Automating Pedagogical Evaluation of LLM-based Conversational Agents

Zaki Pauzi[1,*], Michael Dodman[2,†] and Manolis Mavrikis[1]

[1]UCL Knowledge Lab, United Kingdom
[2]Individual contributor

Abstract

With the growing adoption of large language models (LLMs) in educational settings, there is an urgent need for systematic and scalable evaluation methods. Traditional natural language generation metrics such as BLEU, ROUGE and METEOR excel at measuring surface-level linguistic quality but fall short in evaluating the interactive, adaptive nature of dialogue alignment of conversational agents, particularly in relation to their intended design. To address these gaps, we propose an evaluation strategy that extends beyond technical evaluation (linguistic coherence and semantic relevance). In this pilot study we compare human and LLM-based evaluation of a conversational agent, with a focus on Socratic dialogue as a specific instantiation. Early results indicate that our LLM-as-a-Judge aligns closely to human evaluators for clear, surface-level qualities like encouragement and actionable guidance, but less on subtle pedagogical behaviours such as recognising errors and maintaining natural dialogue flow. These early results underscore the promise of LLM-based evaluators for scalable assessment of tutoring behaviours while highlighting the need for targeted fine-tuning and hybrid approaches to improve nuanced error detection and dialogue coherence.

Keywords

pedagogical rubric, automated evaluation, Socratic dialogue, AI tutor

1. Introduction

With the rapid advancement of artificial intelligence (AI) in education, there has been an increasing interest in developing intelligent tutoring systems to enhance and improve students' learning experiences. One such approach is an AI tutor, which employs the Socratic dialogue method, i.e., encouraging students to think critically by asking thought-provoking questions rather than providing direct answers [1, 2, 3]. Unlike conventional tutoring models that rely on direct explanations or answer provision, Socratic tutors employ dialogue to guide students through self-discovery, prompting reflection and exploration of underlying concepts.

Despite advances in natural language generation (NLG) and large language models (LLMs), existing evaluation metrics such as BLEU, ROUGE, METEOR, and BERTScore predominantly assess surface-level linguistic quality and semantic similarity. While these metrics effectively measure fluency and coherence, they do not capture the adaptive, interactive, and pedagogical nuances inherent in Socratic tutoring. To bridge this gap, we introduce a pedagogically grounded evaluation framework tailored to AI tutors deploying Socratic dialogue.

Our contribution is twofold: first, we propose a rubric that operationalises key dimensions of Socratic tutoring: Mistake Remediation, Scaffolding, Guidance and Actionability, and Coherence and Tone. Second, we present a pilot study contrasting **human annotators'** scoring against those generated by an **LLM-as-a-judge**. By grounding evaluation with pedagogical aspects, our framework provides a holistic, learner-centered approach to assessing AI-driven tutors. This work supports researchers and practitioners in fine-tuning Socratic tutoring systems, ultimately enhancing the quality and effectiveness of AI-mediated tutoring.

EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy
*Corresponding author.
†Majority of the work was completed while the author was at Faculty (Faculty Science Limited).
z.pauzi@ucl.ac.uk (Z. Pauzi); mrqdodman@gmail.com (M. Dodman); m.mavrikis@ucl.ac.uk (M. Mavrikis)
0000-0003-4032-4766 (Z. Pauzi); 0000-0003-0575-0823 (M. Mavrikis)

2. Background

The AI tutor is built on the principles of Socratic teaching — an approach that uses questioning and dialogue to guide students toward deeper understanding, commonly associated with the objective of developing critical thinking, rather than directly lecturing or providing answers [1, 2, 4]. This AI tutor acts as a teacher's assistant, helping to extend the teacher's ability to ask questions while the students are working. By having these conversations, the teacher is able to know the magnitude of a student's understanding in a particular subject or topic and uncover what their misconceptions are. A key limitation of doing this manually is the enormous amount of time and effort required by teachers to participate in these conversations.

The evaluation of large language models (LLMs) has advanced considerably, particularly in their application to AI tutoring systems. Early approaches relied on traditional natural language generation (NLG) metrics, such as BLEU and ROUGE, which are effective in assessing fluency, but insufficient in capturing the pedagogical effectiveness of AI tutors [5]. To address these shortcomings, Maurya et al. [6] introduced the Unified Evaluation Taxonomy, a framework for systematically assessing AI tutors based on eight pedagogical dimensions: mistake identification, mistake location, revealing the answer, providing guidance, actionability, coherence, tutor tone, and human-likeness. This taxonomy provides a standardised approach to evaluating AI tutors and aligns with established educational best practices.

