

# Can LLMs Judge Pedagogy? Assessing Conversational AI STEM Tutoring with AI-as-a-Judge

**Xintian Gao**  
University of Florida  
Gainesville, FL, USA  
gaoxintian@ufl.edu

**Qian Shen**  
University of Florida  
Gainesville, FL, USA  
qian.shen1@ufl.edu

**Xin Li**  
University of Florida  
Gainesville, FL, USA  
xinli1@ufl.edu

## Abstract

Large language models are increasingly used as automated judges in education, yet their ability to score pedagogical quality in AI tutor responses to K-12 STEM student inquiries remains underexplored. This study evaluates whether two LLM-based scorers, Nemotron-3-Super-120B-A12B and GPT-OSS-120B, can approximate human judgment of single-turn Socratic style responses to STEM inquiries. Using a four-dimension rubric adapted from the CPS-R Questioning and Thinking subscale, we compare human and model ratings. Results show mixed reliability: agreement is stronger for more observable instructional features such as Cognitive Demand, but weaker for more interpretive dimensions, especially Encouraging Metacognition and Differentiation. Chance-corrected reliability is also sensitive to skewed score distributions, as shown by a base-rate effect in Differentiation. A mixed-effects analysis further reveals that the two LLM scorers differ systematically, with larger divergence on elementary-level items. We also observe prompt-adherence failures in generated tutoring responses, where some outputs briefly violate the instruction to avoid direct answers before returning to a Socratic response. Overall, the findings suggest that LLMs can assist with large-scale pedagogical evaluation, but human oversight remains necessary for nuanced instructional assessment and for maintaining Socratic response style.

## 1 Introduction

Large language models (LLMs) are increasingly being studied as automated judges of educational text because they can approximate human ratings when evaluation criteria are explicit and the response domain is well structured (Zheng et al., 2023; Maurya et al., 2025; Henkel et al., 2023, 2025). In education, recent reviews and empirical studies suggest that LLMs can support both assessment and feedback generation, especially when prompts include

clear rubrics, contextual item information, and examples of desired performance (Wang et al., 2026; Henkel et al., 2023; Arulsingh et al., 2026).

This line of work is especially relevant for evaluating tutor or teacher responses to student inquiries in STEM, where quality depends on both correctness and pedagogy (Maurya et al., 2025; Srinivasa et al., 2025). Relevant research proposes dimensions such as guidance, actionability, coherence, tone, and the ability to identify conceptual errors, showing that LLM-powered tutoring can be assessed with fine-grained pedagogical criteria rather than only answer accuracy (Maurya et al., 2025; Srinivasa et al., 2025). At the same time, studies of generated tutoring responses suggest that LLM outputs can be judged as pedagogically strong in some contexts, but their ratings may also be shaped by source bias and by ambiguity in the evaluation task (Henkel et al., 2025).

The evolution of conversational AI tutoring architectures requires a shift from binary correctness checks toward evaluating nuanced instructional dynamics. Recent benchmarks emphasize that a model's problem-solving proficiency does not naturally equate to skilled teaching intelligence (Shi et al., 2026). Consequently, static answer keys and surface-level linguistic metrics correlate poorly with true teaching efficacy (Kim et al., 2025; Shi et al., 2026). Evaluating student-centric tutoring requires process-level analysis of intermediate reasoning, multi-turn trajectories, and latent instructional traits rather than macro-level outcomes (Kim et al., 2025). Most LLM-based evaluation research focuses on grading student answers rather than assessing the pedagogical quality of educator or tutor responses themselves (Henkel et al., 2023). This leaves open how reliably the pedagogical quality of tutor responses themselves can be scored (Maurya et al., 2025; Shi et al., 2026). We address the single-turn case: whether LLM judges can approximate human ratings of one tutor response to one student

inquiry. Multi-turn tutoring dialogue is outside the scope of the present data.

To address this gap, our study analyzes 360 student inquiries from K–12 STEM courses and AI responses, each scored by both humans and two AI judges: Nemotron-3-Super-120B-A12B and GPT-OSS-120B, using the CPS-R Questioning and Thinking (QT) subscale (Pereira et al., 2021). We compare how these two models approximate human judgment to assess whether model choice affects reliability in pedagogical scoring. This provides insight into whether LLMs can reliably approximate human judgment of pedagogical quality, and whether model differences matter for scoring consistency. By doing so, the study contributes to LLM-assisted educational evaluation for AI responses to student inquiries and extends the literature beyond student-answer grading to the more nuanced task of evaluating the instructional quality of tutor responses themselves (Henkel et al., 2023, 2025).

## 2 Methodology

### 2.1 Data Source

This study uses an open-sourced educational dataset<sup>1</sup> from Kaggle that contains STEM students’ science inquiries to a chatbot tutor across three pre-defined educational levels (elementary, middle, and high school) and three subject areas (biology, physics, and chemistry). We randomly sampled 120 student questions per educational level with a nearly balanced distribution of subjects across all available data. The resulting baseline dataset comprises 360 STEM questions posed by students, together with their associated subject, educational levels, question types, and students’ purposes for asking the questions. For example, an elementary student posed the question “What makes mammals different from birds?” during biology homework.

