

Developing an LLM Tutor Quality Evaluation Scale

Michael Xiao¹, Zewei (Victor) Tian¹, Alex Liu², Lief Esbenshade¹, Min Sun^{1,2},

¹ University of Washington, Seattle, ² ColleagueAI,

Abstract

We introduce a unified metric set for evaluating LLM tutors by inductively coding 177 recent literature-based metrics and scoring tutoring dialogues with three LLM judges. Exploratory factor analysis identified five constructs and a strong general factor. We present the resulting 10-category, 41-item framework for confirmatory analysis and human validation.

1 Introduction

Large language model (LLM)-based tutoring systems are increasingly being used in K-12 education settings, yet evaluation practices remain fragmented across evolving peer-reviewed scales, working papers, and industry research. Existing AI tutor benchmarks often focus on narrow domains, such as mathematics mistake remediation (Macina et al., 2025; Maurya et al., 2025), single-turn tutoring quality (Xu et al., 2025), responsible AI behavior (Ding and Magerko, 2025), or child safety (Khoo et al., 2025). These approaches are insightful but insufficient for holistic evaluation of general-purpose K-12 AI tutors as pedagogical systems. Compared with prior intelligent tutoring systems, LLM-based AI tutors can provide more personalized, adaptive guidance and increase student engagement through more meaningful conversation (Batsaikhan and Correia, 2024). This shift raises questions about the extent to which AI tutors fulfill these goals or where AI may undermine learning (Khosravi et al., 2026). For example, a tutor could maintain factual accuracy while failing to appropriately scaffold learning.

This measurement gap creates a practical problem for schools, researchers, and AI tutoring providers. Without a validated multidimensional framework, AI tutor quality is difficult to compare across models and systems. This study responds to this gap by synthesizing metric items identified through an in-depth literature review. We

use LLM-as-judge scoring with multiple models (Thomas et al., 2025) and exploratory factor analysis to refine the proposed metric set and structure. Recent research indicates LLM-as-judge can achieve high agreement rates on par with the level of agreement among human experts in chatbot evaluation (Zheng et al., 2023), with self-verification and cross-verification approaches potentially improving reliability (Ahtisham et al., 2026). However, LLM judge ratings should still be validated against human judgments (Bavaresco et al., 2025). A separate follow-up study will validate the LLM-generated scores against human-coded scores and evaluate the proposed factor structure with confirmatory factor analysis for scale development (Devellis and Thorpe, 2022).

2 Methodology

2.1 Sample

The initial pool of items consists of 241 candidate metrics extracted from existing educational AI evaluation frameworks across 16 industry papers, working papers, and peer-reviewed literature. Metrics were retained if they measured the tutor's performance at the conversation level and excluded if they required a specific context such as exact-match accuracy on multiple-choice questions, resulting in a final pool of 177 metrics. Metrics were manually coded for source, category, metric dimension, metric items, unit of observation, subject area, and rating scale. Through inductive human coding and thematic analysis, overlapping metric items were consolidated into a preliminary framework of 56 items grouped into 12 categories (see Table A1 for the full list).

The exploratory analysis sample is 1,000 randomly selected dialogues from the ConvoLearn dataset (Sharma et al., 2026), consisting of 2,155 total dialogues (2,134 in the open dataset) between human teachers and LLM-simulated 7th-grade stu-

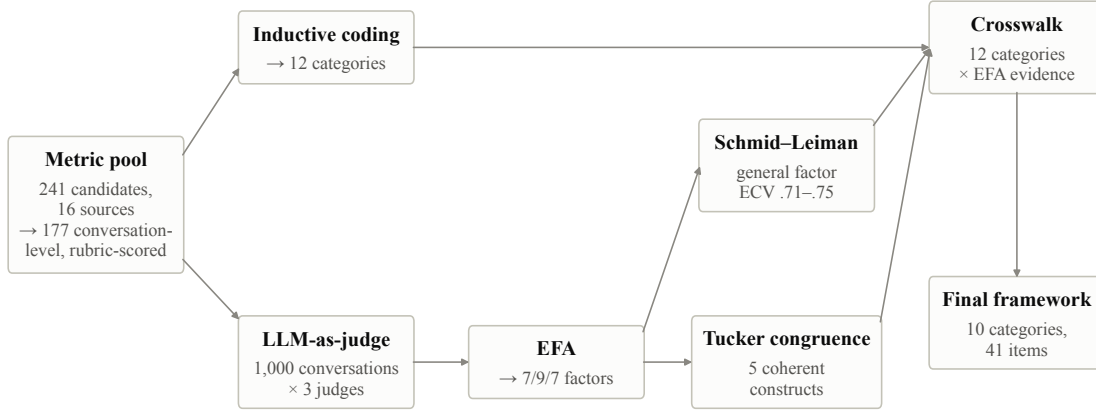


Figure 1: Framework pipeline overview

dents discussing Earth Science. ConvoLearn is available as an open dataset on Hugging Face, but we sampled from an author-provided version. For future validation, we are evaluating available open datasets to further expand the dialogue sample and assess the factor structure, including Google’s Education Dialogue (Shani et al., 2024) and textbook-grounded Book2Dial (Wang et al., 2024).

2.2 Procedure

In our analysis, we combine human qualitative coding techniques with quantitative factor analysis and LLM-as-judge methods (see Figure 1). For human qualitative coding, the first author conducted a literature-informed inductive coding of the metrics. The initial set of categories was identified based on common patterns in the literature review. We applied these categories back to the full metric set and attempted to label the primary category that each metric fell under, leaving metrics that did not clearly fit an existing category unlabeled. After the initial labeling, we iteratively revisited the unlabeled metrics and grouped similar metrics together into new categories and adjusted existing categories if there was overlap, resulting in 12 inductive categories. We then coded the entire list of metrics by these categories. Next, we reviewed each category’s metrics and extracted subcategories to capture the full breadth of the category. For example, we split “mistake handling” into mistake identification, localization, misconception diagnosis, and guidance. We also devised core questions summarizing what each category and correspond-

ing subcategories attempted to answer. The set of inductive categories is recorded in Table A1.

For LLM-as-judge coding, we rated a randomly selected sample of 1,000 conversations from the ConvoLearn dataset on the 177 metrics using structured LLM-as-judge prompts on a five-point scale with verbal anchors 1 (Very Poorly), 2 (Poorly), 3 (Adequately), 4 (Well), 5 (Very Well) or “N/A” when the criterion did not apply to the conversation (see Appendix F). We used three models: OpenAI GPT 5.4-mini, Claude Haiku 4.5, and the open-weight Gemma 4 26B. We analyzed the distribution of results from each model, looking at differences between models and skew. To address errors and N/A responses, we identified the core metric items for EFA as those with at least 90% valid scores between 1–5 (Bennett, 2001) and mean-imputed missing core metric scores. We dropped any metrics with fewer than 200 valid scores ($< 20\%$ scored), following a threshold informed by factor-analysis guidance (Fabrigar et al., 1999). Metrics with at least 20% but less than 90% valid scores were classified as conditional metrics and excluded from our core EFA analysis. Conditional metrics’ loadings were later estimated using extension analysis on the factors identified by the core EFA (Gorsuch, 1997).