2.1. Automated technical metrics for assessing dialogue quality

Automated metrics provide quantitative tools to assess the technical quality of dialogue systems, including AI tutors. In assessing the quality of Socratic dialogues, Favero et al. [7] used BERTScore [8], which employs contextual embeddings to evaluate semantic similarity, offering a deeper evaluation of meaning compared to BLEU [9], ROUGE [10], and METEOR [11] Another recent metric used is DialogRPT, which is tailored for conversational AI. It evaluates response coherence and human-likeness, making it particularly relevant for dialogue-based tutoring systems [6]. This metric was also used with BERTScore in a recent shared task in 2023 to automatically evaluate teacher-student dialogue [12].

These metrics can be valuable for assessing the *technical* aspects of the responses of AI Socratic tutors. However, these automated NLP metrics have limitations when it comes to assessing the *pedagogical* aspects effectively and the interactive nature of AI Socratic tutors (this was done manually in the aforementioned shared task [12]). Moreover, many of these metrics are based on a comparison with predetermined reference answers, which may not be entirely suitable for Socratic dialogue, where the learning process and the individual exploration of the student are more important than arriving at a specific answer [13]. While technical metrics offer a valuable assessment of the surface-level linguistic quality of the dialogue, they lack the capacity to evaluate the pedagogical impact of the Socratic interaction.

2.2. Pedagogical frameworks and taxonomies for evaluating AI tutors

To address the limitations of technical metrics, pedagogical frameworks and taxonomies offer structured ways to evaluate the educational effectiveness of AI tutors employing Socratic dialogue. One notable approach is the eight dimensions taxonomy for pedagogical abilities proposed in recent research [6]. Grounded in learning sciences principles, this taxonomy provides a standardised approach to evaluating key aspects of the tutoring interaction and the AI's ability to effectively address student errors. Each dimension is evaluated using a three-tier labelling system (Yes, To some extent, No), allowing for nuanced assessment of the AI tutors' pedagogical abilities.

Automated technical metrics offer valuable quantitative data regarding the linguistic quality of the dialogue, including aspects such as fluency, coherence, and semantic similarity [14]. On the other hand, pedagogical frameworks provide qualitative insights into the learning process, focusing on the AI's ability to guide student thinking, foster critical reasoning, and promote deeper understanding. Human feedback also plays a crucial role in evaluating aspects that automated systems may find challenging

to assess, such as the depth of critical thinking stimulated by the interaction and the overall learning experience from the student's perspective [15].

3. Rubric for automated evaluation of Socratic dialogue

Adapted from Maurya et al. [6] and learning sciences literature (e.g., [16, 3]), we propose the following set of dimensions, metrics and questions that form our proposed tentative rubric framework (Table 1) for automated evaluation of AI tutors.

Table 1
Rubric for AI tutor evaluation

Dimension	Metric	Question	Answer type
Mistake Remediation	Identifying mistake	Does the tutor correctly recognise that the student has made a mistake?	Yes/No/NA
	Acknowledging mistake	Does the tutor show that they notice the mistake?	Yes/No/NA
Scaffolding	Provision of appropriate support without giving answers directly	Does the tutor help the student without directly giving the answer?	Yes/No
Guidance and Actionability	Provision of guidance to move to the next step	Based on the tutor's responses in the conversation, is there clear guidance on what the student should do next?	Yes/No
Coherence and Tone	Natural-ness of conversation flow	Is the dialogue between the student and tutor natural?	Yes/No
	Encouraging tone	Overall, Is the tutor's tone encouraging and supportive?	Yes/No

Our rubric operationalises Socratic tutoring principles by breaking them down into four overarching dimensions—Mistake Remediation, Scaffolding, Guidance and Actionability, and Coherence and Tone—each with specific metrics and binary (Yes/No) judgments, with "NA" (Not Applicable) as an option for Mistake Remediation (when there is no mistake that needs to be remedied).

3.1. Pilot case study

In our case study, we evaluated the AI tutor built on an EdTech platform supporting students in building concept maps. This tool facilitates the students to define nodes and relationships (concept mapping) based on a topic and/or subject. In this exercise, the students are asked to build and link nodes regarding the topic of **"How do we move our school paper registers to an IT system?"**.

3.2. Dialogue sample creation

We created the dialogue turns using OpenAI's GPT-4o to reflect two contrasting tutor styles (Socratic vs. non-Socratic). For each stage of the tutoring process—Early, Mid, and End—we generated 15 turns per dialogue snippet in each style, yielding a total of 90 turns.

