Book of Papers presented at the IFKAD 2026
Intelligent Knowledge for Sustainable Organizations
Budapest, Hungary 1-3 July 2026
ISBN: 978-88-96687-20-8

Explainable AI as Decision Support in Automated Assessment: An Exploratory Study

Gabriele Creta, Andrea De Mauro
LUISS Guido Carli University, Rome, Italy

Abstract

This paper investigates how rubric-based Explainable Artificial Intelligence (XAI) outputs influence instructor trust and grading behavior in AI-assisted academic assessment. As AI technologies become increasingly integrated into educational evaluation workflows, understanding the cognitive mechanisms through which instructors interact with AI-generated scores is critical for designing effective human–AI collaboration in grading and for the broader governance of knowledge assets in higher education (Khosravi et al., 2022). Two competing hypotheses motivate the study: explanations may either mitigate automation bias by enabling critical evaluation, or amplify it by providing a rationalization for uncritical acceptance of AI recommendations. We present a within-subject pilot experiment in which university-level evaluators (n = 3) assessed 12 short student responses to Computer Science questions under three conditions: (a) AI-only, where only a numerical AI score was displayed; (b) AI + Explanation, where the AI score was accompanied by a four-criterion rubric breakdown and a textual justification; and (c) Anchoring, where evaluators first graded independently and then were exposed to the AI output with an option to revise. AI grading outputs were generated using a Retrieval-Augmented Generation (RAG) pipeline calibrated against three open-access essay scoring datasets (GRE Analytical Writing, DREsS, and ASAP2) and indexed in a FAISS vector store for similarity retrieval. Preliminary results from this pilot are consistent with rubric-based explanations reducing deviation between human and AI grades (mean |delta| = 1.2 vs. 1.9 for AI-only), doubling the observed revision rate in the anchoring condition (67% vs. 33%), and shifting all observed revisions toward the AI score. Qualitative evidence indicates that evaluators cite specific AI observations when revising with explanations but resist adjustment without them. A discrepancy between self-reported and observed AI influence further suggests a metacognitive blind spot consistent with the automation-bias literature. These preliminary findings are consistent with the hypothesis that structured XAI outputs may enhance evaluator reliance on AI grading systems and reshape evaluative judgment through anchoring mechanisms, with implications for the governance of human–AI assessment workflows in higher education.

Keywords: Explainable Artificial Intelligence; Human–AI Decision-Making; Automated Assessment; Decision Support Systems; Knowledge Governance

Paper type: Academic Research Paper

1 Introduction

Assessment represents a core educational practice through which knowledge is validated, academic standards are enforced, and institutional legitimacy is maintained. Grading decisions do not merely certify learning outcomes; they shape perceptions of fairness, influence trust in educational institutions, and contribute to the governance of academic systems. As artificial intelligence technologies become increasingly integrated into educational environments, assessment has emerged as a particularly critical domain where automated tools promise efficiency gains while simultaneously raising significant ethical, organizational, and governance-related concerns (Kasneci et al., 2023).

The rapid diffusion of Large Language Models (LLMs) within educational environments has introduced new challenges for automated assessment practices, including difficulties in understanding how grading decisions are produced, the emergence of potential biases embedded in model outputs, and a reduced visibility of human judgment in evaluative processes that were traditionally grounded in expert interpretation and contextual awareness (Baker and Hawn, 2021). As a result, educators may develop reduced trust in automated grading systems or, conversely, rely on them inappropriately—either by delegating evaluative responsibility uncritically or by rejecting potentially useful tools due to a lack of interpretability. Both dynamics risk compromising the effective and responsible integration of AI-based assessment systems in educational practice.

