on this need for constructive alignment, participants advocated for the ability to provide the LLM with course materials so that both the assessment and the explanations can be grounded in the course materials. *DR2 - Combine natural language explanations with performance metrics and confidence indicators.* Participants, that during that cycle were only presented with natural language explanations, explicitly expressed the need for other explanations, mainly those that give information about the performance of the system, and how certain the LLM is about its scoring-decision, known as performance metrics and confidence scores. *DR3 - Display confidence indicators in a visual, non-numeric way.* When presenting different confidence indicators to HCI specialist, they stressed that uncertainty should be communicated visually, avoiding numerical scores. Graphs with different color cues were preferred over other formats. *DR4 - Ensure human oversight for each scoring decision.* HCI experts also stressed that each rubric criterion should require explicit confirmation by the educator, for example via a checkbox. This can support accountability and reduce automation bias, whereby users tend to over-rely on AI outputs by accepting them uncritically.

Current Interface Design. In the current interface design, **natural-language explanations** (A) are provided for each rubric criterion, in line with the literature. It may be argued that the natural language explanations generated by our system do not constitute XAI in a strict technical sense, as they do not directly reveal the model's underlying mechanisms or decision process (e.g., definition of European Commission [6]). However, these explanations can be considered XAI from a broader user-centered perspective, as they aim to enhance the understandability of the model predictions, here the scoring decisions (e.g., definition of Mangold [14]). When applicable, RAG incorporates relevant course

materials from a curated knowledge base into the explanation (as a response to DBR 1). Embedded reference numbers make these sources transparent to educators, who can inspect them via hover-based pop-up previews. The novelty lies not in the use of RAG itself, but in exposing retrieved sources to users as an additional explainability mechanism. Furthermore, the interface presents **confidence indicators and performance metrics** (as a response to DBR 2). For each Gen-AI score the interface shows a graph, that indicates the certainty of the score (B) (as a response to DR3). By clicking on the information button below that graph, educators can get a more detailed explanation of how confidence was derived (C). This expanded view also includes GenAI-generated, actionable suggestions that indicate where human judgment is particularly important. **Performance metrics** are given in the quality report (D). By interacting with the footer, educators can access a historical overview showing how AI-generated scores aligned with human scores in previous academic years. Finally, every GenAI-score must be either accepted or edited by the educator via visible checkboxes and edit buttons next to each rubric criterion (as a response to DR4). This approach aligns with the "human oversight" requirement of the AI Act (Art. 14).

3. Participatory design session with educators

Seven teaching assistants (TAs) in educational sciences participated in a participatory design session, interacting with a prototype of the *X-AIssessment Assistant.* Although the system was not fully operational, the AI-generated scores and natural language explanations were actually AI produced (see section 2.1) to ensure the proposed interface and functionality reflect a feasible real-world AI application rather than a purely conceptual design. The study was approved by the KU Leuven's ethical committee (G-2024-8446). The same methodological approach as in previous DBR cycles was adopted. The session was video-recorded, transcribed, and pseudonymized (P1–P7). The analysis of this design cycle was informed by the **Theory of Planned Behavior (TPB)**, serving as framework to structure factors that may influence TAs' intention to engage with the *X-AIssessment Assistant.* Within TPB, a first determinant is the attitude towards the behavior, which in this context reflects how favorably a person evaluates the use of the *X-AIssessment Assistant.* A second determinant is the subjective norm that refers to the perceived social pressure to use or avoid the *X-AIssessment Assistant.* A third determinant is perceived behavioral control, which refers to educators' perceptions of how easy it is to effectively use the platform [2].

