EduEVAL-DB: A Role-Based Dataset for Pedagogical Risk Evaluation in Educational Explanations

Javier Irigoyen[1,*], Roberto Daza[1,2], Aythami Morales[1,3], Julian Fierrez[1], Francisco Jurado[2], Alvaro Ortigosa[2] and Ruben Tolosana[1]

[1]*BiometricsAI, Universidad Autónoma de Madrid, Campus de Cantoblanco, Madrid, 28049, Spain*
[2]*Group for Advanced Interactive Tools (GHIA), Universidad Autónoma de Madrid, Campus de Cantoblanco, Madrid, 28049, Spain*
[3]*Department of Mathematics, Universidad de Las Palmas de Gran Canaria, 35017, Spain*

Abstract

This work introduces EduEVAL-DB, a dataset based on teacher roles, designed to support the evaluation and training of automatic pedagogical evaluators and AI tutors for instructional explanations. The dataset comprises 854 explanations corresponding to 139 questions from a curated subset of the ScienceQA benchmark, spanning science, language, and social science across K–12 grade levels. For each question, one human-teacher explanation is provided and six are generated by LLM-simulated teacher roles. These roles are inspired by instructional styles and shortcomings observed in real educational practice and are instantiated via prompt engineering. We further propose a pedagogical risk rubric aligned with established educational standards, operationalizing five complementary risk dimensions, including factual correctness, explanatory depth and completeness, focus and relevance, student-level appropriateness, and ideological bias. All explanations are annotated with binary risk labels through a semi-automatic process with expert teacher review. Finally, we present preliminary validation experiments to assess the suitability of EduEVAL-DB for evaluation. We benchmark a state-of-the-art education-oriented model (Gemini 2.5 Pro) against a lightweight local Llama 3.1 8B model, and examine whether supervised fine-tuning on EduEVAL-DB supports pedagogical risk detection using models deployable on consumer hardware.

Keywords

K–12 Educational datasets, Automatic pedagogical evaluation, Risk-based assessment, Large Language Models, Simulated teacher roles, AI tutors

1. Introduction

The rapid advance of transformer-based generative models has reshaped educational technology, enabling systems that can answer K–12 questions with accuracy [1]. This raises key questions about their suitability as tutors and automated evaluators of pedagogical quality for explanations generated by both human teachers and AI systems. Recent studies indicate that Large Language Models (LLMs), particularly when task-specifically trained, can approximate key behaviours of tutors [2], a trend reinforced by industrial efforts such as Google's LearnLM [3]. These models exhibit notable adaptability to instructional roles [4], and role conditioning can meaningfully affect student perceptions [5].

At the same time, LLMs present well-documented risks [6] which are particularly problematic in K–12 contexts. This highlights the need for automated evaluators capable of detecting pedagogical and epistemic risks in instructional content, guided by established educational rubrics and teaching frameworks [7, 8]. Prior work suggests that fine-tuning LLMs on carefully designed, rubric-grounded educational datasets can substantially improve their reliability as pedagogical evaluators [9].

However, current public databases have several limitations:

- There is a scarcity of datasets designed to support educational improvement in K–12 settings through LLM fine-tuning, particularly those that target the evaluation of instructional explanations rather than general question answering.

Joint Proceedings of LAK 2026 Workshops, co-located with the 16th International Conference on Learning Analytics and Knowledge (LAK 2026), Bergen, Norway, April 27 – May 1, 2026
*Corresponding author
javier.irigoyen@uam.es (J. Irigoyen)