Explainable AI (XAI) offers a potential mechanism for promoting calibrated trust rather than blind reliance. In the context of AI-assisted assessment, XAI encompasses techniques that make AI grading decisions interpretable to human evaluators, from feature importance scores to natural language explanations of the reasoning behind an AI grade. Arrieta et al. (2020) provide a comprehensive taxonomy of XAI approaches, emphasizing that the goal of explainability is not complete access to a model's internal reasoning processes, but rather the ability to produce justifications that are understandable and verifiable by domain experts. In educational settings specifically, Khosravi et al. (2022) argue that explainability is essential for maintaining

ISBN: 978-88-96687-20-8

accountability and trust when AI systems are deployed in contexts where automated decisions have direct consequences for individuals.

However, a parallel concern arises from cognitive psychology. Research on the anchoring-and-adjustment heuristic has established that individuals are susceptible to anchoring effects when making numerical judgments under uncertainty (Tversky and Kahneman, 1974). In the context of AI-assisted grading, an AI-generated score may serve as an anchor that systematically biases the instructor's final grade toward the AI suggestion, regardless of whether the AI assessment is accurate. This phenomenon is related to automation bias, where human decision-makers over-rely on automated systems without critically evaluating their outputs (Parasuraman and Riley, 1997). Explanations may either mitigate this bias by enabling critical evaluation or amplify it by providing a rationalization for uncritical acceptance of AI recommendations (Bansal et al., 2021).

This study addresses these competing hypotheses through a controlled within-subject experiment. Adopting a human-centered perspective, we conceptualize automated grading not as a replacement for educators' judgment, but as a decision-support mechanism integrated into existing assessment practices. In line with a human-centered interpretation of XAI, explainability is operationalized as the ability to justify scoring decisions through explicit, rubric-based criteria that are understandable and verifiable by educators. Specifically, we investigate two research questions:

- ***RQ1:*** *Does the presence of rubric-based XAI explanations increase evaluator reliance on AI grades compared to score-only output?*
- ***RQ2:*** *Do XAI explanations amplify or mitigate anchoring effects when evaluators are exposed to AI grades after forming their own initial assessment?*

2 Theoretical Background

2.1 Automated assessment and AI in education

Automated essay scoring (AES) systems have evolved significantly since their introduction in the 1960s. Contemporary systems leverage deep learning architectures and large language models to evaluate student writing across multiple dimensions including content accuracy, argumentation quality, and language proficiency (Ke and Ng, 2019). Despite improvements in accuracy, adoption in summative assessment contexts remains limited, largely due to instructor skepticism about the ability of AI to capture the nuanced qualitative judgments that characterize expert evaluation (Hockly, 2019). Recent advances in generative AI have introduced new capabilities for rubric-based grading that go beyond holistic scoring: LLMs can generate structured evaluations aligned with predefined criteria, producing not only scores but also detailed justifications that approximate the format of human feedback (Mizumoto and Eguchi, 2023). Yoo et al. (2025) demonstrate the viability of rubric-based essay evaluation using structured datasets such as DREsS, confirming that scoring rubrics can serve as structured knowledge representations that formalize evaluative criteria and make assessment logic explicit.

2.2 Explainable AI, trust, and automation bias

The field of XAI has emerged in response to concerns about the opacity of machine learning models and the need for transparency in high-stakes decision-making contexts. Arrieta et al. (2020) provide a foundational survey of XAI concepts and taxonomies, categorizing approaches by scope, methodology, and explanation type. In educational contexts specifically, Khosravi et al. (2022) identify explainability as a critical enabler for the responsible adoption of AI, arguing that without interpretable outputs, AI-driven educational systems risk undermining both pedagogical objectives and institutional trust.

Trust calibration—the alignment between a user's trust in an AI system and the system's actual reliability—is a central goal of XAI design (Lee and See, 2004). Miscalibrated trust manifests as either under-reliance (rejecting accurate AI outputs) or over-reliance (accepting inaccurate outputs without scrutiny). Explanations have been theorized to promote appropriate reliance by enabling users to evaluate the quality of AI reasoning on a case-by-case basis (Bansal et al., 2021). However, empirical evidence remains mixed: Vasconcelos et al. (2023) find that while explanations can reduce overreliance in some decision-making contexts, they may also increase automation bias by providing a rationalization for uncritical acceptance. Parasuraman and Manzey (2010) provide a comprehensive integration of the automation bias literature, demonstrating that the gap between perceived and actual reliance on automated systems is a robust and persistent phenomenon across professional domains.