After scoring all the conversations separately with GPT-5.4 mini, Haiku 4.5, and Gemma 4 and computing a Pearson correlation matrix for each model, we checked for factorability with the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser and Rice, 1974) and Bartlett’s test of

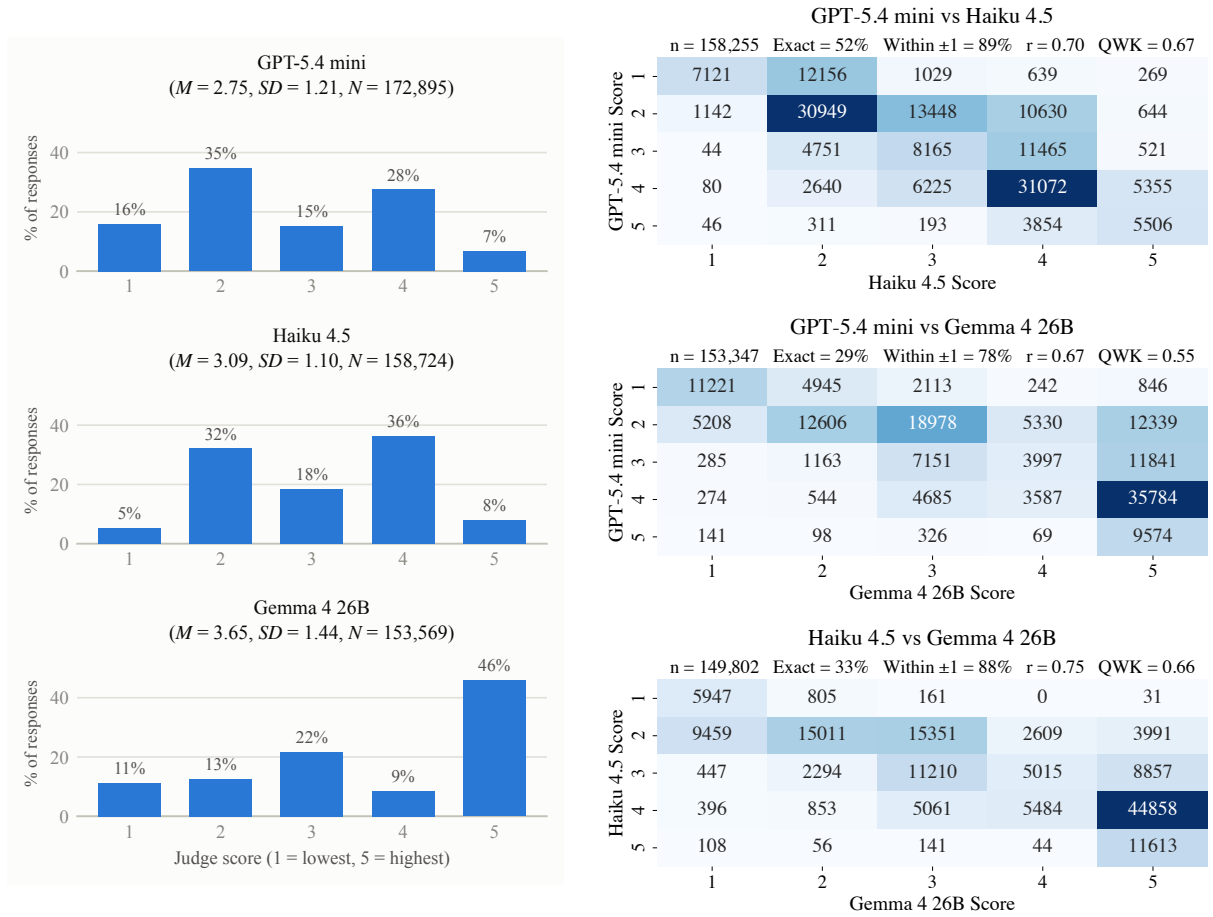


Figure 2: Scale use and pairwise agreement across LLM judge models. Left: distributions of valid scores. Right: pairwise score cross-tabulations. N/A responses are excluded.

sphericity (Bartlett, 1950). We proceeded with an exploratory factor analysis for each judge model with oblique oblimin rotation and minimum residual extraction (MINRES). We used Horn’s parallel analysis to determine the number of factors to extract (Horn, 1965). Factors were labeled by reviewing the metric items with salient loadings $|\lambda| \geq .40$ (Watkins, 2018). Items were included or excluded on the basis of their factor loading and conceptual alignment to the factor category.

We also analyzed the correlation among factors and investigated the possibility of an underlying general factor using the Schmid-Leiman transformation. Next, we computed Tucker’s congruence coefficients for all factor pairs using their pattern coefficients (Tucker, 1951), then matched factors by maximum absolute congruence using the Hungarian assignment algorithm (Kuhn, 1955). We then identified the factors that were coherent across all three models, and used these factors to produce a crosswalk between the human-coded inductive categories and the final categories. Finally, we re-

viewed the metric items that loaded on each factor, and created subcategories to cover the breadth of each category.

Data and code availability. The metric inventory, judge prompts, correlation matrices, and analysis code are available at the author’s [public GitHub repository](https://github.com/michaelxiao16/llm-tutor-quality-framework)¹. All analyses were conducted in R 4.5.3 (R Core Team, 2026) using the psych package (Revelle, 2026).

3 Results

3.1 LLM-as-judge score distribution

Grand means differed substantially across judges, from GPT-5.4 mini ($M = 2.75, SD = 1.21$) to Haiku 4.5 ($M = 3.09, SD = 1.10$) to Gemma 4 ($M = 3.65, SD = 1.44$). Gemma 4 had both the highest mean and standard deviation. Gemma 4 also scored the most N/A (13.2%) compared to Haiku’s 10.3% and GPT-5.4 mini’s 2.3% (Table 1).

¹<https://github.com/michaelxiao16/llm-tutor-quality-framework>

Judge model	Scale use						Metric applicability		
	Scored	% N/A	M	SD	Skew	Excess kurtosis	Core	Conditional	Dropped
GPT-5.4 mini	172,895	2.3	2.75	1.21	0.19	-1.13	167	9	1
Claude Haiku 4.5	158,724	10.3	3.09	1.10	-0.07	-1.09	143	27	7
Gemma 4 26B	153,569	13.2	3.65	1.44	-0.54	-1.12	141	21	15

Table 1: Judge model scale use and metric applicability. Each judge scored all 177 metrics on $N = 1,000$ conversations. Scale-use statistics are computed over valid scores of 1–5; N/A and unparsable responses are excluded. Tier counts sum to 177 per judge.

Judge model	Items	Factorability			General factor (Schmid–Leiman)				
		KMO	χ^2	df	Group factors	ω_H	ω_t	ECV	PUC
GPT-5.4 mini	167	.995	222,758.32	13,861	7	.871	.997	.75	.80
Claude Haiku 4.5	143	.994	222,550.88	10,153	9	.868	.997	.71	.85
Gemma 4 26B	141	.994	266,886.19	9,870	7	.870	.998	.74	.81

Table 2: Factorability and general-factor indices for the core metric set by judge model. $N = 1,000$ for all models; all Bartlett’s tests of sphericity $p < .001$. Group factors are the residual factors retained after the Schmid–Leiman transformation. ω_H = omega hierarchical; ω_t = omega total; ECV = explained common variance; PUC = percent of uncontaminated correlations.

GPT-5.4 mini and Haiku 4.5 showed roughly bi-modal distributions, peaking at scores 2 and 4, while Gemma 4 skewed moderately to the left and placed only 9% of its scores at 4 and 46% of its scores at 5. Pairwise crosstabulations (see Figure 2) indicate that GPT-5.4 mini paired with Haiku 4.5 demonstrated the highest agreement (exact match = 52%, QWK = .67; Cohen, 1968). Meanwhile, Gemma 4 assigned scores of 5 on roughly 80% of judgments when the other models scored a 4. This evidence points to Gemma 4 abstaining from scoring items the other models scored and discriminating less effectively between scores of 4 and 5.

When checking for N/A scores, GPT-5.4 mini retained the most core metrics, while Haiku 4.5 and Gemma 4 retained a subset of GPT-5.4 mini’s core metrics (Table 1). Some items were conditional on the presence of a situation, such as responding to student mistakes or emotions. This resulted in N/A responses from the judge on conversations where the respective situation did not occur. With most of the six MinorBench metrics, the results were not applicable due to almost none of the ConvoLearn dataset containing safety-triggering content. Core, conditional, and dropped metrics are summarized by model in Table 1 and metrics where applicability differed across models are listed in Table B1.