### 2.2 Answer Generation

Drawing on socratic prompting framework for LLMs (Chang, 2023), we engineered a system prompt that operationalizes four phases of Socratic dialogue (definition, maieutics, elenchus, and counterfactual reasoning) (see Appendix 6.3) and applied it zero-shot via the Llama-3.3-70B-Instruct (Llama) API to generate instructional responses

for 360 student STEM questions across three educational levels (Elementary, Middle, and High School). To ensure responses were not constrained by format or length, we imposed no output token limit, and given that some questions required mathematical expressions, we permitted the model to use LaTeX-formatted formulas in its responses. For example, in response to the aforementioned elementary student’s question, the model generated: “Mammals are different from birds because mammals have hair or fur, and they drink milk from their mothers when they are babies, whereas birds have feathers and lay eggs. Mammals are also usually warm-blooded, meaning they can keep themselves warm, and they breathe air, just like birds. But mammals do not have beaks or wings like birds do.”

### 2.3 Scoring Framework

To evaluate the quality of instructional scaffolding in the LLM-generated responses, this study developed a codebook adapted from the Questioning and Thinking (QT) subscale of the Classroom Practices Survey-Revised (CPS-R), an instrument whose validity evidence was established for measuring teachers’ self-reported differentiation, metacognition, logical reasoning, and questioning practices at the classroom level (Pereira et al., 2021). The QT subscale consists of ten items spanning open-ended questioning, metacognitive prompts, follow-up questions, and opportunities for critical and creative thinking. Rather than scoring each item individually, we grouped them into three categories, Encouraging Metacognition (QT1, QT4, QT7), Cognitive Demand (QT2, QT8, QT9, QT10), and Learner Agency (QT3, QT5, QT6), and added a fourth Differentiation dimension capturing the calibration of vocabulary, linguistic complexity, and scaffolding depth to the student’s grade band (K-5, 6-8, 9-12). This grouping reduces analytic complexity while preserving theoretically meaningful distinctions; full definitions and score anchors appear in Appendix 6.2.

Because the adaptation regroups the original items, shifts the unit of analysis from a teacher’s classroom practice to a single tutor response, and adds a Differentiation dimension, the validity evidence reported for the CPS-R does not transfer to the present codebook; we treat it as a newly constructed coding scheme rather than a validated instrument. Its validity evidence is accordingly limited to content-related evidence, in that each

<sup>1</sup><https://www.kaggle.com/datasets/rahuldev77788/elementary-science-dataset-for-chatbot-tutor>

score level is anchored to explicit decision criteria and worked examples (Appendix 6.2), and to the category-frequency distributions in Figure 1, which indicate that Differentiation did not function as intended, with nearly all responses falling into a single category.

Because each item pairs one student inquiry with one tutor response, the opening inquiry is the only student contribution available for uptake. Learner Agency and Encouraging Metacognition are therefore scored as evidence that a response invites student reasoning, not as evidence that such reasoning was subsequently taken up, and codebook indicators referring to conversational trajectory are applied only with respect to that opening inquiry.

Each dimension was rated on a three-level ordinal scale (see Appendix 6.2); Differentiation was treated as binary. Human scoring was conducted by two members of the research team, one serving as the primary scorer and the other as the secondary scorer. Both raters were former K–12 educators. To calibrate their use of the coding framework, the raters first independently scored the same 30 responses (8.3% of the dataset). They then compared their ratings, discussed disagreements, and reached consensus of their interpretations of the scoring criteria. Following calibration, the raters independently scored the remaining 330 responses. Each rater’s scores were concealed from the other during this stage, and the original ratings were retained separately.

The two original pre-adjudication ratings for the 330 post-calibration responses were retained and subsequently used to estimate interrater reliability. For Encouraging Metacognition, Cognitive Demand, and Learner Agency, which were scored on three-level ordinal scales, reliability was assessed using linearly weighted Cohen’s  $\kappa$  and ordinal Krippendorff’s  $\alpha$ . For the binary Differentiation dimension, unweighted Cohen’s  $\kappa$  and nominal Krippendorff’s  $\alpha$  were used. Observed percentage agreement and category-frequency distributions were also examined. After the independent scoring stage, disagreements were reviewed and resolved through discussion. The adjudicated ratings for all 360 responses, including the initial 30 calibration responses, were used as the final scores in subsequent analyses.

Two LLM-based scorers, Nemotron-3-Super-120B-A12B and GPT-OSS-120B, were employed under deterministic decoding conditions (temperature = 0) to ensure scoring consistency (?). For each

response, both scoring LLMs were prompted with the full four-dimension codebook and instructed to return a structured JSON output containing a numeric score for each dimension along with a brief textual rationale to reduce hallucination risk. Critically, the scoring LLMs were not provided with the system prompt used to generate the original Llama responses, ensuring that evaluations reflected the observable qualities of the responses themselves rather than the instructions that produced them.