ISBN: 978-88-96687-20-8

2.3 Anchoring effects in evaluative judgment

The anchoring-and-adjustment heuristic, first described by Tversky and Kahneman (1974), refers to the tendency for individuals to insufficiently adjust their estimates from an initial reference value. In assessment contexts, exposure to prior evaluations or suggested scores has been shown to systematically influence subsequent judgments (Mussweiler and Strack, 1999). The introduction of AI-generated grades creates a novel anchoring scenario in which the anchor is produced by an automated system rather than a human peer, raising questions about whether the source of the anchor modulates its influence on evaluator behavior. Few studies have examined anchoring in the specific context of AI-assisted educational assessment, and fewer still have investigated how the presence or absence of AI explanations moderates anchoring strength. This study addresses this gap directly.

3 Research Design and Methodology

3.1 Experimental design

We employed a within-subject experimental design in which each participant evaluated 12 short student responses under three conditions. The within-subject approach controls for individual differences in grading severity, as each participant serves as their own control. The three conditions were: (a) AI-first AI-only (4 tasks), where participants viewed the student response alongside an AI-generated total score with no further explanation; (b) AI-first AI + Explanation (4 tasks), where participants viewed the response alongside the AI score, a criterion-level breakdown, and a textual explanation highlighting key strengths and weaknesses; and (c) Anchoring (4 tasks, 2 with score only and 2 with full explanation), where participants first assigned an initial grade without AI input, then on a subsequent page were shown the AI output and given the option to revise with a brief justification.

3.2 Stimuli: student responses

Twelve short written responses (50–150 words) to fundamental Computer Science questions were collected from undergraduate students in the final year of a Bachelor's degree program. Topics covered core CS concepts including data structures, algorithm complexity, databases, programming paradigms, version control, machine learning, neural networks, and data preprocessing. Responses were reviewed for authenticity; five responses exhibiting characteristics of AI generation (e.g., formulaic analogies, perfectly balanced paragraph structures) were rewritten to match natural student writing style, including realistic imperfections such as minor spelling errors, informal phrasing, and uneven argumentation depth.

3.3 AI grading pipeline

AI grades were generated using a Retrieval-Augmented Generation (RAG) pipeline designed to produce calibrated, rubric-aligned assessment outputs. Three open-access essay scoring datasets were ingested as calibration sources: the GRE Analytical Writing scoring guide with sample essays and reader commentaries at each level (0–6); the DREsS dataset comprising approximately 29,000 student essays scored on three analytic dimensions (Yoo et al., 2025); and the ASAP2 dataset containing approximately 24,000 student essays with holistic scores (1–6). Source texts were segmented into overlapping chunks, embedded using a sentence-level transformer model (all-MiniLM-L6-v2), and indexed in a FAISS vector store for cosine-similarity retrieval. For each student response, the system retrieved the most semantically relevant reference excerpts and injected them into the grading prompt. The Large Language Model used for grading was Claude Opus 4 (Anthropic), operating at low temperature (0.1) to ensure deterministic outputs. All outputs were frozen prior to participant exposure.