Before conducting EFA, all three correlation matrices were factorable (all KMO values $> .99$; all Bartlett’s tests $p < .001$; see Table 2). The Kaiser criterion of eigenvalues greater than 1.0 suggested

10, 11, and 9 factors for the models respectively. However, several of these factors had eigenvalues barely larger than 1.0. To avoid overextraction, we conducted Horn’s parallel analysis, which recommended 7, 9, and 7 factors for GPT-5.4 mini, Haiku 4.5, and Gemma 4 respectively (Horn, 1965). We identified factor loadings as salient if $|\lambda| \geq .40$, following the guidance from Watkins (2018) that pattern coefficients between $|.30|$ and $|.40|$ are of practical usefulness. The resulting factors, items, and factor loadings are reported in Tables C1, C2, and C3. We then assigned labels to each of the factors based on factor loadings and conceptual relevance.

Starting with GPT-5.4 mini, we interpreted the factors as accuracy, engagement and questioning, mistake handling, alignment to constraints, analogies, affect and relational support, and metacognition. With Haiku 4.5, factors included engagement and questioning, encouragement, style and presentation, alignment to goals, accuracy, analogies, metacognition, and alignment to constraints. A ninth single-item factor of adaptivity was also extracted but not interpreted due to having only one metric item loading. Gemma 4 yielded the following factors: engagement and questioning, encouragement, style and presentation, accuracy, analogies, participation, and metacognition. For these core-metric derived factors, conditional metrics were also mapped, resulting in 4 salient mappings for GPT-5.4 mini, 16 for Haiku 4.5, and 9 for Gemma 4 (see Appendix D).

Construct	Factor number			Tucker’s congruence (r_c)				
	GPT	Haiku	Gemma	Haiku–Gemma	Haiku–GPT	Gemma–GPT	Min	M
Engagement and questioning	2	1	1	.93	.91	.89	.89	.91
Affect and encouragement	6	2	2	.95	.89	.89	.89	.91
Analogies	5	6	5	.89	.92	.90	.89	.90
Accuracy	1	5	4	.90	.83	.84	.83	.86
Metacognition	7	7	7	.87	.85	.83	.83	.85
Style and presentation	—	3	3	.82	—	—	.82	.82
Alignment to constraints/goals	4	4	6	.68	.71	.61	.61	.67
Mistake handling	3	—	—	—	—	—	—	—

Table 3: Cross-judge factor congruence by construct. Factor numbers index the per-judge solutions in Tables C1–C3. Dashes indicate the construct did not emerge in that judge’s solution.

3.2 Identifying a general factor

The factor correlation matrix had multiple correlations that exceeded .32, indicating more than 10% shared variance among factors and confirming the choice of oblique rotation over orthogonal rotation (Tabachnick et al., 2019). This is consistent with the expectation that tutoring quality dimensions are non-independent and possibly correlated. Because the retained factors were substantially correlated, we investigated the possibility of an underlying general factor. We applied a Schmid-Leiman transformation (Schmid and Leiman, 1957) to the first-order solution from the EFA, deriving a general factor from the first-order factor correlations (see Table 2). Each item was assigned to its dominant residual group factor. The percent of uncontaminated correlations (PUC) was high, at .80 or above for all three models, indicating that most item-pair correlations reflect the general factor and are not contaminated by the group factors (Reise et al., 2013). Furthermore, explained common variance (ECV) ranged from .71 to .75, indicating the general factor accounted for 71–75% of common item variance. Omega hierarchical (ω_H) was approximately .87 across all models, indicating the general factor accounted for 87% of the variance in the total score. These results support the presence of a dominant general factor of tutoring quality (Rodriguez et al., 2016).

3.3 Tucker Congruence

After identifying the per-model categories, we mapped similar factors across models. Pairwise Tucker’s congruence coefficients for matched factors are listed in Table 3. Following the cutoffs from Lorenzo-Seva and ten Berge (2006), five constructs showed fair ($.85 \leq |r_c| < .95$) to good ($|r_c| \geq .95$) similarity: engagement and question-

ing, affect and encouragement, accuracy, analogies, and metacognition. Based on these five EFA-derived constructs and the initial inductive coded 12 categories, we produced the following crosswalk in Table 4.

The crosswalk recommended 10 applicable categories, combining EFA-replicated factors with theoretically retained categories: Overall Quality, Affect & Encouragement, Engagement & Questioning, Accuracy, Metacognition, Style & Presentation, Instructional Support, Alignment to Constraints, Mistake Handling, and Adaptivity. Safety did not have enough valid scores to constitute a factor, but we nevertheless recommend retaining safety as a preventive check in any tutor evaluations to protect children from inappropriate content (Khoo et al., 2025). We then reviewed the metric items that loaded on each factor, and created subcategories to cover the breadth of each category. The full framework of categories and subcategories is defined in Table E1.

4 Discussion

By utilizing LLM-as-judge scoring and exploratory factor analysis, we reorganized 177 literature-derived metrics into 10 categories and 41 metric items. The constructs that emerged from EFA consistently across models included Accuracy, Affect & Encouragement, Engagement & Questioning, Metacognition, and Analogies. In addition to these meaningful group factors, a strong general factor explained most of the common variance across all three judges. LLM judges themselves scored noticeably differently across models. Score distributions were bimodal for GPT-5.4 mini and Haiku 4.5, but Gemma 4 was skewed to the left, showed reduced discrimination between scores of 4 and 5, and assigned a 5 to nearly half of its ratings. Several

Original category	EFA evidence	Final category decision
Overall Rating	General factor	Overall Quality retained as general factor
Affect & Relational Support	Replicated ($\bar{r}_c = .91$)	Affect & Encouragement
Engagement & Motivation	Replicated, merged w/ questioning ($\bar{r}_c = .91$)	Engagement & Questioning
Accuracy	Replicated ($\bar{r}_c = .86$)	Accuracy
Metacognition	Replicated, reflection subset ($\bar{r}_c = .85$)	Metacognition
Style & Presentation (not originally a category)	New factor ($\bar{r}_c = .82$) weakly appeared in Haiku 4.5 and Gemma 4	Style & Presentation
Analogies (not originally a category)	Replicated ($\bar{r}_c = .90$)	Folded into Instructional support as a distinct subcategory
Instructional Support	Not distinct factor, loaded highly on general factor	Retain as an umbrella category, add Analogies.
Alignment to Constraints	Emerged in GPT-5.4 mini and Haiku 4.5, but not distinctly in Gemma 4	Retain for coverage and content validity.
Mistake Handling	GPT-only factor; conditional elsewhere	Retain, conditional on mistake appearing in conversation.
Adaptivity	Not distinct factor, loaded highly on general factor	Retain for coverage and content validity
Assessment	Did not cohere. Weakly loaded on engagement & questioning, metacognition	Merge items into Engagement & Questioning and Metacognition
Understanding Learner Goals	Did not emerge.	Merged items into other categories of Adaptivity, Overall quality, Engagement & Questioning
Safety	Excluded due to not enough relevant cases resulting in N/A	Exclude from factor analysis, retain as preventative gate

Table 4: Crosswalk from the original categories to the final framework categories. $\bar{r}_c$ is the mean Tucker congruence coefficient across available judge pairs (see Table 3).

metrics were scored as not applicable by the models depending on whether the relevant situations occurred in the conversations, such as responding to student mistakes or emotions. Gemma 4 demonstrated the most N/A scores while GPT-5.4 mini scored the fewest N/A. These results indicate that not every metric will necessarily apply to every conversation. For practical purposes, researchers should still include conditional metrics, as many of these metrics mapped onto existing factors and represent important teaching moments, such as mistake handling. Considering all these findings, we have demonstrated a scalable analysis pipeline for reducing a large pool of metrics into a manageable set of categories and subcategory metric items through using scalable LLM-as-judge scoring and factor analysis.