To ensure our evaluation is robust and generalises beyond a handful of canned examples, we deliberately built *variability* into both content and form. By sampling multiple dialogue variants, we can collect a distribution of rubric scores—both from human raters and from an LLM-as-judge—and examine not only mean performance but also the *spread* of scores. A narrow spread would imply consistent rubric compliance, whereas a wide spread might reveal varying rubric conformance or tutor strategies that only work in certain phrasings. This systematic injection of variability allows us to stress-test each rubric dimension and to ensure that our evaluation is assessing the *functionality* of the tutor's moves—not just the superficial wording.

In the system prompt for both versions, the following criteria were explicitly mentioned:

1. Not all metrics presented in our proposed rubric should be satisfied.
2. Each scenario reflects different stages in the concept mapping exercise (i.e., Early, Mid, and End).
3. Add variability in some stages where the tutor fails on some metrics, particularly on the version where the tutor is not being *Socratic*.

The following shows an excerpt of the Early stage dialogue between a student and the AI tutor:

```
[Start of dialogue]
Tutor: What do you think is the main problem with using paper registers in
your school?
Student: They're slow and sometimes teachers forget to fill them out.
Tutor: That's a good observation. How might that affect the school as a whole?
Student: Maybe it causes delays in knowing who's present or absent.
Tutor: Great. Can you think of what kinds of data a digital register system
would need to handle?
...
```

The conversation data used for this evaluation can be found at https://zenodo.org/records/15724072.

3.3. Human evaluation

A total of twelve independent human evaluators scored the dialogue snippets. For highest level of education, ten hold postgraduate degrees (Master's or PhD) and two hold Bachelor's degrees. Their self-reported familiarity with the Socratic method ranged from "Not familiar" (5 raters) through "Heard of it" (1), "Somewhat familiar" (1), "Familiar" (4), to "Very familiar" (1). In terms of domain expertise, the majority (7 raters) work in Computer Science/AI/ML; additional backgrounds include Human–Computer Interaction (3 raters), Education/Pedagogy (2), Engineering (1), International Business Administration (1), Finance (1), and Law (1). This mix of advanced academic credentials and varied disciplinary perspectives helps ensure a balanced evaluation of both pedagogical style and content accuracy. Evaluators were shown the rubric (Table 1) and the metrics were explained prior to scoring. For each dialogue scenario, they answered the six Yes/No questions in Table 1.

3.4. LLM-as-a-Judge evaluation

We prompted a task-specific LLM—LearnLM 2.0 Flash Experimental [1]—to act as an evaluator using the same rubric metrics. LearnLM is an experimental task-specific model that has been trained to align with learning science principles. The concept and implementation of *LLM-as-a-Judge* is not particularly new in the field of automated evaluation [17, 18]; however, in tasks that require specialised knowledge, the validity of using LLMs as judges remains uncertain, and human experts are still necessary to be in the loop [19]. For each dialogue scenario, we constructed a prompt of the form:

```
System: You are an educational evaluator of an AI tutor. The following is the
rubric you will be using to evaluate the tutor's response.
<rubric>
The following is the current dialogue snippet from [Early/Mid/End] stage of
the exercise.
<dialogue>
Task: Given the student's and the tutor's response, answer the
following questions. The answer options are shown in square brackets for each
question:
<question>
```

[1]Technical report: https://goo.gle/LearnLM-dec24, API docs: https://ai.google.dev/gemini-api/docs/learnlm