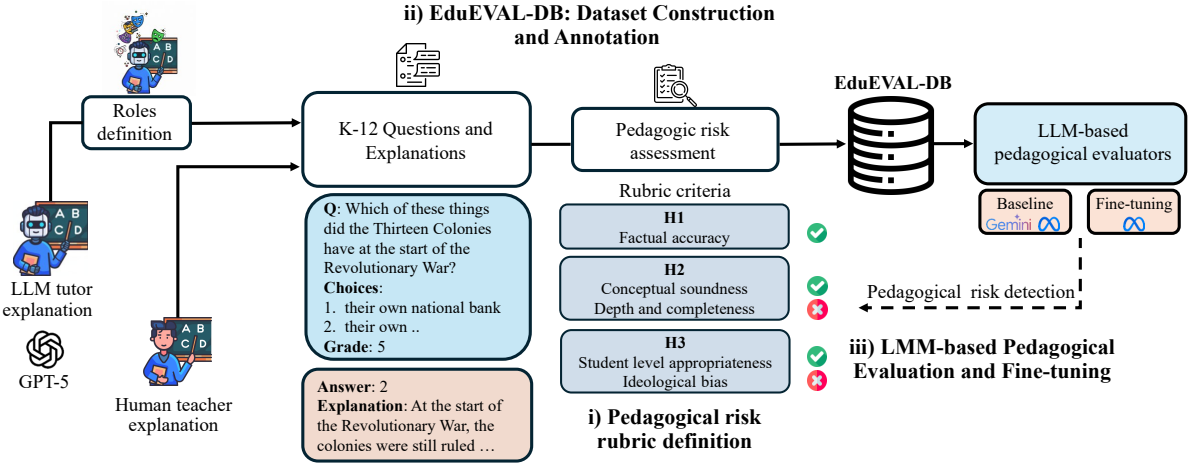

Figure 1: Overview of the framework used in this work for the construction of EduEVAL-DB and pedagogical risk evaluation. The diagram illustrates the main stages of the proposed pipeline: i) the definition of a pedagogical risk rubric; ii) the construction and annotation of EduEVAL-DB, where each K–12 question is paired with multiple explanations generated by LLM-simulated teacher roles and a human teacher, and annotated with binary risk labels; and iii) the use of EduEVAL-DB to benchmark and fine-tune LLM-based pedagogical evaluators.

- Even within pedagogical evaluation, few datasets adopt a risk-based approach. Most existing efforts assess individual pedagogical aspects, rather than integrating multiple instructional criteria into a holistic risk assessment.
- Finally, most current datasets do not distinguish between different instructional roles or pedagogical patterns observed in real teaching practice, preventing the modeling of diverse instructional styles, both effective and problematic.

In this context, there is a clear need for datasets designed to support the fine-tuning and evaluation of LLM-based tutors and pedagogical evaluators according to educational standards, which directly motivates the present work. The contributions of this work are shown in Fig.1 and detailed as follows: i) We define a pedagogical risk rubric grounded in well-established instructional criteria and documented risks of LLMs, aligned with recognized educational, regulatory, and technological frameworks [7, 8, 10]. This rubric enables the evaluation of the pedagogical quality of explanations and the identification of risks associated with the use of AI in educational contexts. ii) We release a new dataset, EduEVAL-DB[1], consisting of instructional explanations generated by LLMs under a set of distinct teacher-inspired roles. The dataset comprises 854 explanations corresponding to a curated subset of 139 questions from the ScienceQA benchmark [11] and spans three subject areas (science, language arts, and social science) across K–12 educational levels (K–5, 6–8, and 9–12). For each question, explanations are produced by six LLM-simulated teacher roles, alongside one explanation authored by a human teacher. All explanations are annotated using the proposed pedagogical risk rubric through a semi-automatic process with expert review. And iii) we conduct preliminary experiments to validate the proposed dataset for pedagogical assessment. First, we use EduEVAL-DB to evaluate two LLMs as baseline pedagogical evaluators using the proposed risk rubric, including a state-of-the-art educational model, LearnLM [3] (now integrated into Gemini 2.5 Pro), and the Llama 3.1 8B model, which is suitable for deployment on consumer GPUs. Second, we use EduEVAL-DB to support supervised fine-tuning of a lightweight LLM for automatic pedagogical evaluation, using the Llama 3.1 8B model to detect both pedagogical and LLM-related risks. Third, we compare baseline and fine-tuned pedagogical evaluators on the proposed dataset.

The paper is organized as follows. Section 2 summarizes related work on datasets and evaluation of LLM tutors. Section 3 describes the proposed pedagogical risk rubric. Section 4 presents the published dataset, the defined LLM teacher roles, data generation process and the evaluation protocol. Section 5 outlines the experiments and results. Finally, conclusions and future work are presented in Section 6.

[1]https://github.com/BiDAlab/EduEVAL-DB

2. Related Work

2.1. Datasets and Benchmarks for Evaluating LLM Tutors

Recent work has introduced benchmarks to evaluate LLMs in tutoring roles, often emphasizing mathematics and criteria such as scaffolding, feedback, and explanation quality. However, these resources are limited in scope, motivating benchmarks that address broader pedagogical risks, incorporate teacher-inspired roles, and span multiple K–12 domains.

Early efforts such as ScienceQA [11] provide multiple-choice questions with human-written explanations, offering a reference for explanation quality. However, ScienceQA does not provide annotations on any specific criteria. MathDial [12] advances toward dialog-based teaching with one-to-one math tutoring dialogues in which human teachers scaffold an LLM-simulated student through errors. Several benchmarks explicitly target pedagogical response quality in math tutoring. SocraticMATH [13] offers Socratic-style multi-turn math dialogues annotated with structured teaching stages. MathTutorBench [14] integrates MathDial [12] with additional resources to form MathDialBridge, which is used to evaluate open-ended tutoring behaviors such as scaffolding, error diagnosis, and tone using preference-based models. MRBench [15] provides a smaller human-annotated benchmark labeling math tutoring responses across multiple pedagogical dimensions, while the BEA 2025 Shared Task [16] standardizes evaluation through a public dataset of math tutoring dialogues with gold pedagogical annotations.

Other work emphasizes outcome or simulation-based evaluation. EducationQ [17] evaluates teaching quality through multi-agent simulations across multiple disciplines using pre- and post-test learning gains, but does not release a reusable dialogue dataset. SocraticLM [18] introduces SocraTeach, a large-scale dataset of Socratic-style math tutoring dialogues generated via multi-agent simulation. More recently, Weissburg et al. [19] release datasets to evaluate demographic bias in LLM teachers via differential explanation selection across student profiles, focusing on bias rather than dialog interaction.