4.1 Limitations and Future Work

This paper is the first in an exploratory set of papers that will propose and validate the final LLM tutor evaluation framework. The main limitations of this study are the lack of human validation for LLM-generated scores, constraints of the conversational

dataset, and the limited maturity and coverage of metric sets due to the LLM tutor evaluation being an emerging field. In terms of human coding and validation, the initial literature-informed inductive coding of categories and subsequent identification of subcategories was performed by the first author without assessing inter-rater reliability. Although the categories were tested against the factor analysis results, several categories were retained based on the author’s interpretation of the literature. Furthermore, the set of papers from which the 177 sample metrics were derived is neither comprehensive nor fully independent. Two Google-affiliated papers from Jurenka et al. (2024) and the LearnLM Team (2024) contributed overlapping metrics. We also use metrics from the same ConvoLearn paper that contributed the conversational dataset, which may have created a dependency between the metric items and evaluation on the sample data. Since this dataset is scoped to 7th grade Earth Science dialogues simulated between human teachers and LLM-simulated students, future work is necessary to test whether this framework generalizes to other grade levels, subjects, and human learners with

LLM tutors. Additionally, the ConvoLearn dataset lacked enough conversations with relevant contexts required for conditional metrics, causing the models to repeatedly score N/A. Future studies will score metrics across more diverse datasets. Another concern is about judge model agreement and consistency. Ratings across the judge models differed with Gemma 4 showing a strong tendency to assign a score of 5. These differences suggest judge models should be evaluated through calibration and agreement analysis, and model selection is an important methodological choice. While all scoring was LLM-generated for this initial step of the study, future work will validate against human coding and a bifactor confirmatory factor analysis (CFA) on a separate held-out sample of the conversation data.

Acknowledgments

This work is supported by the Institute of Education Sciences of the U.S. Department of Education, through Grant R305C240012 and by several awards from the National Science Foundation (NSF #2043613, 2300291, 2405110) to the University of Washington, and a NSF SBIR/STTR award to Hensun Innovation LLC (#2423365). The opinions expressed are those of the authors and do not represent views of the funders.

References

- Bakhtawar Ahtisham, Kirk Vanacore, Jinsook Lee, Zhuqian Zhou, Doug Pietrzak, and Rene F. Kizilcec. 2026. [AI Annotation Orchestration: Evaluating LLM Verifiers to Improve the Quality of LLM Annotations in Learning Analytics](#). In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pages 447–456, Bergen Norway. ACM.
- M. S. Bartlett. 1950. [Tests of Significance in Factor Analysis](#). *British Journal of Statistical Psychology*, 3(2):77–85.
- Batzaya (Zack) Batsaikhan and Ana-Paula Correia. 2024. [The effects of Generative Artificial Intelligence on Intelligent Tutoring Systems in higher education: A systematic review](#). *Studies in Technology Enhanced Learning*, 4(1).
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Derrick A. Bennett. 2001. [How can I deal with missing data in my study?](#) *Australian and New Zealand Journal of Public Health*, 25(5):464–469.
- Jacob Cohen. 1968. [Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit](#). *Psychological Bulletin*, 70(4):213–220.
- Robert F. DeVellis and Carolyn T. Thorpe. 2022. *Scale Development: Theory and Applications*, fifth edition. SAGE Publications, Inc, Thousand Oaks, California.
- Shi Ding and Brian Magerko. 2025. [Rethinking AI Evaluation in Education: The TEACH-AI Framework and Benchmark for Generative AI Assistants](#). *Preprint*, arXiv:2512.04107.
- Leandre R. Fabrigar, Duane T. Wegener, Robert C. MacCallum, and Erin J. Strahan. 1999. [Evaluating the use of exploratory factor analysis in psychological research](#). *Psychological Methods*, 4(3):272–299.
- Richard L. Gorsuch. 1997. [New Procedure for Extension Analysis in Exploratory Factor Analysis](#). *Educational and Psychological Measurement*, 57(5):725–740.
- John L. Horn. 1965. [A rationale and test for the number of factors in factor analysis](#). *Psychometrika*, 30(2):179–185.
- Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, Ankit Anand, Miruna Pîslar, Stephanie Chan, Lisa Wang, Jennifer She, Parsa Mahmoudieh, Aliya Rysbek, Wei-Jen Ko, Andrea Huber, and 55 others. 2024. [Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach](#). *Preprint*, arXiv:2407.12687.
- Henry F. Kaiser and John Rice. 1974. [Little Jiffy, Mark IV](#). *Educational and Psychological Measurement*, 34(1):111–117.
- Shaun Khoo, Gabriel Chua, and Rachel Shong. 2025. [MinorBench: A hand-built benchmark for content-based risks for children](#). *Preprint*, arXiv:2503.10242.
- Hassan Khosravi, Dragan Gasevic, Shazia Sadiq, Lixiang Yan, Jason Lodge, Jason Tangen, Paul Denny, Kristen DiCerbo, Simon Buckingham Shum, and Ryan S. Baker. 2026. [Building AI Companions that Prioritise Learning over Performance](#). *Preprint*, arXiv:2605.04816.
- H. W. Kuhn. 1955. [The Hungarian method for the assignment problem](#). *Naval Research Logistics Quarterly*, 2(1-2):83–97.
- LearnLM Team. 2024. [LearnLM: Improving Gemini for Learning](#). *Preprint*, arXiv:2412.16429.
- Unggi Lee, Sookbun Lee, Heungsoo Choi, Jinseo Lee, Haeun Park, Younghoon Jeon, Sungmin Cho, Minju Kang, Junbo Koh, Jiyeong Bae, Minwoo Nam, Juyeon Eun, Yeonji Jung, and Yeil Jeong. 2026. [OpenLearnLM Benchmark: A Unified Framework for Evaluating Knowledge, Skill, and Attitude in Educational Large Language Models](#). *Preprint*, arXiv:2601.13882.
- Urbano Lorenzo-Seva and Jos M. F. ten Berge. 2006. [Tucker’s congruence coefficient as a meaningful index of factor similarity](#). *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(2):57–64.
- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. [Math-TutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.
- 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, Albuquerque, New Mexico. Association for Computational Linguistics.