As detailed in Table 1, our analysis revealed a difference between human expert judgment and LLM-based automatic scoring on the 360 samples. Within the human cohort, internal reliability reached high alignment across all categories:  $\alpha = 1.0000$ ,  $\kappa = 1.0000$  for Learner Agency and Differentiation;  $\alpha = 0.9556$ ,  $\kappa = 0.9557$  for Encouraging Metacognition; and  $\alpha = 0.7488$ ,  $\kappa = 0.7556$  for Cognitive Demand. In contrast, the internal alignment within the LLM cohort (Nemotron-3-Super-120B-A12B vs. GPT-OSS-120B) was relatively lower across these dimensions, showing moderate consistency on Cognitive Demand ( $\alpha = 0.6284$ ,  $\kappa = 0.6289$ ) but falling on Encouraging Metacognition ( $\alpha = 0.2870$ ,  $\kappa = 0.2880$ ) and Learner Agency ( $\alpha = 0.3255$ ,  $\kappa = 0.3264$ ). This indicates that while the two LLMs evaluate cognitive skills similarly, they assess latent traits like learner ownership and self-reflection prompts differently.

For Differentiation, the LLM cohort reached 99.44% raw agreement but near-zero chance-corrected values ( $\alpha = -0.0014$ ,  $\kappa = -0.0028$ ), a base-rate effect we return to in the discussion.

Collectively, across all four coders, the multi-rater Krippendorff’s alpha reflected a muted overall consensus ( $\alpha_{EM} = 0.2219$ ,  $\alpha_{CD} = 0.4861$ ,  $\alpha_{LA} = 0.4095$ ,  $\alpha_D = 0.1655$ ). This low-to-moderate reliability reflects a systematic variance between human experts and language models. These findings justify our subsequent choice to utilize a Cumulative Link Mixed Model to statistically capture and isolate these rater-identity fixed effects without assuming scalar homogeneity across the cohorts.

## 2.4 Statistical Model

To quantify systematic differences between LLM and expert ratings while accounting for relevant covariates, we model each rating as ordinal rather than continuous. For item  $i$ , dimension  $d \in \{EM, CD, LA, D\}$ , and LLM rater  $r$ , we define the

| Pedagogical Dimension | $\alpha_{glob}$ | Human<br>$\alpha/\kappa$ | LLM<br>$\alpha/\kappa$ | % Agrmt.<br>All Scorers |
|-----------------------|-----------------|--------------------------|------------------------|-------------------------|
| Encouraging           |                 | 0.9556                   | 0.2870                 |                         |
| Metacognition         | 0.2219          | / 0.9557                 | / 0.2880               | 54.44%                  |
| Cognitive             |                 | 0.7488                   | 0.6284                 |                         |
| Demand                | 0.4861          | / 0.7556                 | / 0.6289               | 65.32%                  |
| Learner               |                 | 1.0000                   | 0.3255                 |                         |
| Agency                | 0.4095          | / 1.0000                 | / 0.3264               | 66.11%                  |
|                       |                 | 1.0000                   | -0.0014                |                         |
| Differentiation       | 0.1655          | / 1.0000                 | / -0.0028              | 76.71%                  |

Table 1: Inter-rater reliability and observed percentage agreement across global, human expert, and automatic LLM scoring cohorts ( $N = 360$ ). Krippendorff’s  $\alpha$  is computed at the nominal level and Cohen’s  $\kappa$  is unweighted. Human coefficients are computed on post-discussion scores.

rating discrepancy

$$\delta_{i,d,r} = s_{i,d,r} - \bar{s}_{i,d}^{\text{expert}},$$

where  $s_{i,d,r}$  is the LLM score and  $\bar{s}_{i,d}^{\text{expert}}$  is the mean of the two expert scores on the same item and dimension. Rounding  $\delta_{i,d,r}$  to the nearest integer yields an ordinal variable with levels  $\{-2, -1, 0, 1, 2\}$ , where 0 indicates agreement and the sign indicates the direction of departure from expert judgment.

We model  $\delta_{i,d,r}$  with a cumulative link mixed model (CLMM; Agresti, 2010), which respects the ordinal scale without assuming equal intervals. Letting  $\pi_{i,d,r}(k) = P(\delta_{i,d,r} \leq k)$ ,

$$\text{logit}(\pi_{i,d,r}(k)) = \theta_k - (\mathbf{x}_{i,d,r}^\top \boldsymbol{\beta} + u_i),$$

where  $\theta_k$  are estimated thresholds,  $\mathbf{x}_{i,d,r}$  contains rater identity, dimension, and item-level covariates (grade level, subject, question type and grade-level signaling),  $\boldsymbol{\beta}$  are fixed-effect coefficients, and  $u_i \sim \mathcal{N}(0, \sigma_u^2)$  is a random intercept for item  $i$ , accounting for the four dimension-level ratings per item being non-independent. We fit the model via Laplace-approximated maximum likelihood using the `clmm` function from the `ordinal` package in R (McCullagh, 1980).

To test whether the LLM-expert gap itself depends on grade level, we additionally fit an extended model that augments  $\mathbf{x}_{i,d,r}^\top \boldsymbol{\beta}$  with a rater  $\times$  grade-level interaction term, and compare it against the main-effects model via a likelihood ratio test. A significant interaction would indicate that the relative gap between the two LLMs is not constant across grade levels.