3.4 Rubric design

Table 1. Analytic rubric structure

Criterion	Scale	Description
Completeness / Coverage	0–8	Addresses all parts of the question; covers key concepts
Correctness / Conceptual Accuracy	0–8	Factual accuracy; appropriate use of terminology
Reasoning & Structure	0–7	Logical coherence; argument development; organization
Clarity of Language	0–7	Clarity; grammar; readability; conciseness
Total	0–30	Sum of all four criteria

ISBN: 978-88-96687-20-8

Scoring rubrics were treated as structured knowledge representations that formalize evaluative criteria and make assessment logic explicit. Prior research has shown that rubric-based assessment supports transparency, consistency, and pedagogical effectiveness in grading practices (Yoo et al., 2025). The four-criterion rubric was designed to evaluate both content mastery (Completeness, Correctness) and communicative competence (Reasoning, Language), reflecting the multi-dimensional nature of academic assessment.

3.5 Participants and procedure

Three university-level evaluators participated in the study: two Researchers/Assistant Professors specializing in Artificial Intelligence (10 and 5 years of experience respectively) and one Associate Professor in Data Science (2 years of experience). All held teaching and assessment responsibilities at the university level and were familiar with the Computer Science topics covered. The experiment was administered through a Qualtrics online survey. After providing demographic information, participants completed the 12 evaluation tasks in randomized order. Backward navigation was disabled to prevent modification of initial grades after AI exposure. The survey concluded with a seven-item trust and perception questionnaire (7-point Likert scale). Completion time ranged from 9 to 25 minutes (mean: 17 minutes).

4 Results

4.1 AI-first conditions: effect of explanations on reliance

Table 2. Human grades versus AI scores in AI-first conditions

Condition	Topic	AI	P1	P2	P3	M	Δ
AI-only	Train/Val/Test	25	24	25	22	23.7	−1.3
AI-only	Struct. vs Unstruct.	28	28	27	29	28.0	0.0
AI-only	Relational vs NoSQL	24	25	25	23	24.3	+0.3
AI-only	API	17	25	22	18	21.7	+4.7
AI+Expl	Compiled vs Interp.	26	26	25	24	25.0	−1.0
AI+Expl	Version Control	24	26	24	24	24.7	+0.7
AI+Expl	Overfitting	28	30	27	27	28.0	0.0
AI+Expl	Data Normalization	24	28	24	25	25.7	+1.7

The mean absolute deviation from the AI score was 1.92 points in the AI-only condition and 1.17 points in the AI + Explanation condition, a pattern consistent with rubric-based explanations being associated with greater evaluator reliance on AI grades in this pilot sample. The effect is particularly visible in cases where the AI and human assessments diverge substantially: for the API question (AI score: 17), all three evaluators assigned markedly higher grades in both conditions (18–25), but the deviation was larger in the AI-only condition (mean delta: +4.7) where evaluators lacked access to the AI's reasoning about the response's circularity and superficiality.

4.2 Anchoring analysis: grade revision after AI exposure

Table 3. Anchoring effect: initial grades, AI scores, and revisions

Cond.	Topic	AI	P	Init.	Rev.	Adj.	Dir.
Only	Big O	20	P1	24	23	−1	Toward AI
Only	Big O	20	P2	24	—	0	—
Only	Big O	20	P3	16	18	+2	Toward AI
Only	Neural Net.	22	P1	27	—	0	—
Only	Neural Net.	22	P2	10	—	0	—
Only	Neural Net.	22	P3	22	—	0	—
Expl	API	17	P1	26	24	−2	Toward AI
Expl	API	17	P2	22	20	−2	Toward AI
Expl	API	17	P3	22	21	−1	Toward AI
Expl	Sup./Unsup.	28	P1	30	29	−1	Toward AI
Expl	Sup./Unsup.	28	P2	27	—	0	—
Expl	Sup./Unsup.	28	P3	29	—	0	—

In the AI-only anchoring condition, 2 out of 6 evaluator-task pairs (33%) resulted in grade revision. In the AI + Explanation condition, 4 out of 6 (67%) resulted in revision. All six observed revisions moved in the direction

ISBN: 978-88-96687-20-8

of the AI score, confirming the presence of an anchoring effect consistent with the predictions of Tversky and Kahneman (1974). The mean adjustment magnitude was 1.5 points for AI + Explanation revisions compared to 0.5 for AI-only, suggesting that explanations not only increase the likelihood of revision but also produce larger adjustments toward the AI anchor.