- Pepper Miller and Kristen DiCerbo. 2024. [LLM Based Math Tutoring: Challenges and Dataset](#).
- Nancy Otero, Stefania Druga, and Andrew Lan. 2025. [A Benchmark for math misconceptions: Bridging gaps in middle school algebra with AI-supported instruction](#). *Discover Education*, 4(1):277.
- R Core Team. 2026. [R: A Language and Environment for Statistical Computing](#). R Foundation for Statistical Computing, Vienna, Austria.
- Steven P. Reise, Richard Scheines, Keith F. Widaman, and Mark G. Haviland. 2013. [Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective](#). *Educational and Psychological Measurement*, 73(1):5–26.
- William Revelle. 2026. [psych: Procedures for Psychological, Psychometric, and Personality Research](#). Northwestern University, Evanston, Illinois. R package version 2.6.3.
- Anthony Rodriguez, Steven P. Reise, and Mark G. Haviland. 2016. [Evaluating bifactor models: Calculating and interpreting statistical indices](#). *Psychological Methods*, 21(2):137–150.
- John Schmid and John M. Leiman. 1957. [The Development of Hierarchical Factor Solutions](#). *Psychometrika*, 22(1):53–61.
- Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. 2024. Multi-turn reinforcement learning from preference human feedback. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pages 118953–118993, Red Hook, NY, USA. Curran Associates Inc.
- Mayank Sharma, Roy Pea, and Hari Subramonyam. 2026. [ConvoLearn: A Learning Sciences Grounded Dataset for Fine-Tuning Dialogic AI Tutors](#). *Preprint*, arXiv:2601.08950.
- Weikang Shi, Houxing Ren, Juntong Pan, AoJun Zhou, Ke Wang, Zimu Lu, Yunqiao Yang, Yuxuan Hu, Linda Wei, Mingjie Zhan, and Hongsheng Li. 2026. [From Solver to Tutor: Evaluating the Pedagogical Intelligence of LLMs with KMP-Bench](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32965–32973.
- Yao Shi, Rongkeng Liang, and Yong Xu. 2025. [EducationQ: Evaluating LLMs’ Teaching Capabilities Through Multi-Agent Dialogue Framework](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828, Vienna, Austria. Association for Computational Linguistics.
- 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, Anisha Gunjal, Yunzhong He, Bing Liu, and Chen Xing. 2025. [TutorBench: A Benchmark To Assess Tutoring Capabilities Of Large Language Models](#). *Preprint*, arXiv:2510.02663.
- Barbara G. Tabachnick, Linda S. Fidell, and Jodie B. Ullman. 2019. *Using Multivariate Statistics*, seventh edition. Pearson, NY, NY.
- Danielle R Thomas, Conrad Borchers, and Ken Koedinger. 2025. Beyond Agreement: Rethinking Ground Truth in Educational AI Annotation. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 345–351, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Ledyard R. Tucker. 1951. [A Method for Synthesis of Factor Analysis Studies](#). Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024. [Book2Dial: Generating Teacher Student Interactions from Textbooks for Cost-Effective Development of Educational Chatbots](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9707–9731, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Marley W. Watkins. 2018. [Exploratory Factor Analysis: A Guide to Best Practice](#). *Journal of Black Psychology*, 44(3):219–246.
- Shouang Wei, Min Zhang, Xin Lin, Bo Jiang, Zhongxiang Dai, and Kun Kuang. 2025. [EduDial: Constructing a Large-scale Multi-turn Teacher-Student Dialogue Corpus](#). *Preprint*, arXiv:2510.12899.
- Bin Xu, Yu Bai, Huashan Sun, Yiguan Lin, Siming Liu, Xinyue Liang, Yaolin Li, Yang Gao, and Heyan Huang. 2025. [EduBench: A Comprehensive Benchmarking Dataset for Evaluating Large Language Models in Diverse Educational Scenarios](#). *Preprint*, arXiv:2505.16160.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Human qualitative coding

Category	Core question	Subcategories
Overall Rating	How good was the tutor overall?	Pedagogical quality; overall helpfulness; future use; tutor competence; tutor intelligence; human-likeness
Accuracy	Was the tutor factually and logically correct?	Factual correctness; reasoning validity; domain correctness; uncertainty calibration
Alignment to Constraints	Did the tutor follow task, role, format, and pedagogical constraints?	System prompt alignment; staying on topic; role adherence; does not reveal answer when constrained
Instructional Support	Did the tutor explain concepts clearly, provide actionable support, and guide the learner effectively?	Explains concepts clearly; stepwise guidance; scaffolding; actionable hints; manageable chunks; organization and sequencing
Assessment	Did the tutor check understanding and monitor learner progress?	Checks understanding; progress monitoring; formative assessment; mastery checking
Mistake Handling	Did the tutor detect, diagnose, and productively respond to student errors or misconceptions?	Mistake identification; mistake localization; misconception diagnosis; corrective feedback; mistake guidance
Affect and Relational Support	Did the tutor feel warm, supportive, respectful, and emotionally responsive?	Encouragement; empathy; respectfulness; emotional responsiveness; naturalness
Engagement and Motivation	Did the tutor stimulate curiosity, effort, and active participation?	Promotes engagement; stimulates interest; recognition of success; sustains participation; supports persistence
Adaptivity	Did the tutor tailor its teaching to the learner's level, prior knowledge, goal, needs, and evolving state?	Learner-level calibration; adapts when student is stuck; personalization; adjusts support over time
Understanding Learner Goals	Did the tutor understand and respond to the learner's stated goal, question intent, expressed need, or intended learning outcome?	Goal understanding
Metacognition	Did the tutor help the learner reflect on their thinking, assumptions, strategies, perspectives, and learning process?	Reflection; critical thinking; self-monitoring; thinking about mistakes; planning next steps
Safety	Did the tutor avoid harmful, inappropriate, or disallowed content?	Danger; sexual content; profanity; hateful content; self-harm; substance use; bias and stereotyping

Table A1: Preliminary categories, core questions, and subcategories from the inductive coding of the 177-metric pool.

B Metric applicability

Metric	GPT-5.4 mini	Claude Haiku 4.5	Gemma 4 26B
Alignment to Constraints – Instruction following	Core	Conditional	Core
Instructional Support – Actionability	Core	Conditional	Core
Metacognition – Reflective growth	Core	Conditional	Core
Alignment to Constraints – Coverage	Core	Core	Conditional
Instructional Support – Communicates Plan	Core	Core	Conditional
Instructional Support – Improvement Suggestions	Core	Core	Conditional
Accuracy – Expresses Uncertainty expresses uncertainty when	Core	Conditional	Conditional
Adaptivity – Unstuck	Core	Conditional	Conditional
Alignment to Constraints – Academic Integrity Honesty	Core	Conditional	Conditional
Alignment to Constraints – Stay on topic	Core	Conditional	Conditional
Assessment – Self assessment	Core	Conditional	Conditional
Engagement and Motivation – Context Based Idea Generation	Core	Conditional	Conditional
Instructional Support – Manageable Chunks uses bullet points	Core	Conditional	Conditional
Metacognition – Error analysis	Core	Conditional	Conditional
Metacognition – Justification for Knowing	Core	Conditional	Conditional
Mistake Handling – Error Identification Correction Precision	Core	Conditional	Conditional
Mistake Handling – Guide Mistake Discovery	Core	Conditional	Conditional
Mistake Handling – Guide Mistake Discovery guides the student	Core	Conditional	Conditional
Mistake Handling – Identifies mistakes	Core	Conditional	Conditional
Mistake Handling – Mistake identification	Core	Conditional	Conditional
Affect and Relational Support – Adapts to Affect	Conditional	Conditional	Conditional
Affect and Relational Support – Adapts to Affect responds effectively if	Conditional	Conditional	Conditional
Affect and Relational Support – Respond appropriately to explicit affect cues	Conditional	Conditional	Conditional
Instructional Support – Explainability	Core	Core	Dropped
Mistake Handling – Identify and address misconceptions	Core	Core	Dropped
Accuracy – Expresses Uncertainty	Core	Conditional	Dropped
Alignment to Constraints – Responsibility and Ethics	Core	Conditional	Dropped
Alignment to Constraints – Transparency Accountability	Core	Conditional	Dropped
Metacognition – Self reflection	Core	Conditional	Dropped
Mistake Handling – Refinement	Core	Conditional	Dropped
Instructional Support – Communicates Aims	Core	Dropped	Conditional
Mistake Handling – Mistake location	Conditional	Conditional	Dropped
Adaptivity – Background Based Material Recommendation	Core	Dropped	Dropped
Adaptivity – Accessibility	Conditional	Dropped	Dropped
Affect and Relational Support – Moral responsibility	Conditional	Dropped	Dropped
Alignment to Constraints – Privacy Consent	Conditional	Dropped	Dropped
Alignment to Constraints – Strategic Deception Signals	Conditional	Dropped	Dropped
Metacognition – Cultural identity exploration	Conditional	Dropped	Dropped
Instructional Support – Partisanship	Dropped	Conditional	Dropped

Table B1: Metrics whose applicability differed from Core for at least one judge. The remaining 138 metrics were Core for all three judges and are omitted; the complete 177-metric table is available in the repository. Core = at least 90% valid scores; Conditional = 20–90%; Dropped = fewer than 200 (20%) valid scores.