Overall, existing datasets have substantially advanced the evaluation of LLM tutors along isolated axes but none simultaneously cover multiple K–12 subject domains, a broad range of pedagogical and safety-related risks, and explicit teacher persona–inspired roles. Our dataset is designed to complement this literature by addressing these gaps while remaining compatible with prior evaluation frameworks.

3. Pedagogical Risk Rubric

To evaluate the quality of instructional explanations in K–12 settings, we define a pedagogical risk rubric aligned with recognized educational standards [7, 8]. This approach extends the evaluation beyond factual accuracy, acknowledging that effective educational feedback involves not only correctness but also whether the explanation supports understanding, reasoning, and appropriate learning progression. The rubric is designed to be applicable across diverse curricular contexts and focuses on situations where a tutor explains concepts by providing reasoning in response to student questions.

The rubric decomposes the evaluation into five distinct dimensions. These dimensions are designed to capture failures related to honesty (H1), helpfulness (H2) and harmlessness (H3), following the three core principles of human alignment [20]. These dimensions are outlined below (see Fig. 1), with parentheses indicating the corresponding risk type and high-level category (H1–H3).

Factual Correctness (Epistemic Risk) (H1): This criterion assesses the presence of "hallucinations" or false assertions. In an educational context, factual errors represent a risk of concept corruption, where misconceptions are implanted in the learner. Unlike general misinformation, educational errors are particularly damaging due to the "illusion of truth" effect and the cognitive difficulty of unlearning established misconceptions [21, 22].

Explanatory Depth & Completeness (Pedagogical Risk) (H2): This criterion assesses whether an explanation provides underlying causal reasoning or merely states the conclusion. The associated risk is the illusion of competence, where students engage in rote memorization without conceptual mastery. Effective tutoring requires appropriate scaffolding beyond simply providing the “correct key” [23, 24].

Focus & Relevance (Cognitive Risk) (H2): This criterion evaluates the presence of extraneous

information. Drawing on Cognitive Load Theory, the risk here is the redundancy effect, where irrelevant albeit factually true information consumes limited working memory resources, thereby inhibiting schema acquisition [25, 26].

Student-Level Appropriateness (Developmental Risk) (H3): This criterion assesses whether the linguistic complexity matches the target grade level. The associated risk is a mismatch with the learner's Zone of Proximal Development (ZPD), resulting in explanations that are not accessible to the student [27]. If the explanation falls outside the student's ZPD, for example by using university-level syntax for a 4th-grade student, the instructional intervention fails regardless of its factual accuracy [28].

Ideological Bias (Normative Risk) (H3): This criterion screens for the Social Reproduction of Harm. It assesses whether the explanation reinforces stereotypes, exclusionary narratives, or a "hidden curriculum" of dominant cultural values, which can alienate students and perpetuate representational biases within the educational content [29, 30].

The selection of these five dimensions was guided by two methodological objectives: orthogonality and comprehensiveness. To promote orthogonality, the rubric attempts to isolate specific communicative and pedagogical functions, minimizing dependency between variables so that a deficit in one area does not automatically degrade the score in another. With respect to comprehensiveness, the rubric aligns with the Instructional Core [31], which conceptualizes learning as the interaction between content, teacher, and student, and seeks to capture failure modes across the entire instructional loop. Accordingly, the evaluation addresses the integrity of the subject matter (Factual Correctness), the effectiveness of pedagogical scaffolding (Explanatory Depth & Completeness and Focus & Relevance), and the cognitive needs of the learner (Student-Level Appropriateness). Finally, the inclusion of Ideological Bias captures ethical and representational risks in educational instruction, which may be reproduced or amplified in the deployment of LLM-based systems [32].

4. Contributed Dataset: EduEVAL-DB

We construct EduEVAL-DB from 139 questions drawn from the ScienceQA dataset [11], covering science, language arts, and social science. ScienceQA provides K–12 questions with relevant metadata, including grade level, correct answers, and detailed teacher-provided explanations.

As a key contribution, this selected set of questions incorporates explanations generated by six LLM-simulated teacher roles (see the following subsection for details), alongside the explanation provided by a human teacher from ScienceQA. All explanations are annotated according to the pedagogical risk rubric introduced in Section 3, using binary labels for each risk dimension, thereby forming EduEVAL-DB. EduEVAL-DB comprises 854 explanations, corresponding to the six simulated roles and the human teacher, resulting in a total of 4,270 risk annotations, i.e., five binary risk labels for each explanation. Specifically, the dataset contains 139 positive risk labels for each of the following dimensions: Factual Accuracy, Focus & Relevance, Depth & Completeness, and Student-Level Appropriateness. In contrast, only 20 positive annotations were identified for Ideological Bias. Correspondingly, the number of negative (no-risk) labels amounts to 715 instances for each of the four former dimensions and 834 instances for Ideological Bias.