Regarding **attitude towards the behavior**, participants (3 out of 7) expressed positive attitudes towards the use of the platform, indicating that AI can save time on form-related criteria. As P1 noted, "*APA style and language, those are time savers. You can almost blindly accept it.*"[2] This preference was explicitly contrasted with content-related criteria, which these participants felt required human judgment. This distinction persisted even after participants were informed that the quality report indicated lower AI–human alignment for language than for content-related criteria. At the same time, participants (4 out of 7) raised concerns about the workload. Although each natural language explanation was concise, the combined amount of text required too much time to read. P2 said: *"In fact, it will not save time because, you need to read a lot more."* Participants agreed that they would not adopt an AI tool if it was more time-consuming than manual work. The **subjective norms** surrounding GenAI-assisted assessment appeared to restrict participants' intention to use the platform. Participants (2 out of 7) emphasized that grading is seen as a core responsibility by educators. P7 stated: *"We as teachers have not only the duty but also the right to score and to graduate students. So, it is really deeply rooted in our pedagogical task."* However, TAs expressed greater openness to use the *X-AIssessment Assistant* for formative assessment (rather than for summative assessment), related to this 5 out of 7 participants suggested including feedback. P5 said: "*AI can also help in formulating feedback, because now the practice is that the student doesn't get feedback.*" Regarding **perceived behavioral control,** all participants complimented the interface design. P3 stated: *"I really enjoyed the minimalistic interface. It was really clean and intuitive*." However, the perceptions of control were not uniformly positive. While

[2]For the purposes of this article, quotes in Dutch were translated into English by the author(s).

some participants valued the ability to adjust GenAI-generated scores and explanations, others felt that the system constrained their professional autonomy. This concern was expressed by P7, who stated: "*In some way it is changing something with us and our autonomy as teachers.*" P5 suggested: *"the sequence is very important. It's not the AI first, it's the teacher first."*

4. Discussion

The findings resonate with patterns identified in prior research, and informed the formulation of design guidelines for our platform, while pointing to avenues for future research.

A key finding emerging from the participatory sessions is that, while natural-language explanations combined with confidence indications and their breakdown were appreciated in principle. However, their cumulative volume increased the perceived workload and reduced the intention to use the platform. This mirrors findings by De Groot [4], who showed that explanations can increase cognitive load and may ultimately discourage the use of explainability features. Future iterations of the *X-AIssessment assistant* should therefore *balance explainability and cognitive load by avoiding numerous and lengthy explanations,* and user studies should explicitly study cognitive load as outcome variable. The finding that TAs have a stronger intention to use the platform for form-related criteria, than for content-related criteria aligns with Hall et al. [9] who show that instructors tend to trust AI support for plagiarism detection or word counts, while remaining more skeptical toward criteria that require interpretation. Further research should also take the role of existing mental models of the system into account, as suggested by Hoffman [10]. The finding that educators are more likely to use the *X-Assessment Assistant* when adaptable feedback is included, aligns with concerns raised by AlGhamdi et al. [1]. The authors note that although scoring and feedback are closely intertwined in educational practice, research on automated assessment systems often addresses only one of these components. Future versions of the platform will *integrate scoring and adaptable feedback within a single workflow***.** Lastly, based on the concerns around autonomy and the remark of P5, a future version can be proposed in which an *AI-suggested score is displayed alongside the teacher-assigned score, with the option to show or hide the AI score.* Similar design choices have been adopted in other XAI assessment platforms, such as Deep Dives studied by Hall et al. [9] and CheckMates studied by De Groot [4].

While small participant groups are appropriate in DBR, a limitation of this study is that only proxy users, teaching assistants from the field of educational sciences, were involved. This choice was made deliberately, since in a next design cycle we will include the didactic team of the target course. A think-aloud and semi-structured interview study will be conducted with the didactic team to unravel different (interface-related) factors influencing the perceived acceptance of the system, the perceived autonomy, and the perceived *appropriate* trust. Examining the latter is particularly important, as explanations may increase users' trust in AI systems and can inadvertently lead to overreliance on AI-generated recommendations [7].

Acknowledgments

This study was funded by Bijzonder Onderzoeksfonds (BOF)/Special Research Fund (grant number: IDN/24/004).

Declaration on Generative AI

During the preparation of this work, the author(s) used Microsoft Co-Pilot for language checking and GPT-4o for the generation of scores and explanations (via an evidence-based prompting approach).