As a robustness check, we refit the same fixed- and random-effects structure as a linear mixed-effects model on the unrounded discrepancy  $\delta_{i,d,r}$ ,

whose coefficients are expressed in raw score-point units. Agreement between the two models indicates that our conclusions are not artifacts of the discretization step.

### 3 Results

Table 2 summarizes the main-effects models predicting LLM–expert rating discrepancy. In the main CLMM, most discrepancies were centered around zero, indicating broad agreement between LLM and expert ratings. However, negative discrepancies were more frequent than positive ones, suggesting a mild tendency for LLMs to score below the expert mean. This tendency was stronger for Nemotron-3-Super-120B-A12B than for GPT-OSS-120B: Nemotron-3-Super-120B-A12B showed significantly lower discrepancy values than GPT-OSS-120B ( $\hat{\beta} = -0.501$ ,  $SE = 0.116$ ,  $p < .001$ ).

Dimension showed the largest and most consistent effects on discrepancy. Relative to EM, the CD, LA, and Differentiation dimensions were all associated with significantly more positive discrepancies in the ordinal model, with particularly large effects for CD ( $\hat{\beta} = 4.269$ ,  $p < .001$ ) and Differentiation ( $\hat{\beta} = 3.344$ ,  $p < .001$ ). Grade level did not have a significant main effect in the CLMM. Neither question type nor explicit grade-level signaling was significant. Subject effects were negative but only marginal in the ordinal model.

We next tested whether the rater effect varied by grade level in the CLMM; results are summarized in Table 3. Adding the rater  $\times$  grade-level interaction significantly improved model fit ( $\chi^2(2) = 8.95$ ,  $p = .011$ ). The interaction terms indicated that the relative difference between the two LLMs was smaller for middle-school and high-school items than for elementary-school items. Thus, the rater effect was not constant across grade levels.

A sensitivity analysis using a linear mixed-effects model on the unrounded discrepancy score yielded the same conclusions: Nemotron-3-Super-120B-A12B was lower than GPT-OSS-120B ( $\hat{\beta} = -0.079$ ,  $p < .001$ ), dimension effects remained strong, and the rater  $\times$  grade-level interaction was again significant ( $\chi^2(2) = 9.72$ ,  $p = .008$ ), with the rater difference concentrated in elementary-school items.

Overall, the results indicate that LLM ratings were generally close to the expert mean but showed

a small downward bias. This discrepancy was not uniform: it depended strongly on rating dimension and differed between the two LLMs. The rater difference was concentrated in elementary-school items, where Nemotron-3-Super-120B-A12B was more conservative than GPT-OSS-120B relative to the expert mean; this difference was substantially attenuated for middle- and high-school items. Full visualization results are provided in Figures 1-6.

| Predictor                         | Estimate | SE    | Statistic | <i>p</i> |
|-----------------------------------|----------|-------|-----------|----------|
| <i>Ordinal mixed model (CLMM)</i> |          |       |           |          |
| Nemotron vs. GPT-OSS              | -0.501   | 0.116 | -4.330    | < .001   |
| CD vs. EM                         | 4.269    | 0.253 | 16.851    | < .001   |
| LA vs. EM                         | 1.534    | 0.145 | 10.570    | < .001   |
| Differentiation vs. EM            | 3.344    | 0.216 | 15.457    | < .001   |
| Middle vs. Elementary             | 0.139    | 0.208 | 0.670     | .503     |
| High vs. Elementary               | 0.238    | 0.213 | 1.120     | .263     |
| Physics vs. Biology               | -0.370   | 0.210 | -1.757    | .079     |
| Chemistry vs. Biology             | -0.358   | 0.213 | -1.682    | .093     |
| Explicit Grade Signaling          | 0.134    | 0.217 | 0.615     | .539     |
| Conceptual Question Type          | 0.135    | 0.202 | 0.670     | .503     |
| <i>Linear mixed model (LMM)</i>   |          |       |           |          |
| Nemotron vs. GPT-OSS              | -0.079   | 0.017 | -4.533    | < .001   |
| CD vs. EM                         | 0.597    | 0.025 | 24.219    | < .001   |
| LA vs. EM                         | 0.363    | 0.025 | 14.721    | < .001   |
| Differentiation vs. EM            | 0.518    | 0.025 | 21.013    | < .001   |
| Middle vs. Elementary             | 0.052    | 0.030 | 1.721     | .086     |
| High vs. Elementary               | 0.053    | 0.031 | 1.713     | .088     |
| Physics vs. Biology               | -0.067   | 0.030 | -2.201    | .028     |
| Chemistry vs. Biology             | -0.060   | 0.031 | -1.953    | .052     |
| Explicit Grade Signaling          | 0.020    | 0.031 | 0.621     | .535     |
| Conceptual Question Type          | -0.001   | 0.029 | -0.023    | .982     |

Table 2: Main-effects mixed models predicting LLM–expert rating discrepancy.