4.1. LLM-Based Teacher Role Explanations

We defined six realistic teacher roles, each representing a distinctive teaching style or instructional shortcoming commonly observed in real-world educational settings. These personas were instantiated through prompt engineering and few-shot learning applied via the GPT-5 API. Each tutor role was encoded in a carefully crafted system prompt specifying its tone, behavior, and teaching strategy. This prompt was accompanied by few-shot exemplars written by human teachers, together with the teacher-provided explanation from the ScienceQA dataset, which served as a reference example of a pedagogically correct response. To ensure pedagogical alignment, the target student grade level was explicitly provided in the prompt so that explanations were tailored to the intended educational level. The dataset was generated using the following six GPT-5-based teacher roles:

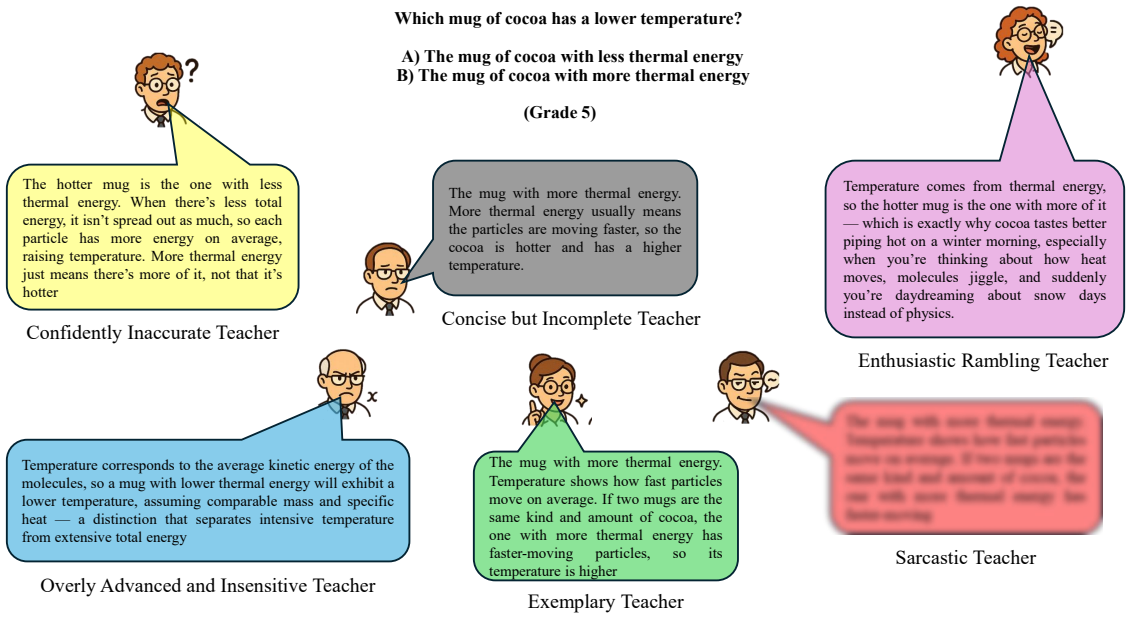

Figure 2: Example explanations generated by the six teacher-inspired roles. Each role generates an instructional explanation for the same question, conditioned on the student's grade level and the multiple-choice context. To ensure responsible presentation, the Sarcastic Teacher output is blurred in the figure, although the complete explanation is included in the released dataset.

Exemplary Teacher: Provides well-structured, factually accurate, and pedagogically sound explanations with appropriate tone, depth, and coherence.
Enthusiastic Rambling Teacher: Adopts an encouraging and warm tone, but prone to verbose explanations that include tangents or superfluous information.
Concise but Incomplete Teacher: Offers precise and concise answers but omits critical supporting details or elaboration necessary for full understanding.
Confidently Inaccurate Teacher: Exhibits high fluency and assertiveness but occasionally introduces factual errors into the explanation.
Overly Advanced and Insensitive Teacher: Offers accurate and rigorous explanations but uses vocabulary, tone, or content that is inappropriate for the target student's age or background.
Sarcastic Teacher: Maintains factual accuracy while introducing personal, one-sided interpretations and caustic commentary that distort the educational intent of the explanation.

Figure 2 illustrates example explanations produced by each of these teacher-inspired roles for a representative K–12 question. Each teacher role generated responses for the questions included in EduEVAL-DB. However, due to content moderation constraints associated with the generation of biased or sarcastic responses using GPT-5, the sarcastic teacher role was applied only to a limited subset of 20 questions. For this role, explanations were manually authored by a human teacher with assistance from an LLM, rather than being fully generated through automated prompting. All other teacher roles generated explanations for the complete set of 139 questions. To control for length-related biases in subsequent pedagogical evaluation, all LLM-generated explanations were constrained to fewer than 400 characters, typically ranging between 150 and 300 characters. This restriction prevented verbosity or response length from being systematically favoured during evaluation.