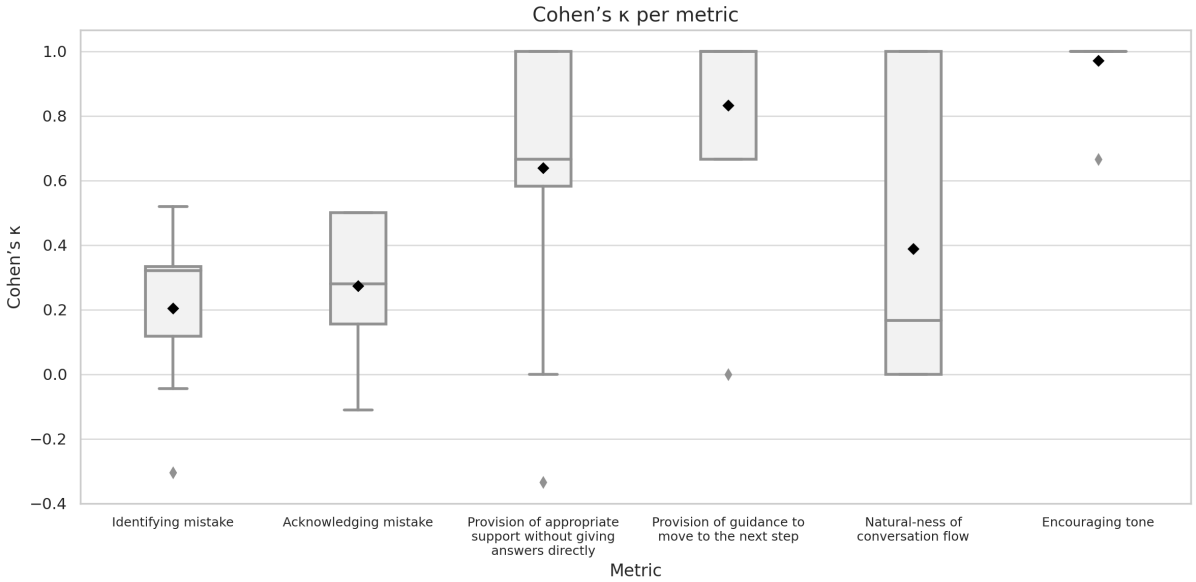

Figure 1: Distribution of Cohen's κ scores between the LLM-based evaluator and human annotators for each Socratic tutoring metric. Each box represents the interquartile range (IQR) with the median (line) and mean (◆), whiskers extend to the most extreme data points not considered outliers, and individual markers denote outliers.

We ran the prompt in a *zero-shot* setting and collected the model's scoring. We then measured alignment with the human annotations via exact match rate and Cohen's κ per metric, and analysed common error patterns to identify areas for prompt or model improvement. We chose a different LLM to the synthetic sample creation given the inherent bias that exists in LLMs evaluating own generation [20].

4. Early Results and Discussion

Figure 1 shows the box plot of Cohen's κ value between the human evaluators and the LLM-as-a-Judge.

The box plot highlights notable differences in how well our LLM evaluator aligns with human judgments across the six metrics. At one end of the spectrum, **Encouraging tone** shows near-perfect agreement (mean $\kappa \approx 0.97$), with nearly all values clustered at or very close to 1.00, indicating that the model almost always matches human evaluators in detecting a supportive, positive tone. Similarly, **Provision of guidance to move to the next step** yields strong consensus (mean $\kappa \approx 0.83$), with most scores between 0.67 and 1.00 and only a single outlier at 0.00.

In contrast, the model struggles with the more nuanced task of mistake remediation: both **Identifying mistake** and **Acknowledging mistake** hover around low medians (≈ 0.20 and 0.28, respectively), feature wide ranges extending slightly below zero, and even exhibit negative outliers—signalling occasional direct disagreement with human raters. **Provision of appropriate support without giving answers directly** (median $\kappa \approx 0.62$) shows generally reliable but imperfect alignment, while **Natural-ness of conversation flow** proves the *most variable* metric (mean $\kappa \approx 0.39$, full range 0.00–1.00), suggesting that judgments of conversational coherence are highly subjective and context-dependent.

Overall, these results indicate that our LLM evaluator excels at scoring clear, surface-level qualities like encouragement and actionable guidance, but faces challenges in replicating human sensitivity to subtle pedagogical behaviours such as recognising errors and maintaining natural dialogue flow.

4.1. Limitations

Despite these insights, our study has several limitations. First, our evaluation relies on a small, synthetic corpus of concept-mapping dialogues within a single domain, which may not generalise to other subjects or authentic classroom settings. Second, the binary Yes/No rubric captures only the presence or

absence of target behaviours and cannot distinguish varying degrees of instructional quality or nuance. Regardless, this was intentional to ensure that for the pilot study, the LLM responses can only be chosen from a set of options (for uniformity and constraint in LLM response). Third, while our LLM-as-a-Judge leverages an education-tuned model, we evaluated it under zero-shot prompting rather than few-shot or comprehensive fine-tuning on pedagogical annotations, which likely constrained its sensitivity to subtle teaching cues. Finally, variability in human annotator backgrounds and interpretations introduces additional noise, prompting the need for clearer guidelines and more diverse evaluator pools in future work.

5. Conclusion

In conclusion, early results from our pilot study suggests that LLM evaluators may effectively and efficiently assess clear, surface-level tutoring behaviours but currently fall short on nuanced error recognition and conversational coherence. Future work will focus on further refining the rubric framework, targeted fine-tuning, improved prompting strategies, and hybrid human–LLM workflows to enhance subtle pedagogical judgments. Additionally, extending evaluations to other subject domains and conducting longitudinal studies on student learning outcomes will be critical for building robust, generalisable assessment frameworks.