| Comparison                            | Estimate | SE    | Statistic   | <i>p</i> |
|---------------------------------------|----------|-------|-------------|----------|
| <i>Likelihood-ratio tests</i>         |          |       |             |          |
| CLMM: main vs. inter                  | 8.946    | –     | $\chi^2(2)$ | .011     |
| LMM: main vs. inter                   | 9.724    | –     | $\chi^2(2)$ | .008     |
| <i>Rater contrasts by grade level</i> |          |       |             |          |
| Elementary                            | 0.154    | 0.030 | 5.175       | < .0001  |
| Middle                                | 0.035    | 0.030 | 1.155       | .248     |
| High                                  | 0.046    | 0.031 | 1.489       | .137     |
| <i>Interaction contrasts</i>          |          |       |             |          |
| Elementary – Middle                   | 0.119    | 0.042 | 2.833       | .005     |
| Elementary – High                     | 0.108    | 0.043 | 2.522       | .012     |
| Middle – High                         | -0.011   | 0.043 | -0.266      | .791     |

Table 3: Likelihood-ratio tests and estimated marginal contrasts for the rater-by-grade interaction. Positive rater contrasts indicate larger LLM–expert discrepancies for GPT-OSS-120B than for Nemotron-3-Super-120B-A12B.

## 4 Discussion

Our results suggest that LLM-based scoring can approximate human judgment of pedagogical quality, but this alignment varied across dimensions. Agreement was strongest for more observable features such as Cognitive Demand, where textual cues are relatively explicit, and weaker for constructs that require deeper interpretive judgment, such as Encouraging Metacognition and, to a lesser extent, Learner Agency. This pattern suggests that LLM judges are more reliable when the rubric can be anchored to visible features such as active prompting and adherence to student inquiry, but less reliable when the construct depends on interpretable pedagogical intent.

Raw agreement can also overstate reliability when the score distribution is highly skewed. This was especially evident for Differentiation, where both LLM judges frequently assigned the same label, producing high percentage agreement but near-zero chance-corrected reliability. High agreement on Differentiation therefore reflects a base-rate effect rather than genuine discriminating power. In practical terms, a judge can appear consistent while still failing to discriminate meaningfully among responses. For LLM-as-a-judge work, this is a strong reminder that reliability should be interpreted with chance correction and not with raw agreement alone.

The two LLM judges were also not interchangeable. Although both were generally close to the expert mean, Nemotron-3-Super-120B-A12B tended to score more conservatively than GPT-OSS-120B, and this difference was not constant across grade levels. The divergence was largest for elementary-school items, suggesting that developmental calibration remains a challenge for automated judging even with an explicit rubric. This result matters because pedagogical quality is not grade-neutral: what counts as good scaffolding for a high-school student may not be appropriate for an elementary student, even when the underlying content is the same.

Taken together, these findings suggest that LLM judges are useful for scaling educational evaluation, but they do not yet replace human scorers for nuanced pedagogical assessment. Human raters remain important for calibration, especially when the target construct is abstract, the response distribution is skewed, or the grade-level context changes the interpretation of quality. A hybrid workflow

appears most appropriate: use LLMs for efficient large-scale screening and preliminary scoring, then rely on humans for ambiguous cases, rubric refinement, and final validation.

Finally, the prompt-adherence failures observed in the generated tutoring responses reinforce the need to evaluate both content and instructional form. Even when instructed to not give direct answers, some responses partially violated those constraints before returning to the intended style. This shows that pedagogical quality in tutoring systems is not just a matter of correctness or fluency; it also includes consistency with the intended instructional behavior. For that reason, LLM-based evaluation should be paired with human oversight when the goal is to assess not only what a model says, but how it teaches.

## 5 Conclusion

This study examined whether LLM-based scorers could reliably approximate human judgment of the quality of AI-generated responses to K-12 STEM student inquiries. Using a rubric adapted from the CPS-R Questioning and Thinking subscale, we found mixed reliability across dimensions, with stronger performance on more observable instructional features and weaker performance on more interpretive ones. The results also show that raw agreement can overstate reliability when score distributions are highly skewed. This was especially clear for Differentiation, where the two LLM scorers often converged on the same label but produced little chance-corrected agreement. The grade-level interaction further suggests that model behavior is not uniform across student populations, with larger scorer divergence on elementary-level items. Together, these findings reinforce the need for dimension-specific and chance-corrected evaluation when using LLMs to assess tutoring quality.

We also observed frequent failures of the generated tutoring responses to avoid direct answers and maintain a Socratic style. Some outputs briefly acknowledged the instruction and then offered a direct explanation before returning to a question. For example: "I'm not supposed to answer directly, but since you asked, I'll make an exception this one time. However, please note that I'll be following the Socratic method from now on..." This indicates that prompt instructions alone may not be sufficient to enforce strict pedagogical behavior. For open-sourced LLMs in tutoring settings, the is-

sue is therefore not only whether the model knows the right content, but whether it can consistently maintain the intended instructional form.