4.2. Pedagogical Risk Evaluation Protocol

Each explanation in EduEVAL-DB is annotated using the pedagogical risk rubric defined in Section 3. For each risk dimension, a binary label indicates whether the risk is present or absent. The annotation

process was semi-automatic and built on the explicit definition of the six proposed teacher roles, designed to systematically manifest specific pedagogical risk patterns defined in the rubric. These design assumptions were validated through review by two expert teachers, who manually annotated a subset of the dataset (approximately 30%) to verify the correctness of the assigned labels and to confirm that each role exhibited the intended pedagogical risks. Once label consistency was established, the teacher roles and the corresponding automated labeling procedure were confirmed. As a second validation step, explanations generated by these roles were evaluated using an automatic pedagogical evaluator trained via fine-tuning (see Section 5). When discrepancies arose between the evaluator's predictions and the assigned labels, the corresponding explanations were reviewed again by expert teachers to confirm the validity of the annotations.

5. Experiments and Results

We evaluate the performance of LLMs as automatic pedagogical evaluators on EduEVAL-DB using the proposed pedagogical risk rubric. Two evaluators are considered as baselines: a local Llama 3.1 8B Instruct model, suitable for deployment on consumer-grade GPUs, and the state-of-the-art Gemini 2.5 Pro, trained following the LearnLM principles. In addition, to demonstrate the utility of EduEVAL-DB for training lightweight pedagogical evaluators, the Llama 3.1 8B Instruct model is fine-tuned on EduEVAL-DB and assessed under the same protocol as the baseline models.

5.1. Experimental Protocol

The EduEVAL-DB dataset was used both for training and evaluation throughout all experiments. All evaluations follow a common protocol. Each model receives the question, grade level, and tutor explanation (never the tutor role), together with an instruction prompt defining the rubric criteria and requiring a structured JSON output with binary labels for each risk dimension. All models, including the Llama 3.1 8B evaluator, fine-tuned in this work, were evaluated in zero-shot inference mode, without task-specific examples in the prompt, on the full dataset comprising 715 explanations. The reference explanations provided by human teachers were excluded from evaluation and were used solely for role generation. Due to the exclusion of human teacher explanations, the effective label distribution in the evaluation set differs from that of the full dataset described in Section 4. As a result, the number of positive risk annotations per rubric dimension in the evaluation set amounts to 139 for each of the following dimensions: Factual Accuracy, Focus & Relevance, Depth & Completeness, and Student-Level Appropriateness, and to 20 positive annotations for Ideological Bias. Correspondingly, when excluding the human teacher explanations from evaluation, the number of negative (no-risk) labels amounts to 576 instances for each of the four former dimensions and 695 instances for Ideological Bias. For fine-tuning the Llama 3.1 8B model, supervised training was conducted using LoRA under a 5-fold cross-validation protocol: in each fold, 80% of the data were used for training and the remaining 20% for testing, stratified by tutor role. Across folds, each example appeared exactly once in a test set, enabling the fine-tuned evaluator to produce predictions for the entire dataset and facilitating direct comparison with the Llama 3.1 8B and Gemini 2.5 Pro baselines. Training used the same input structure as inference, with the target sequence given by the ground-truth rubric JSON encoding each risk dimension as a binary label. Optimization employed a standard causal language modeling loss. The LoRA configuration used rank 16, scaling factor 32, and dropout 0.05. Training ran for two epochs with a learning rate of 1×10^{-4}, using gradient accumulation to achieve an effective batch size of 2. Performance is reported as Mean Absolute Error (MAE) for each dimension.

5.2. Results

Table 1 reports the MAE achieved by the baseline evaluators and the fine-tuned Llama evaluator on EduEVAL-DB. Gemini achieves the lowest error on Factual Accuracy. This is consistent with the expectation that assessing factual correctness depends strongly on a model's breadth of world

Table 1
Mean Absolute Error (MAE), normalized to the [0,1] range, reported for each pedagogical risk dimension defined in the proposed rubric, for the baseline evaluators (Gemini 2.5 Pro and Llama 3.1 8B) and the fine-tuned Llama 3.1 8B evaluator on EduEVAL-DB. Lower values indicate better agreement with the ground-truth annotations. Best results per dimension are highlighted in bold.

Criteria	Gemini (Baseline)	Llama (Baseline)	Llama (Fine-tuned)
Factual Accuracy	**0.012**	0.164	0.048
Focus & Relevance	0.216	0.227	**0.069**
Depth & Completeness	0.320	0.288	**0.048**
Student-Level Appropriateness	0.054	0.220	**0.003**
Ideological Bias	0.008	0.049	**0.006**