4.3 Qualitative evidence: the role of explanations in revision reasoning

Analysis of open-ended revision justifications reveals a qualitative difference in reasoning between conditions. In the AI + Explanation condition, evaluators referenced specific observations from the AI explanation to justify their revisions. Participant 1 revised the API question from 26 to 24, stating that the AI "made me think about the shallowness and the lack of key aspects." Participant 2 revised the same question from 22 to 20, noting that "the explanation suggested additional limitations, to which I agree." For the supervised/unsupervised learning question, Participant 1 revised from 30 to 29, acknowledging that "the Netflix example is hybrid, not purely unsupervised."

In the absence of explanations, evaluators expressed explicit resistance to adjustment. Participant 1 stated: "Without explanation I just stick to my grade," providing direct evidence that rubric-based explanations serve as a cognitive mechanism enabling evaluators to identify specific grounds for revision, whereas a numerical score alone is insufficient to overcome confirmation of one's own judgment. This finding is consistent with Bansal et al.'s (2021) observation that explanations enable complementary team performance by providing decision-makers with information they can independently verify.

4.4 Trust and perception questionnaire

Table 4. Trust and perception questionnaire (1–7 Likert scale)

Statement	P1	P2	P3	M
AI grade was accurate and reliable	5	5	5	5.0
Confident in my final grade	6	6	6	6.0
AI influenced my evaluation	5	3	2	3.3
AI made evaluation faster	6	5	5	5.3
Understood AI grading logic	4	7	5	5.3
AI makes evaluation fairer	7	7	4	6.0
Would use AI grading regularly	7	7	3	5.7

The most notable finding is the discrepancy between self-reported and observed AI influence. Evaluators reported that the AI had limited influence on their evaluation (mean: 3.3/7, below the scale midpoint), yet the behavioral data show that 50% of all anchoring task pairs resulted in grade revision, with 100% of revisions moving toward the AI score. This pattern is consistent with the automation bias literature documenting a gap between perceived and actual reliance on automated systems (Parasuraman and Manzey, 2010). At the same time, evaluators expressed high confidence in their final grades (6.0/7) and strong agreement that AI makes evaluation fairer (6.0/7), suggesting that they value AI as a decision-support tool even while underestimating its influence on their own judgment.

5 Discussion

The results of this exploratory study provide preliminary evidence addressing both research questions and contribute to the emerging literature on human–AI collaboration in educational assessment.

Regarding RQ1, the AI + Explanation condition produced closer alignment between human and AI grades than the AI-only condition (mean |delta| = 1.17 vs. 1.92), a pattern consistent with rubric-based XAI outputs being associated with greater evaluator reliance on AI assessments. This finding is consistent with the trust calibration framework proposed by Lee and See (2004), which predicts that transparency in AI decision-making facilitates appropriate reliance. It also aligns with Khosravi et al.'s (2022) argument that explainability is essential for productive human–AI collaboration in educational contexts: structured explanations appear to provide the interpretive scaffolding that enables evaluators to engage meaningfully with AI outputs rather than simply accepting or rejecting them.

Regarding RQ2, the anchoring analysis suggests that, in this pilot sample, explanations may amplify rather than mitigate anchoring effects. Evaluators were twice as likely to revise their initial grade when presented with an explanation compared to a score alone, and all revisions moved toward the AI score. This finding challenges the optimistic view that XAI necessarily promotes critical evaluation. Instead, structured explanations may function as persuasive arguments that facilitate acceptance of the AI perspective, particularly when the

ISBN: 978-88-96687-20-8

explanation identifies specific weaknesses that the evaluator had not independently considered. This interpretation is consistent with Vasconcelos et al.'s (2023) finding that explanations can increase overreliance in certain decision contexts.