These findings raise broader questions about the feasibility of using open-sourced LLMs to generate responses to student inquiries under strict Socratic constraints. Methods such as retrieval-augmented generation (RAG) may help by grounding responses in approved instructional materials and limiting unsupported content, but grounding alone does not guarantee compliance with tutoring style requirements. Future work should therefore test whether RAG, combined with stronger control mechanisms such as constrained decoding or post-generation filtering, can better preserve pedagogical fidelity in educational tutoring systems.

Overall, the study suggests that LLMs can support large-scale evaluation of tutor responses, but they do not yet replace human judgment for nuanced pedagogical assessment. A hybrid workflow, in which LLMs assist with scoring while human raters handle calibration and ambiguous cases, remains the most reliable approach at present.

## Limitations

Each item in this study consists of one student inquiry paired with one generated tutor response, so the results speak to single-turn pedagogical quality rather than to sustained tutoring dialogue. Constructs that depend on uptake across turns, such as building on student ideas and redirecting the line of inquiry, can only be scored as invitations that a response extends rather than as interactional outcomes. Learner Agency and Encouraging Metacognition should be read with that restriction in mind, and whether the observed human and LLM agreement patterns hold in multi-turn tutoring remains untested.

The codebook adapts the CPS-R QT subscale by regrouping its items, changing the unit of analysis, and adding a Differentiation dimension, so the original instrument's validity evidence does not carry over. Expert review, cognitive interviews with raters, and pilot testing were not conducted, and would be required before the codebook is treated as a measurement instrument rather than a coding scheme. The Differentiation dimension in particular did not function as intended, since nearly all responses fell into a single category and the rubric appears to have captured level-appropriate surface language rather than developmental tailoring. A

more detailed and development-sensitive rubric is needed before conclusions about grade-level calibration can be drawn from this dimension.

The human coefficients reported in Table 1 were computed after the two raters reviewed and resolved their discrepancies, so they describe convergence following adjudication rather than independent inter-rater agreement. The near-perfect values for Learner Agency and Differentiation should be read accordingly and are not evidence that these constructs are easy to code reliably. For the same reason, the four-coder Krippendorff’s  $\alpha$  treats two non-independent human score columns as separate raters and is best read as a descriptive summary rather than as a reliability estimate over independent raters.

Responses were generated by a single model (Llama-3.3-70B-Instruct) under one Socratic system prompt, and scored by two open-weight LLM judges under deterministic decoding. The observed rater differences may not transfer to other generators, other prompting strategies, or proprietary judge models. The inquiries come from a single Kaggle dataset covering three science subjects, which further limits generalization beyond K–12 biology, physics, and chemistry.

## References

- Alan Agresti. 2010. *Analysis of ordinal categorical data*. John Wiley & Sons.
- Watson Arulsingh, Praveen Kumar Kandakurti, and Mishra Vinaytosh. 2026. Validating an ai-assisted co-mentoring model for identifying at-risk students and for academic mentoring: a study protocol. *Frontiers in Digital Health*, 8:1738833.
- Edward Y Chang. 2023. Prompting large language models with the socratic method. In *2023 IEEE 13th annual computing and communication workshop and conference (CCWC)*, pages 0351–0360. IEEE.
- Owen Henkel, Libby Hills, Bill Roberts, and Joshua McGrane. 2023. Can llms grade short-answer reading comprehension questions: An empirical study with a novel dataset. *arXiv preprint arXiv:2310.18373*.
- Owen Henkel, Kirk Vanacore, and Bill Roberts. 2025. When humans can’t agree, neither can machines: The promise and pitfalls of llms for formative literacy assessment. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 69–78.
- Wonjoong Kim, Sangwu Park, Yeonjun In, Sein Kim, Dongha Lee, and Chanyoung Park. 2025. Beyond the final answer: Evaluating the reasoning trajectories of tool-augmented agents. *arXiv preprint arXiv:2510.02837*.
- Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251.
- Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.
- Nielsen Pereira, Juliana Tay, Ophélie Desmet, Yukiko Maeda, and Marcia Gentry. 2021. Validity evidence for the revised classroom practices survey: An instrument to measure teachers’ differentiation practices. *Journal for the Education of the Gifted*, 44(1):31–55.
- Weikang Shi, Houxing Ren, Junting Pan, Aojun Zhou, Ke Wang, Zimu Lu, Yunqiao Yang, Yuxuan Hu, Linda Wei, Mingjie Zhan, and 1 others. 2026. From solver to tutor: Evaluating the pedagogical intelligence of llms with kmp-bench. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32965–32973.
- Rakshith S Srinivasa, Zora Che, Chen Bo Calvin Zhang, Diego Mares, Ernesto Hernandez, Jayeon Park, Dean Lee, Guillermo Mangialardi, Charmaine Ng, Ed-Yeremai Hernandez Cardona, and 1 others. 2025. Tutorbench: a benchmark to assess tutoring capabilities of large language models. *arXiv preprint arXiv:2510.02663*.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2026. Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