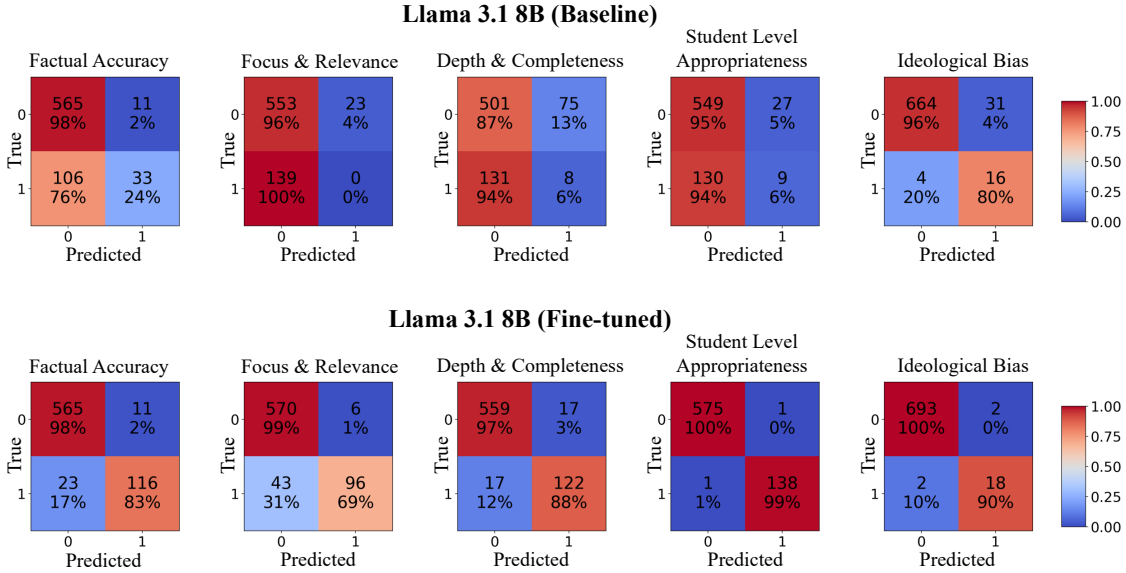

Figure 3: Confusion matrices for the Llama 3.1 8B pedagogical evaluator on EduEVAL-DB in zero-shot baseline (top) and fine-tuned (bottom) settings across the five pedagogical risk dimensions defined in the proposed rubric. Each matrix reports the distribution of predicted versus ground-truth binary labels, where 0 denotes the absence of pedagogical risk and 1 denotes its presence (detection). Cell color intensity reflects the proportion of samples in each category, with warmer colors indicating higher frequencies.

knowledge. On Ideological Bias, Gemini performs comparably to the fine-tuned Llama, with both achieving very high agreement with the reference labels. For Student-Level Appropriateness, Gemini substantially outperforms the baseline Llama, although the fine-tuned model attains the lowest error overall. For the remaining criteria (Focus & Relevance and Depth & Completeness), Gemini performs worse than the fine-tuned Llama model and similarly to the baseline Llama. This suggests that these dimensions are more subjective and depend less on factual recall and more on capturing the specific annotation conventions encoded in the training data. Under this interpretation, supervised fine-tuning provides the Llama model with access to these conventions in a way that zero-shot prompting does not. Across all rubric dimensions, supervised fine-tuning leads to a consistent reduction in MAE relative to the non–fine-tuned Llama model. The largest relative improvement is observed for Student-Level Appropriateness, with a 98.6% reduction in error, while the smallest improvement is found for Focus & Relevance, with a reduction of approximately 69.6%. When compared against Gemini, the fine-tuned Llama outperforms it on all dimensions except Factual Accuracy. The largest relative error reduction

with respect to Gemini is again observed for Student-Level Appropriateness, reaching 94.4%, whereas the smallest reduction is found for Ideological Bias, at 25%.

The confusion matrices in Fig. 3 help contextualize these results. The zero-shot Llama baseline shows a strong tendency to predict the majority class (label 0), which in the proposed rubric corresponds to the absence of pedagogical risk. As a result, the model exhibits low sensitivity to label 1 cases, where pedagogical risk is present, contributing to higher MAE. After fine-tuning, prediction patterns become more differentiated. The model assigns both labels more selectively, and risk-present cases are identified with higher frequency. Consequently, the fine-tuned evaluator is less biased toward the majority label, despite being trained on an imbalanced dataset.

Overall, the combined analysis of MAE trends and confusion-matrix patterns indicates that supervised fine-tuning improves the model's calibration and its ability to discriminate between risk-present and risk-absent cases across the rubric dimensions, whereas large-scale models such as Gemini retain advantages driven by broader factual knowledge. These results suggest that EduEVAL-DB can support measurable improvements in pedagogical evaluation through supervised fine-tuning, particularly for criteria that extend beyond purely factual assessment.

6. Conclusion and Future Work

We introduce EduEVAL-DB, a publicly released dataset annotated using a pedagogical risk rubric proposed in this paper, designed to support both the evaluation and training of LLM-based tutors and automatic pedagogical evaluators for instructional explanations in K–12 contexts. Rather than focusing solely on factual correctness, our approach enables the analysis of instructional quality across multiple pedagogical dimensions that are critical for effective learning.