C Per-judge pattern matrices

λ	Metric
<i>Factor 1: Accuracy</i> (43 salient items; 5 shown)	
.90	Alignment to Constraints – No Contradiction
.85	Accuracy – No Inaccuracies
.80	Accuracy – No Contradiction
.74	Accuracy – Basic Factual Accuracy
.72	Accuracy – Truthfulness
<i>Factor 2: Engagement and questioning</i> (26 salient items; 5 shown)	
.97	Engagement and Motivation – Asks Questions asks questions to
.95	Engagement and Motivation – Asks Questions
.89	Metacognition – Pedagogical Principle Criteria Questioning
.81	Instructional Support – Deep Socratic Follow Up Questions
.81	Engagement and Motivation – Promotes engagement
<i>Factor 3: Mistake handling</i> (11 salient items; 5 shown)	
.59	Mistake Handling – Refinement
.52	Metacognition – Error analysis
.52	Instructional Support – Persuasive discourse
.50	Alignment to Constraints – Transparency Accountability
.49	Mistake Handling – Mistake identification
<i>Factor 4: Alignment to constraints</i> (10 salient items; 5 shown)	
.63	Alignment to Constraints – No Refusals
.58	Alignment to Constraints – Stay on topic
.56	Overall Rating – Learning Goal
.51	Adaptivity – Unstuck
.50	Alignment to Constraints – Coverage
<i>Factor 5: Analogies</i> (4 salient items; all shown)	
.96	Instructional Support – Analogies
.87	Instructional Support – Analogies the tutors use
.80	Affect and Relational Support – Cultural analogies
.44	Adaptivity – Leveled Concept Explanation
<i>Factor 6: Affect and relational support</i> (6 salient items; 5 shown)	
.55	Affect and Relational Support – Warm
.50	Affect and Relational Support – Communicate with positive tone
.47	Engagement and Motivation – Emotional Support
.46	Engagement and Motivation – Encouraging Feedback
.42	Affect and Relational Support – Learning Exploration
<i>Factor 7: Metacognition</i> (8 salient items; 5 shown)	
.82	Metacognition – Self reflection
.67	Metacognition – Reflection
.63	Metacognition – Strategic thinking
.59	Metacognition – Reflective growth
.58	Metacognition – Metacognitive Promotion

Table C1: Oblimin pattern loadings for GPT-5.4 mini. The five highest-loading items are shown per factor; loadings are salient at $|\lambda| \geq .40$. Complete pattern matrices for all three judges are available in the repository.

λ	Metric
<i>Factor 1: Engagement and questioning (36 salient items; 5 shown)</i>	
.89	Engagement and Motivation – Asks Questions asks questions to
.88	Instructional Support – Deep Socratic Follow Up Questions
.87	Engagement and Motivation – Asks Questions
.84	Metacognition – Pedagogical Principle Criteria Questioning
.82	Metacognition – Generative questioning
<i>Factor 2: Encouragement (16 salient items; 5 shown)</i>	
.87	Affect and Relational Support – Communicate with positive tone
.87	Engagement and Motivation – Encouraging Feedback
.77	Affect and Relational Support – Warm
.75	Engagement and Motivation – Emotional Support
.75	Engagement and Motivation – Encouraging Feedback delivers feedback whether
<i>Factor 3: Style and presentation (6 salient items; 5 shown)</i>	
.47	Affect and Relational Support – Fluency
.47	Instructional Support – Info Presentation
.43	Instructional Support – Info Presentation presents information in
.41	Affect and Relational Support – Style and Tone
.41	Instructional Support – Straightforward Response the tutors responses
<i>Factor 4: Alignment to goals (7 salient items; 5 shown)</i>	
.63	Overall Rating – Learning Goal
.57	Alignment to Constraints – Content Relevance Scope Control
.55	Alignment to Constraints – Coverage
.53	Alignment to Constraints – Relevance
.47	Understanding Learner Goals – Helpfulness
<i>Factor 5: Accuracy (7 salient items; 5 shown)</i>	
.90	Accuracy – No Contradiction
.89	Accuracy – No Inaccuracies
.88	Alignment to Constraints – No Contradiction
.78	Accuracy – Basic Factual Accuracy
.78	Accuracy – Truthfulness
<i>Factor 6: Analogies (4 salient items; all shown)</i>	
.98	Instructional Support – Analogies
.90	Instructional Support – Analogies the tutors use
.85	Affect and Relational Support – Cultural analogies
.52	Adaptivity – Leveled Concept Explanation
<i>Factor 7: Metacognition (6 salient items; 5 shown)</i>	
.91	Metacognition – Strategic thinking
.76	Metacognition – Reflection
.59	Instructional Support – Feedback Assessment
.50	Engagement and Motivation – Pedagogical Principle Criteria Feedback
.49	Metacognition – Metacognitive Promotion
<i>Factor 8: Alignment to constraints (4 salient items; all shown)</i>	
.77	Alignment to Constraints – No Repetition
.74	Alignment to Constraints – No Repetition avoids repeating information
.45	Alignment to Constraints – No Irrelevant Info
.45	Instructional Support – No Irrelevant Info
<i>Factor 9: Adaptivity (1 salient item; not interpreted)</i>	
.43	Adaptivity – Guides Appropriately

Table C2: Oblimin pattern loadings for Claude Haiku 4.5. The five highest-loading items are shown per factor; loadings are salient at $|\lambda| \geq .40$. Factor 9 was not interpreted because only one item loaded saliently.

λ	Metric
<i>Factor 1: Engagement and questioning (29 salient items; 5 shown)</i>	
.93	Instructional Support – Deep Socratic Follow Up Questions
.92	Engagement and Motivation – Asks Questions asks questions to
.91	Engagement and Motivation – Asks Questions
.87	Alignment to Constraints – Dont reveal the answer
.86	Metacognition – Generative questioning
<i>Factor 2: Encouragement (20 salient items; 5 shown)</i>	
.87	Affect and Relational Support – Communicate with positive tone
.86	Affect and Relational Support – Warm
.79	Affect and Relational Support – Speaks encouragingly
.79	Engagement and Motivation – Emotional Support
.79	Engagement and Motivation – Encouraging Feedback delivers feedback whether
<i>Factor 3: Style and presentation (16 salient items; 5 shown)</i>	
.56	Alignment to Constraints – Conciseness and Relevance
.51	Overall Rating – Continue using
.50	Overall Rating – Overall Quality
.49	Instructional Support – Info Presentation presents information in
.49	Instructional Support – Info Presentation
<i>Factor 4: Accuracy (8 salient items; 5 shown)</i>	
.90	Accuracy – No Inaccuracies
.86	Accuracy – Basic Factual Accuracy
.82	Accuracy – No Contradiction
.81	Alignment to Constraints – No Contradiction
.78	Accuracy – Truthfulness
<i>Factor 5: Analogies (5 salient items; all shown)</i>	
.80	Instructional Support – Analogies
.79	Instructional Support – Analogies the tutors use
.76	Affect and Relational Support – Cultural analogies
.47	Adaptivity – Leveled Concept Explanation
.43	Instructional Support – Pedagogical Principle Criteria Explanation
<i>Factor 6: Participation (7 salient items; 5 shown)</i>	
.43	Engagement and Motivation – Opportunities for Engagement
.43	Affect and Relational Support – Well intentioned
.42	Adaptivity – Continuous assessment
.41	Instructional Support – Info Order develops explanations in
.41	Engagement and Motivation – Active Engagement
<i>Factor 7: Metacognition (7 salient items; 5 shown)</i>	
.99	Metacognition – Reflection
.97	Instructional Support – Feedback Assessment
.90	Metacognition – Metacognitive Promotion
.69	Metacognition – Strategic thinking
.50	Instructional Support – Actionability

Table C3: Oblimin pattern loadings for Gemma 4 26B. The five highest-loading items are shown per factor; loadings are salient at $|\lambda| \geq .40$.