The qualitative data provide a mechanistic account: when explanations are present, evaluators engage in a form of deliberative anchoring, evaluating the plausibility of specific AI observations and adjusting when they find the reasoning convincing. When explanations are absent, evaluators lack a cognitive basis for adjustment and default to their initial judgment. The trust questionnaire reveals an important metacognitive blind spot: evaluators perceive AI as accurate and fair but underestimate its influence, suggesting that AI-assisted grading may introduce systematic biases that evaluators are not consciously aware of. This has significant implications for the governance of AI-enabled assessment systems, aligning with broader concerns about accountability and transparency in algorithmic decision-making (Baker and Hawn, 2021).

5.1 Implications for practice

These findings carry practical implications for the design of AI-assisted assessment systems. First, structured explanations should be considered a design requirement: they meaningfully increase evaluator engagement with AI outputs and enable more informed grading decisions. Second, institutions deploying AI grading tools should implement awareness training to help evaluators recognize anchoring effects and automation bias. Third, assessment workflows should consider the sequencing of human and AI evaluation carefully, as the order of exposure fundamentally shapes the interaction between human judgment and AI suggestion. Finally, rubric-based explanations may serve a dual function as both a transparency mechanism and a calibration tool, helping to align grading standards across multiple evaluators.

6 Future Research Directions

As an exploratory pilot, this study opens several promising avenues for further investigation. The sample size (n = 3) is consistent with the proof-of-concept nature of this work: the within-subject design provides multiple observations per participant and supports the identification of behavioral patterns, while inferential statistical analysis is reserved for a planned full-scale extension with a substantially larger and more diverse evaluator pool. A second extension concerns disciplinary scope: all participants specialized in AI or Data Science, and replication with evaluators from non-technical disciplines (e.g., humanities, social sciences) will help establish whether the observed patterns generalize across academic fields. Third, the present pilot focused on short-answer responses in Computer Science; subsequent work will examine longer-form assessments and a wider range of disciplines, where rubric structure and evaluative complexity differ substantially. Fourth, the AI pipeline employed a single model and configuration, and future iterations will benchmark alternative LLMs and retrieval strategies to assess the robustness of the observed effects. Finally, while this pilot intentionally compared AI-only and AI + Explanation conditions, a follow-up design will incorporate a pure human-only baseline to quantify the marginal contribution of AI assistance to overall assessment quality. Taken together, these directions define a structured research agenda that builds directly on the present findings.

7 Conclusion

This study provides preliminary evidence that rubric-based Explainable AI outputs may meaningfully shape evaluator behavior in AI-assisted academic grading. In this pilot sample, the presence of structured explanations was associated with reduced deviation between human and AI grades, a higher likelihood of grade revision after AI exposure, and revisions that consistently moved toward the AI score. Qualitative evidence further suggests that explanations provide the cognitive scaffolding that enables evaluators to engage critically with AI reasoning, while the absence of explanations leads to resistance to adjustment. These findings contribute to the broader discourse on human–AI collaboration and knowledge governance in educational organizations, supporting the view that explainability is not an intrinsic property of the model alone, but rather an emergent characteristic shaped by the provision of contextual information, evaluative guidance, and structured justification. As AI tools become increasingly prevalent in educational assessment, designing systems that promote calibrated trust rather than uncritical adoption will be essential for preserving the integrity and fairness of academic evaluation.

ISBN: 978-88-96687-20-8

8 Ethics declaration

This research involved voluntary participation in an anonymous online survey. No personal identifying information was collected beyond professional role and years of experience. Participation was voluntary and participants could withdraw at any time.

9 AI declaration

AI tools (Claude, Anthropic) were used in the following capacities: (1) as the grading engine within the RAG pipeline to generate AI assessment outputs used as experimental stimuli; (2) as an editorial aid for manuscript preparation. All substantive research design, data collection, analysis, and interpretation were conducted by the authors.