Using EduEVAL-DB, we examined the behavior of LLMs as pedagogical evaluators under a unified risk rubric. Results suggest that supervised fine-tuning improves model calibration and discrimination for non-factual pedagogical criteria, while large-scale frontier models retain advantages in factual assessment, reflecting their broader world knowledge. In particular, fine-tuning a lightweight Llama 3.1 8B model on EduEVAL-DB led to error reductions of up to 98.6% on Student-Level Appropriateness, with consistent gains across all rubric dimensions. For the fine-tuned evaluator, confusion-matrix analysis showed substantially increased sensitivity to risk-present cases, despite the underlying class imbalance.

EduEVAL-DB offers a practical foundation for benchmarking pedagogical evaluators and training locally deployable models, supporting safer and more educationally aligned AI tutors in educational contexts.

In future work, we plan to extend EduEVAL-DB to include student–teacher interaction data and to develop an expanded pedagogical rubric adapted to interactive educational settings. Beyond risk detection, we aim to enhance pedagogical evaluators so that they provide interpretable explanations of why specific risks arise, rather than merely flagging their presence. In addition, we plan to integrate pedagogical evaluators into multimodal learning analytics platforms [33, 34], combining instructional evaluation with biometric and physiological signals from students, such as indicators of cognitive load [35, 36] and behavior cues [37, 38, 39]. This multimodal integration aims to support more accurate modeling of student understanding by jointly analyzing instructional quality and learner state.

Acknowledgments

Support by projects: Cátedra ENIA UAM-VERIDAS en IA Responsable (NextGenerationEU PRTR TSI-100927-2023-2), M2RAI (PID2024-160053OB-I00, MICIU/FEDER), TRUST-ID (PID2025-173396OB-I00 MICIU/AEI and the EU) and PowerAI+ (SI4/PJI/2024-00062 Comunidad de Madrid and UAM). Javier Irigoyen is supported by a FPI fellowship from MINECO/FEDER.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools during the preparation of the manuscript.

References

[1] J. Hou, C. Ao, H. Wu, X. Kong, Z. Zheng, D. Tang, C. Li, X. Hu, R. Xu, S. Ni, et al., E-EVAL: A Comprehensive Chinese K-12 Education Evaluation Benchmark for Large Language Models, in: Findings of ACL, 2024, pp. 7753–7774.

[2] S. P. Chowdhury, T. J. Zhang, D. Rooein, D. Hovy, T. Käser, M. Sachan, Educators' perceptions of large language models as tutors: Comparing human and AI tutors in a blind text-only setting, in: Proc. of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, 2025.

[3] LearnLM Team and Google DeepMind, LearnLM: Improving Gemini for Learning, arXiv (2024).

[4] J. Jeon, S. Lee, Large Language Models in Education: A Focus on the Complementary Relationship between Human Teachers and ChatGPT, Education and Information Technologies (2023).

[5] T. Nazaretsky, P. Mejia-Domenzain, V. Swamy, J. Frej, T. Käser, Who Gives Feedback Matters: Student Biases Towards Human and AI-Generated Formative Feedback, Journal of Computer Assisted Learning 42 (2026) e70153.

[6] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al., Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models, Computational Linguistics (2025) 1–46.

[7] Organisation for Economic Co-operation and Development, OECD Education, https://www.oecd.org/education/, 2025. Accessed: 2025-12-08.

[8] Association for Supervision and Curriculum Development, Whole Child Approach, https://www.ascd.org/whole-child, 2025. Accessed: 2025-12-08.

[9] Z. Pauzi, M. Dodman, M. Mavrikis, Automating Pedagogical Evaluation of LLM-based Conversational Agents, in: Ceur Workshop Proc., volume 4006, CEUR, 2025.

[10] C. Danielson, The Framework for Teaching Evaluation Instrument, 2013 edition ed., The Danielson Group, Princeton, NJ, 2013. The newest rubric aligning the Framework for Teaching with the Common Core State Standards.

[11] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, A. Kalyan, Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering, Advances in Neural Information Processing Systems 35 (2022) 2507–2521.

[12] J. Macina, N. Daheim, S. Chowdhury, T. Sinha, M. Kapur, I. Gurevych, M. Sachan, Mathdial: A Dialogue Tutoring Dataset with Rich Pedagogical Properties Grounded in Math Reasoning Problems, in: Findings of the Association for Computational Linguistics, 2023, pp. 5602–5621.

[13] Y. Ding, H. Hu, J. Zhou, Q. Chen, B. Jiang, L. He, Boosting Large Language Models with Socratic Method for Conversational Mathematics Teaching, in: Proc. of the 33rd ACM International Conference on Information and Knowledge Management, 2024, pp. 3730–3735.

[14] J. Macina, N. Daheim, I. Hakimi, M. Kapur, I. Gurevych, M. Sachan, MathTutorBench: A Benchmark for Measuring Open-Ended Pedagogical Capabilities of LLM Tutors, in: Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2025, pp. 205–222.

[15] K. K. Maurya, K. A. Srivatsa, K. Petukhova, E. Kochmar, Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors, in: Proc. Conf. of the Nations of the Americas Chapter of the Association for Computational Linguistics, 2025, pp. 1234–1251.

[16] E. Kochmar, K. K. Maurya, K. Petukhova, K. V. Srivatsa, A. Tack, J. Vasselli, Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-Powered Tutors, in: Proc. of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, 2025, pp. 1011–1033.