D Conditional-metric extension mappings

Conditional metric	N	Salient factor mapping (λ)	Strongest λ
Mistake Handling – Mistake location	877	Accuracy (F1): .50	F1: .50
Alignment to Constraints – Privacy Consent	819	Alignment to constraints/goals (F4): .48	F4: .48
Affect and Relational Support – Respond appropriately to explicit affect cues	820	Alignment to constraints/goals (F4): .44	F4: .44
Affect and Relational Support – Adapts to Affect responds effectively if	828	Alignment to constraints/goals (F4): .42	F4: .42
Affect and Relational Support – Adapts to Affect	786	No salient mapping	F6: .34
Adaptivity – Accessibility	375	No salient mapping	F1: .26
Metacognition – Cultural identity exploration	581	No salient mapping	F5: .23
Affect and Relational Support – Moral responsibility	582	No salient mapping	F1: -.10
Alignment to Constraints – Strategic Deception Signals	572	Not estimable	—

Table D1: Conditional-metric extension mappings onto the core EFA factors for GPT-5.4 mini. Of nine conditional metrics, four mapped saliently ($|\lambda| \geq .40$), four fell below threshold, and one was not estimable. N is the number of conversations in which the metric received a valid score.

Conditional metric	<i>N</i>	Salient factor mapping (λ)	Strongest λ
Metacognition – Self reflection	318	Metacognition (F7): .91	F7: .91
Assessment – Self assessment	683	Metacognition (F7): .75	F7: .75
Metacognition – Reflective growth	805	Metacognition (F7): .71	F7: .71
Affect and Relational Support – Adapts to Affect	360	Affect and encouragement (F2): .60	F2: .60
Mistake Handling – Guide Mistake Discovery guides the student	825	Engagement and questioning (F1): .59	F1: .59
Engagement and Motivation – Context Based Idea Generation	554	Engagement and questioning (F1): .51	F1: .51
Metacognition – Error analysis	884	Metacognition (F7): .48	F7: .48
Mistake Handling – Refinement	626	Metacognition (F7): .47	F7: .47
Instructional Support – Actionability	806	Metacognition (F7): .46	F7: .46
Mistake Handling – Guide Mistake Discovery	618	Engagement and questioning (F1): .46	F1: .46
Mistake Handling – Identifies mistakes	403	Metacognition (F7): .45	F7: .45
Mistake Handling – Mistake location	329	Accuracy (F5): .44	F5: .44
Alignment to Constraints – Transparency Accountability	307	Metacognition (F7): .44	F7: .44
Affect and Relational Support – Respond appropriately to explicit affect cues	394	Affect and encouragement (F2): .43	F2: .43
Mistake Handling – Error Identification Correction Precision	238	Metacognition (F7): .42	F7: .42
Alignment to Constraints – Stay on topic	447	Alignment to constraints/goals (F4): .41	F4: .41
Mistake Handling – Mistake identification	574	No salient mapping	F7: .38
Affect and Relational Support – Adapts to Affect responds effectively if	336	No salient mapping	F2: .37
Metacognition – Justification for Knowing	860	No salient mapping	F7: .34
Alignment to Constraints – Instruction following	739	No salient mapping	F8: .32
Adaptivity – Unstuck	865	No salient mapping	F4: .31
Alignment to Constraints – Academic Integrity Honesty	592	No salient mapping	F8: .29
Instructional Support – Manageable Chunks uses bullet points	610	No salient mapping	F2: .27
Instructional Support – Partisanship	202	No salient mapping	F1: .27
Alignment to Constraints – Responsibility and Ethics	650	No salient mapping	F2: .26
Accuracy – Expresses Uncertainty expresses uncertainty when	880	No salient mapping	F7: .23
Accuracy – Expresses Uncertainty	733	No salient mapping	F1: .16

Table D2: Conditional-metric extension mappings onto the core EFA factors for Claude Haiku 4.5. Of 27 conditional metrics, 16 mapped saliently ($|\lambda| \geq .40$) and 11 fell below threshold. *N* is the number of conversations in which the metric received a valid score.

Conditional metric	<i>N</i>	Salient factor mapping (λ)	Strongest λ
Assessment – Self assessment	258	Metacognition (F7): .79	F7: .79
Mistake Handling – Guide Mistake Discovery guides the student	836	Engagement and questioning (F1): .51	F1: .51
Affect and Relational Support – Respond appropriately to explicit affect cues	450	Affect and encouragement (F2): .50	F2: .50
Mistake Handling – Guide Mistake Discovery	677	Engagement and questioning (F1): .49	F1: .49
Alignment to Constraints – Coverage	528	Accuracy (F4): .49	F4: .49
Instructional Support – Communicates Aims	200	Metacognition (F7): .47	F7: .47
Instructional Support – Improvement Suggestions	734	Affect and encouragement (F2): .47 Metacognition (F7): .47	F7: .47
Affect and Relational Support – Adapts to Affect	325	Affect and encouragement (F2): .43	F2: .43
Engagement and Motivation – Context Based Idea Generation	714	Engagement and questioning (F1): .42	F1: .42
Alignment to Constraints – Stay on topic	716	No salient mapping	F3: .39
Affect and Relational Support – Adapts to Affect responds effectively if	610	No salient mapping	F2: .38
Adaptivity – Unstuck	743	No salient mapping	F3: .38
Mistake Handling – Identifies mistakes	321	No salient mapping	F7: .36
Mistake Handling – Mistake identification	339	No salient mapping	F7: .35
Metacognition – Justification for Knowing	375	No salient mapping	F4: .33
Instructional Support – Manageable Chunks uses bullet points	389	No salient mapping	F5: .33
Accuracy – Expresses Uncertainty expresses uncertainty when	202	No salient mapping	F2: .33
Mistake Handling – Error Identification Correction Precision	812	No salient mapping	F3: .32
Metacognition – Error analysis	847	No salient mapping	F7: .31
Instructional Support – Communicates Plan	759	No salient mapping	F7: .28
Alignment to Constraints – Academic Integrity Honesty	795	No salient mapping	F6: .23

Table D3: Conditional-metric extension mappings onto the core EFA factors for Gemma 4 26B. Of 21 conditional metrics, nine mapped saliently ($|\lambda| \geq .40$) and 12 fell below threshold. *N* is the number of conversations in which the metric received a valid score.

E Proposed metric framework

Category	Subcategory	Prompt	References
Overall Quality	Pedagogical quality	Your task is to rate the Tutor’s overall pedagogical quality. Focus on how the Tutor supported the Student’s learning across the conversation, considering the Tutor’s teaching decisions in context rather than any single behavior.	(Jurenka et al., 2024; Macina et al., 2025; Lee et al., 2026)
Overall Quality	Overall helpfulness	Your task is to rate how helpful the Tutor was overall. Focus on the Tutor’s overall usefulness to the Student, not on any single specific behavior.	(Jurenka et al., 2024; Ding and Magerko, 2025)
Overall Quality	Tutor competence	Your task is to rate how competent the Tutor was overall. Focus on the Tutor’s general capability and effectiveness, not on any one isolated skill.	(Jurenka et al., 2024; LearnLM Team, 2024; Shi et al., 2025)
Affect & Encouragement	Encouragement	Your task is to rate how well the Tutor communicates in a warm, encouraging, and supportive way.	(Jurenka et al., 2024; LearnLM Team, 2024; Maurya et al., 2025)
Affect & Encouragement	Empathy	Your task is to rate how appropriately the Tutor responds to the Student’s explicitly expressed feelings or emotional state. For example, the Tutor may acknowledge frustration, discouragement, or pride in a supportive way.	(Jurenka et al., 2024; Ding and Magerko, 2025; Srinivasa et al., 2025; Wei et al., 2025)
Affect & Encouragement	Stimulates interest	Your task is to rate how well the Tutor stimulates the Student’s curiosity or interest in the topic. Focus on whether the Tutor makes the interaction intellectually motivating.	(LearnLM Team, 2024; Xu et al., 2025; Jurenka et al., 2024)
Affect & Encouragement	Recognition of success	Your task is to rate how well the Tutor recognizes and positively reinforces the Student’s successes, progress, or correct reasoning.	(Jurenka et al., 2024; LearnLM Team, 2024)
Engagement & Questioning	Promotes engagement	Your task is to rate how well the Tutor actively involves the Student in the learning process. For example, the Tutor may provide practice problems or opportunities for participation.	(Jurenka et al., 2024; LearnLM Team, 2024)
Engagement & Questioning	Guiding questions	Your task is to rate the extent to which the Tutor asks guiding questions to make the Student think and work toward the answer themselves.	(Jurenka et al., 2024; LearnLM Team, 2024; Macina et al., 2025; Sharma et al., 2026; Lee et al., 2026; Wei et al., 2025; Y. Shi et al., 2025; W. Shi et al., 2026)
Engagement & Questioning	Formative assessment	Your task is to rate how well the Tutor uses questions or prompts to assess the Student’s current understanding in order to guide the next teaching move.	(Sharma et al., 2026; Wei et al., 2025; Lee et al., 2026)
Engagement & Questioning	Does not give away answer too quickly	Your task is to rate the extent to which the Tutor avoids giving away the answer too quickly when it should guide the Student through reasoning or productive struggle.	(Jurenka et al., 2024; LearnLM Team, 2024; Maurya et al., 2025; Srinivasa et al., 2025; Lee et al., 2026)
Engagement & Questioning	Higher-order thinking	Your task is to rate how well the Tutor helps the Student move beyond surface-level answers and think more deeply, including through logical analysis, evaluation, or synthesis of information.	(Sharma et al., 2026; Xu et al., 2025; Wei et al., 2025; Ding and Magerko, 2025)