References

[1] AlGhamdi, E., Li, Y., Gašević, D., & Chen, G. (2025). Leveraging prompt-based LLMs for automated scoring and feedback generation in higher education. Computers & Education, 105511. https://doi.org/10.1016/j.compedu.2025.105511

[2] Ajzen, I. (1991). The theory of planned behavior. Organizational behavior and human decision processes, 50(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

[3] Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. Patterns, 6(6), 101260. https://doi.org/10.1016/j.patter.2025.101260

[4] De Groot, L. (2025). Exploring the relationship between cognitive load and teacher-XAI interaction for XAI-assisted scoring thesis [Masterscriptie, Raboud universiteit Nijmegen]. Raboud educational respository.

[5] Dikli, S. (2006). An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment, 5(1). https://ejournals.bc.edu/index.php/jtla/article/view/1640

[6] European Commission: European Education and Culture Executive Agency, AI report – by the European Digital Education Hub's squad on artificial intelligence in education, Publications Office of the European Union, 2023, https://data.europa.eu/doi/10.2797/828281

[7] Ehsan, U., & Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. Patterns (New York, N.Y.), 5(6), Article 100971. https://doi.org/10.1016/j.patter.2024.100971

[8] Figueras, C., Farazouli, A., Cerratto Pargman, T., McGrath, C., & Rossitto, C. (2025). Promises and breakages of automated grading systems: a qualitative study in computer science education. Education Inquiry, 1–22. https://doi.org/10.1080/20004508.2025.2464996

[9] Hall, E., Seyam, M., & Dunlap, D. (2024). Exploring explainability and transparency in automated essay scoring systems: a user-centered evaluation. In P. Zaphiris & A. Ioannou (Eds.), Learning and Collaboration Technologies (pp. 266–282). Springer Nature. https://doi.org/10.1007/978-3-031-61691-4_18

[10] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5. https://doi.org/10.3389/fcomp.2023.109625

[11] Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. Frontiers in Artificial Intelligence, 7, 1456486. https://doi.org/10.3389/frai.2024.1456486

[12] Li, Y., Raković, M., Srivastava, N., Li, X., Guan, Q., Gašević, D., & Chen, G. (2025). Can AI support human grading? Examining machine attention and confidence in short answer scoring. Computers and Education, 228, Article 105244. https://doi.org/10.1016/j.compedu.2025.105244

[13] Li, Y., Shan, Z., Raković, M., Guan, Q., Gašević, D., & Chen, G. (2025). When AI explains in natural language: Unveiling the impact of generative AI explanations on educators' grading and feedback practices. Education and Information Technologies, 30(17), 24931–24964. https://doi.org/10.1007/s10639-025-13741-z

[14] Mangold, A., Zietz, J., Weinhold, S., & Pannasch, S. (2025). On the Design and Evaluation of Human-centered Explainable AI Systems: A Systematic Review and Taxonomy (arXiv:2510.12201). arXiv. https://doi.org/10.48550/arXiv.2510.12201

[15] McKenney, S., & Reeves, T. (2018). Conducting educational design research. Routledge.

[16] Molenaar, I. (2022). Towards hybrid human-AI learning technologies. European Journal of Education, 57(4), 632–645. https://doi.org/10.1111/ejed.12527

[17] Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or human? Evaluating student feedback perceptions in higher education. In J. A. Ruipérez Valiente, N. Rummel, I. Jivet, R. Ferreira Mello, & G. Pishtari (Eds.), Technology Enhanced Learning for Inclusive and Equitable Quality Education (Vol. 15159, pp. 284–298). Springer. https://doi.org/10.1007/978-3-031-72315-5_20

[18] Schlippe, T., Stierstorfer, Q., Koppel, M. ten, & Libbrecht, P. (2023). Explainability in Automatic Short Answer Grading. In T. Schlippe, G. N. Beligiannis, T. Wang, & E. C. K. Cheng (Eds.), Artificial Intelligence in Education Technologies: New Development and Innovative Practices (Vol. 154, pp.

69–87). Springer. https://doi.org/10.1007/978-981-19-8040-4_5

[19] Seßler, K., Fürstenberg, M., Bühler, B., & Kasneci, E. (2024). Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring (No. arXiv:2411.16337; Version 1). arXiv. https://doi.org/10.48550/arXiv.2411.16337