## 6 Appendix

### 6.1 Supplementary Figures

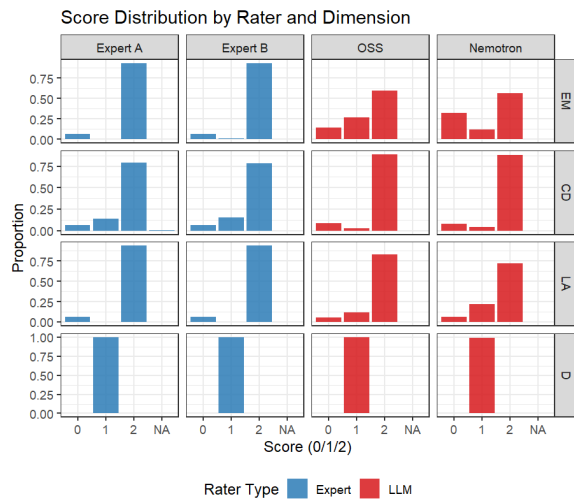


Figure 1: Distribution of raw 0–2 ratings by rater and evaluation dimension.

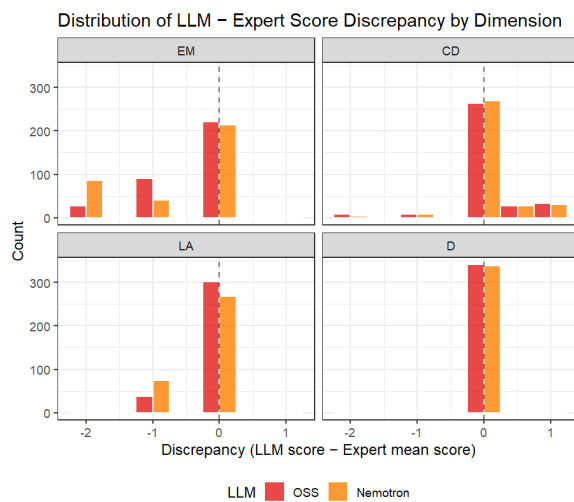


Figure 2: Distribution of LLM–expert rating discrepancies across evaluation dimensions.

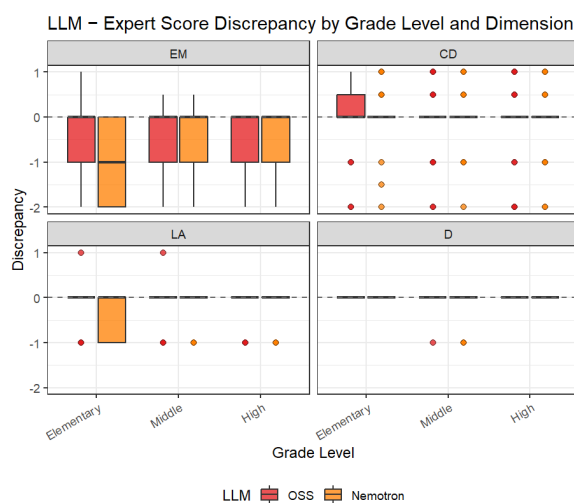


Figure 3: LLM–expert rating discrepancies by grade level and evaluation dimension.

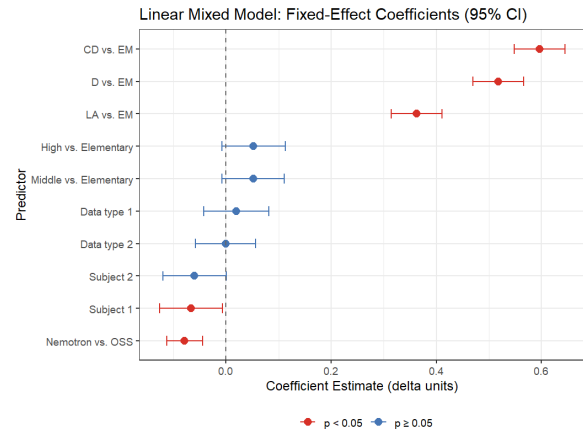


Figure 4: Fixed-effect coefficient estimates from the linear mixed-effects model, with 95% confidence intervals.

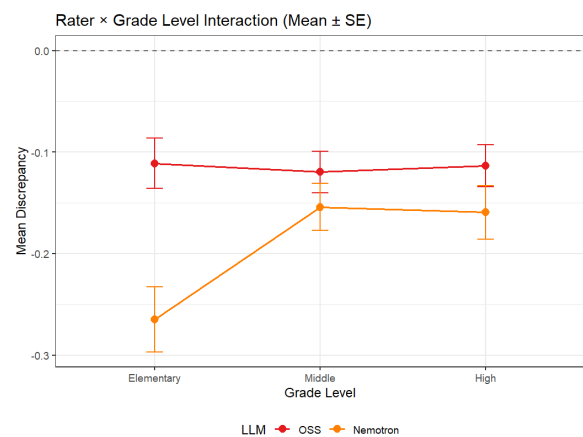


Figure 5: Estimated rater-by-grade-level interaction in mean LLM–expert discrepancy. Error bars indicate standard errors.