References

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82–115.

Baker, R.S. and Hawn, A. (2021) 'Algorithmic bias in education', *International Journal of Artificial Intelligence in Education*, 32, pp. 1052–1092.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T. and Weld, D.S. (2021) 'Does the whole exceed its parts? The effect of AI explanations on complementary team performance', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16.

Hockly, N. (2019) 'Automated writing evaluation', *ELT Journal*, 73(1), pp. 82–88.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023) 'ChatGPT for good? On opportunities and challenges of large language models for education', *Learning and Individual Differences*, 103, 102274.

Ke, Z. and Ng, V. (2019) 'Automated essay scoring: A survey of the state of the art', *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 6300–6308.

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S. and Gašević, D. (2022) 'Explainable artificial intelligence in education', *Computers & Education: Artificial Intelligence*, 3, 100074.

Lee, J.D. and See, K.A. (2004) 'Trust in automation: Designing for appropriate reliance', *Human Factors*, 46(1), pp. 50–80.

Mizumoto, A. and Eguchi, M. (2023) 'Exploring the potential of using an AI language model for automated essay scoring', *Research Methods in Applied Linguistics*, 2(2), 100050.

Mussweiler, T. and Strack, F. (1999) 'Hypothesis-consistent testing and semantic priming in the anchoring paradigm', *Journal of Experimental Social Psychology*, 35(2), pp. 136–164.

Parasuraman, R. and Manzey, D.H. (2010) 'Complacency and bias in human use of automation: An attentional integration', *Human Factors*, 52(3), pp. 381–410.

Parasuraman, R. and Riley, V. (1997) 'Humans and automation: Use, misuse, disuse, abuse', *Human Factors*, 39(2), pp. 230–253.

Tversky, A. and Kahneman, D. (1974) 'Judgment under uncertainty: Heuristics and biases', *Science*, 185(4157), pp. 1124–1131.

Vasconcelos, H., Joshi, M., Bansal, G., Kamar, E. and Weld, D.S. (2023) 'Explanations can reduce overreliance on AI systems during decision-making', *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1), pp. 1–45.

Yoo, H., Han, J., Ahn, S.-Y. and Oh, A. (2025) 'DREsS: Dataset for rubric-based essay scoring on EFL writing', *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13439–13454.

Appendix A. Example of AI grading output (AI + Explanation condition)

The following excerpt illustrates the structure of the AI grading output presented to evaluators in the AI + Explanation condition. It comprises (i) the question prompt, (ii) the verbatim student response, (iii) the AI-suggested score with rubric breakdown, and (iv) the textual explanation generated by the RAG pipeline.

Question (Q25). Explain what data normalization means in the context of data preprocessing. Why is it important before applying machine learning algorithms?

Student response. "Data normalization, used for preprocessing, is a way to convert numerical value scale of a dataset into a standard interval typically between 0 and 1. It ensures that all the features contribute equally to the model and that very large nominal value doesn't influence improperly the training phase. It is used to accelerate the training because algorithms can find optimal solution faster."

AI suggested score: 24/30. Completeness: 6/8 | Correctness: 7/8 | Reasoning: 5/7 | Language: 6/7.

Explanation. The answer covers the core definition (scaling to a standard interval) and two important reasons for normalization (equal feature contribution, faster training convergence), addressing the main parts of the question. Content is accurate, though "nominal value" appears to be a misuse — the student likely means "numerical value" or is confusing "nominal" (categorical) with "large-magnitude." The 0-to-1 range is correct

ISBN: 978-88-96687-20-8

for min-max normalization but the answer doesn't distinguish it from standardization (z-score), which would strengthen completeness. The answer is concise and functional but could develop the reasoning further — for instance, explaining which algorithms are particularly sensitive to feature scaling (e.g., KNN, gradient descent-based methods) versus which are not (e.g., tree-based methods).