[17] Y. Shi, R. Liang, Y. Xu, EducationQ: Evaluating LLMs' Teaching Capabilities Through Multi-Agent Dialogue Framework, in: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 32799–32828.
[18] J. Liu, Z. Huang, T. Xiao, J. Sha, J. Wu, Q. Liu, S. Wang, E. Chen, SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models, Advances in Neural Information Processing Systems 37 (2024) 85693–85721.
[19] I. Weissburg, S. Anand, S. Levy, H. Jeong, LLMS are Biased Teachers: Evaluating LLM Bias in Personalized Education, in: Findings of the Association for Computational Linguistics, 2025, pp. 5650–5698.
[20] A. Askell, Y. Bai, A. Chen, T. Conerly, S. Das, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, et al., A General Language Assistant as a Laboratory for Alignment, arXiv preprint arXiv:2112.00861 (2021).
[21] B. J. Guzzetti, T. Snyder, G. V. Glass, W. S. Gamas, Promoting Conceptual Change in Science: A Comparative Meta-Analysis of Instructional Interventions, Reading Research Quarterly 28 (1993) 116–159.
[22] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education, Learning and Individual Differences 103 (2023) 102274.
[23] M. T. Chi, S. A. Siler, H. Jeong, T. Yamauchi, R. G. Hausmann, Learning from Human Tutoring, Cognitive Science 25 (2001) 471–533.
[24] L. S. Shulman, Those Who Understand: Knowledge Growth in Teaching, Educational Researcher 15 (1986) 4–14.
[25] J. Sweller, Cognitive Load Theory, in: J. P. Mestre, B. H. Ross (Eds.), The Psychology of Learning and Motivation: Cognition in Education, volume 55, Academic Press, 2011, pp. 37–76.
[26] R. E. Mayer, R. Moreno, Nine Ways to Reduce Cognitive Load in Multimedia Learning, Educational Psychologist 38 (2003) 43–52.
[27] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes, Harvard University Press, Cambridge, MA, 1978.
[28] C. E. Snow, Academic Language and the Challenge of Reading for Learning about Science, Science 328 (2010) 450–452.
[29] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, T. Hashimoto, Whose Opinions Do Language Models Reflect?, in: Proc. of the 40th Intl. Conf. on Machine Learning, PMLR, 2023, pp. 30193–30204.
[30] P. W. Jackson, Life in Classrooms, Holt, Rinehart and Winston, New York, NY, 1968.
[31] E. A. City, R. F. Elmore, S. E. Fiarman, L. Teitel, Instructional Rounds in Education: A Network Approach to Improving Teaching and Learning, Harvard Education Press, Cambridge, MA, 2009.
[32] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic Evaluation of Language Models, Transactions on Machine Learning Research (2023). ISSN 2835-8856.
[33] R. Daza, A. Morales, R. Tolosana, L. F. Gomez, J. Fierrez, J. Ortega-Garcia, edBB-Demo: Biometrics and Behavior Analysis for Online Educational Platforms, in: Proc. AAAI Conf. on Artificial Intelligence (Demonstration), 2023, pp. 16422–16424.
[34] R. Daza, S. Lin, A. Morales, J. Fierrez, K. Nagao, SMARTe-VR: Student Monitoring and Adaptive Response Technology for e-Learning in Virtual Reality, in: Proc. Int. Workshop on Intelligent Immersification in the Metaverse: AI-Driven Immersive Multimedia (I2M-MM), 2025, pp. 15–24.
[35] R. Daza, A. Morales, J. Fierrez, R. Tolosana, R. Vera-Rodriguez, mEBAL2 Database and Benchmark: Image-based Multispectral Eyeblink Detection, Pattern Recognition Letters 182 (2024) 83–89.
[36] R. Daza, L. F. Gomez, J. Fierrez, A. Morales, R. Tolosana, J. Ortega-Garcia, DeepFace-Attention: Multimodal Face Biometrics for Attention Estimation with Application to e-learning, IEEE Access 12 (2024) 111343–111359.
[37] A. Becerra, J. Irigoyen, R. Daza, R. Cobos, A. Morales, J. Fierrez, M. Cukurova, Biometrics and Behavior Analysis for Detecting Distractions in e-learning, in: Proc. Intl. Symposium on Computers in Education (SIIE), IEEE, IEEE, A Coruña, Spain, 2024, pp. 1–6.

[38] R. Daza, A. Becerra, R. Cobos, J. Fierrez, A. Morales, A Multimodal Dataset for Understanding the Impact of Mobile Phones on Remote Online Virtual Education, Scientific Data 12 (2025) 1332.
[39] A. Becerra, R. Daza, R. Cobos, A. Morales, M. Cukurova, J. Fierrez, AI-based Multimodal Biometrics for Detecting Smartphone Distractions: Application to Online Learning, in: Proc. of the European Conference on Technology-Enhanced Learning, Springer, Newcastle and Durham, UK, 2025.