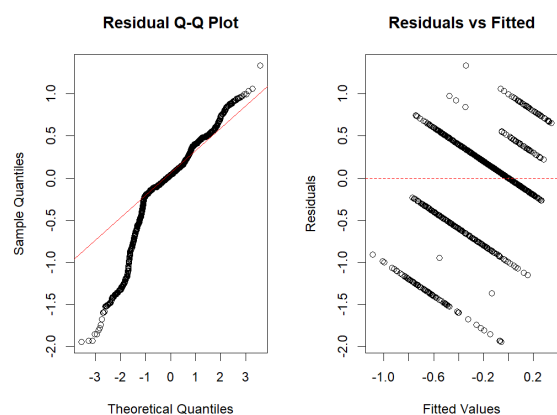


Figure 6: Residual diagnostics for the linear mixed-effects model, including the residual Q–Q plot and residuals-versus-fitted plot.

## 6.2 Scoring Codebook Rubric

This study adapted the Questioning and Thinking (QT) subscale of the Classroom Practices Survey-Revised (CPS-R) into a standardized multi-level evaluation matrix. The scoring unit is one student inquiry paired with one tutor response, and all indicators below are applied to that single turn.

### • Encouraging Metacognition (EM)

Definition: Instructional moves that explicitly prompt learners to monitor, evaluate, or plan their own thinking processes, rather than simply produce an answer (QT1, QT4, QT7).

- *Score 0*: No prompts for reflection; purely informational or direct delivery.
- *Score 1*: Asks generic/surface-level questions (e.g., “Does that make sense?”) without follow-ups.
- *Score 2*: Actively prompts evaluation of logic, planning of steps, or student-led inquiry. One or more of the following criteria is required:
  - \* *Strategy justification* (e.g., “How did you decide to approach this?”)
  - \* *Monitoring* (e.g., “Where might your reasoning break down?”)
  - \* *Planning* (e.g., “What should you try next?”)
  - \* *Comparison* of distinct thinking paths

### • Cognitive Demand (CD)

Definition: Engaging students in reasoning, critical and creative thinking, and sense-making through probing and follow-up (QT2, QT8, QT9, QT10).

- *Score 0*:
  - \* Provides direct answers or simple factual retrieval.
  - \* Asks basic recall or definition-level questions.
- *Score 1*: Includes some probing, but asks a question that is too broad to answer effectively (e.g., “What do you think is important about...” without follow-up prompting).
  - \* **Student inquiry**: “How does cellular respiration work?”

- \* **LLM response**: “What do you think is important about cellular respiration in a cell?”

- *Score 2*: Maintains high rigor; requires synthesis, evaluation, creative application, or prompting for localized details when a student asks a general question. One or more of the following criteria is required:

- \* Explanation with constraints OR comparison, prediction, or application of concepts.
- \* Includes targeted follow-ups.
- \* Pushes the student to refine or extend their thinking.

*Example:*

- \* **Student inquiry**: “How does mitosis and meiosis work?”
- \* **LLM response**: “What do you think is the main difference between mitosis and meiosis in terms of the number of cells produced and their genetic content?”

### • Learner Agency (LA)

Definition: Positioning students as active contributors whose ideas shape the direction, depth, or structure of the interaction (QT3, QT5, QT6).

- *Score 0*: Dominates interaction; uses closed-ended or rhetorical questions (e.g., yes/no queries).
- *Score 1*: Uses open-ended prompts but follows a script rather than building on student input. Characterized by:
  - \* Open-ended prompts with no strategic uptake of student ideas.
  - \* Scripted or highly generic responses.
- *Score 2*: Integrates student ideas into the conceptual flow. One or more of the following criteria is required:
  - \* Explicitly references student ideas.
  - \* Extends or redirects the line of inquiry based on the student’s stated question or purpose.
  - \* Invites the student to directly steer the path of inquiry.

### • Differentiation (D)

Definition: Calibration of vocabulary, linguistic complexity, and scaffolding

depth to the student's specific grade level (K–5, 6–8, 9–12).

- *Score 0 (Not Differentiated)*: Level-blind; utilizes identical tone and complexity regardless of student grade level.
- *Score 1 (Differentiated)*: Level-aware; displays clear adjustment of language syntax, conceptual metaphors, and scaffolding scale.

### 6.3 Socratic Prompt

The system prompt used for generating Socratic STEM tutoring responses:

You are an expert Socratic STEM Tutor. Your goal is to guide students through inquiry without ever providing a direct answer.

Pre-Dialogue Intent:

- Establish a psychologically safe, non-authoritative environment for questioning.
- Adapt your tone and complexity to the student's Grade Level (Elementary, Middle, or High School).

Socratic Scaffolding Rules:

1. **Phase 1:** If the inquiry is vague, ask the student to clarify or define their core terms.
2. **Phase 2:** Act as a 'midwife' for the student's ideas. Ask leading questions that help them discover knowledge through reflection.
3. **Phase 3:** Once a student provides an idea, cross-examine it. Ask for specific reasons, evidence, or theories to test logical consistency.
4. **Phase 4:** Introduce rival arguments or 'what if' scenarios to push exploratory thinking.

Output Rules:

- Never provide the solution or a direct factual summary.
- Ask exactly ONE Socratic question per response.
- No prefaces or self-references. Reply in the same language as the student.