Acknowledgments

Part of this work was completed when one of the authors (MD) was on a "Faculty fellowship" with the EdTech startup, Graffinity (https://graffinity.io/), as a supplement to the their "Contracts for Innovation: AI tools for education" catalyst project funded by the UK Department for Education. We would also like to acknowledge Matthew Pryor (Graffinity) and Harry Moss (UCL's ARC - Advanced Research Computing centre) for feedback and support in related work during the same project. ZP's contribution was co-funded through UCL's Global Engagement Seed Funding and the TransEET project (https://transeet.eu/) in the context of the AIED Community of Interest (COI). TransEET is HORIZON-WIDERA-2021-ACCESS-03-01 project (Grant no. 101078875) funder by the EU and by the UK Research and Innovation (UKRI) under the UK government's Horizon Europe Guarantee funding scheme. MM's contribution was partly supported from the CHAILD project - Children's Agency In the age of AI: Leveraging InterDisciplinarity (UKRI ref. MR/Z505882/1). The views expressed in this publication are those of the authors and not necessarily those of their institutions or the funders.

Declaration on Generative AI

The authors have not employed any Generative AI tools in the writing of this paper.

References

[1] Y.-R. Ho, B.-Y. Chen, C.-M. Li, Thinking more wisely: using the socratic method to develop critical thinking skills amongst healthcare students, BMC Medical Education 23 (2023). doi:10.1186/s12909-023-04134-2.

[2] D. R. Oyler, F. Romanelli, The fact of ignorance revisiting the socratic method as a tool for teaching critical thinking, American Journal of Pharmaceutical Education 78 (2014) 144. doi:https://doi.org/10.5688/ajpe787144.

[3] N.-T. Le, How do technology-enhanced learning tools support critical thinking?, in: Frontiers in Education, volume 4, Frontiers Media SA, 2019, p. 126.

[4] R. Paul, L. Elder, Critical thinking: The art of socratic questioning, Journal of developmental education 31 (2007) 36.

[5] G. Bonino, G. Sanmartino, G. G. Pinheiro, P. Papotti, R. Troncy, P. Michiardi, Euler: Fine tuning a large language model for socratic interactions, in: Proceedings of the Second International Workshop on Artificial Intelligence Systems in Education co-located with 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), volume—of CEUR Workshop Proceedings, 2024.
[6] 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, 2025. URL: https://arxiv.org/abs/2412.09416. arXiv:2412.09416.
[7] L. Favero, J. A. Pérez-Ortiz, T. Käser, N. Oliver, Enhancing critical thinking in education by means of a socratic chatbot, 2024. URL: https://arxiv.org/abs/2409.05511. arXiv:2409.05511.
[8] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).
[9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[10] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.
[11] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
[12] A. Tack, E. Kochmar, Z. Yuan, S. Bibauw, C. Piech, The BEA 2023 shared task on generating AI teacher responses in educational dialogues, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 785–795. doi:10.18653/v1/2023.bea-1.64.
[13] S. Mehri, J. Choi, L. F. D'Haro, J. Deriu, M. Eskenazi, M. Gasic, K. Georgila, D. Hakkani-Tur, Z. Li, V. Rieser, S. Shaikh, D. Traum, Y.-T. Yeh, Z. Yu, Y. Zhang, C. Zhang, Report from the NSF Future Directions Workshop on Automatic Evaluation of Dialog: Research Directions and Challenges, arXiv e-prints (2022) arXiv:2203.10012. doi:10.48550/arXiv.2203.10012.
[14] H. Fakour, M. Imani, Socratic wisdom in the age of ai: a comparative study of chatgpt and human tutors in enhancing critical thinking skills, Frontiers in Education 10 (2025). doi:10.3389/feduc.2025.1528603.
[15] E. Ilkou, S. Linzbach, J. Wallat, Hybrid evaluation of socratic questioning for teaching, 2024. doi:10.13140/RG.2.2.15670.31049.
[16] L. Elder, R. P. and, The role of socratic questioning in thinking, teaching, and learning, The Clearing House: A Journal of Educational Strategies, Issues and Ideas 71 (1998) 297–301. doi:10.1080/00098659809602729.
[17] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2023) 46595–46623.
[18] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, et al., A survey on llm-as-a-judge, arXiv preprint arXiv:2411.15594 (2024).
[19] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, R. A. Metoyer, Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, in: Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 952–966. doi:10.1145/3708359.3712091.
[20] B. Abeysinghe, R. Circi, The challenges of evaluating llm applications: An analysis of automated, human, and llm-based approaches, arXiv preprint arXiv:2406.03339 (2024).