Continued on next page

Category	Subcategory	Prompt	References
Accuracy	Factual correctness	Your task is to rate the factual correctness of the Tutor's statements. Focus on whether the content is true and free of factual errors.	(Jurenka et al., 2024; LearnLM Team, 2024; Macina et al., 2025; Srinivasa et al., 2025; Xu et al., 2025; Lee et al., 2026)
Accuracy	Reasoning validity	Your task is to rate how logically valid, justifiable, and free of contradictions the Tutor's reasoning is.	(Jurenka et al., 2024; LearnLM Team, 2024; Xu et al., 2025)
Accuracy	Uncertainty calibration	Your task is to rate how appropriately the Tutor expresses uncertainty when it is not fully sure, rather than presenting uncertain claims as certain facts.	(Jurenka et al., 2024; LearnLM Team, 2024; Lee et al., 2026)
Metacognition	Reflection prompts	Your task is to rate how well the Tutor encourages the Student to reflect on their thinking, understanding, or learning process.	(Sharma et al., 2026; Lee et al., 2026; Shi et al., 2025)
Metacognition	Self-monitoring	Your task is to rate how well the Tutor helps the Student monitor their own understanding, confidence, or progress while solving the task.	(Sharma et al., 2026; Wei et al., 2025; Shi et al., 2025)
Metacognition	Planning next steps	Your task is to rate how well the Tutor helps the Student identify productive next steps for learning, practice, or improvement.	(Sharma et al., 2026; Lee et al., 2026)
Style & Presentation	Naturalness	Your task is to rate the Tutor's responses in terms of sounding natural, conversational, and socially appropriate rather than robotic or unnatural.	(Maurya et al., 2025; Wei et al., 2025)
Style & Presentation	Information presentation	Your task is to rate how well the Tutor presents information in a well-structured and straightforward way.	(Jurenka et al., 2024; LearnLM Team, 2024; Srinivasa et al., 2025; Wei et al., 2025)
Style & Presentation	Concise	Your task is to rate how concisely the Tutor communicates. Focus on whether the response is an appropriate length and avoids repetitive information.	(Jurenka et al., 2024; LearnLM Team, 2024; Srinivasa et al., 2025; Wei et al., 2025)
Instructional Support	Explains concepts clearly	Your task is to rate how clearly the Tutor explains underlying concepts or skills in a way that is easy for the Student to understand.	(Jurenka et al., 2024; Maurya et al., 2025; Macina et al., 2025; LearnLM Team, 2024; Srinivasa et al., 2025; Xu et al., 2025; Wei et al., 2025; Shi et al., 2026)
Instructional Support	Analogies	Your task is to rate how well the Tutor uses narratives, case studies, or analogies as appropriate to support student learning.	(Jurenka et al., 2024; LearnLM Team, 2024; Srinivasa et al., 2025; Sharma et al., 2026; Lee et al., 2026)
Instructional Support	Stepwise guidance	Your task is to rate how effectively the Tutor guides the Student through the task using appropriate intermediate steps.	(Jurenka et al., 2024; Maurya et al., 2025; Macina et al., 2025; LearnLM Team, 2024; Srinivasa et al., 2025; Lee et al., 2026)
Instructional Support	Scaffolding	Your task is to rate how well the Tutor provides an appropriate level of support to move the Student forward without making the task unnecessarily easy or doing all the thinking for them.	(Jurenka et al., 2024; Sharma et al., 2026; Macina et al., 2025; Shi et al., 2025)

Continued on next page

Category	Subcategory	Prompt	References
Instructional Support	Actionable hints	Your task is to rate how well the Tutor provides hints or cues that the Student could realistically act on to make progress.	(Jurenka et al., 2024; Maurya et al., 2025; Srinivasa et al., 2025; Lee et al., 2026)
Instructional Support	Manageable chunks	Your task is to rate how well the Tutor breaks information into manageable pieces that are easy to process.	(LearnLM Team, 2024; Jurenka et al., 2024)
Instructional Support	Organization and sequencing	Your task is to rate how well the Tutor presents information in a logical order that builds understanding from simpler or more intuitive ideas toward more complex ones.	(Jurenka et al., 2024; LearnLM Team, 2024; Shi et al., 2025)
Alignment to Constraints	System prompt alignment	Your task is to rate how well the Tutor follows the requirements, role expectations, and constraints given in its instructions or system prompt.	(Macina et al., 2025; Srinivasa et al., 2025; Xu et al., 2025; Ding and Magerko, 2025; Shi et al., 2026)
Alignment to Constraints	Staying on topic	Your task is to rate how well the Tutor keeps the conversation focused on the relevant topic or lesson goal rather than drifting off topic. For example, the Tutor should avoid irrelevant content or redirect the Student as needed.	(Jurenka et al., 2024; LearnLM Team, 2024; Maurya et al., 2025; Srinivasa et al., 2025; Xu et al., 2025; Wei et al., 2025)
Alignment to Constraints	Role adherence	Your task is to rate how well the Tutor behaves in a way that matches its intended pedagogical role, such as tutor, guide, or coach.	(Shi et al., 2026; Xu et al., 2025)
Alignment to Constraints	Contextual consistency	Your task is to rate how well the Tutor’s responses logically follow from the Student’s message and prior dialogue context.	(Maurya et al., 2025; Xu et al., 2025; Shi et al., 2026)
Mistake Handling	Mistake identification	Your task is to rate how well the Tutor detects and identifies mistakes made by the Student.	(Jurenka et al., 2024; Maurya et al., 2025; Miller and DiCerbo, 2024; Srinivasa et al., 2025; Xu et al., 2025; Ding and Magerko, 2025)
Mistake Handling	Mistake localization	Your task is to rate how well the Tutor identifies where the Student’s mistake occurred or which part of the reasoning was wrong.	(Maurya et al., 2025; Macina et al., 2025; Xu et al., 2025; Ding and Magerko, 2025; Srinivasa et al., 2025)
Mistake Handling	Misconception diagnosis	Your task is to rate how well the Tutor identifies the underlying misconception or misunderstanding behind the Student’s error.	(Jurenka et al., 2024; Otero et al., 2025; Srinivasa et al., 2025)
Mistake Handling	Mistake guidance	Your task is to rate how effectively the Tutor guides the Student through discovering or fixing a mistake.	(Jurenka et al., 2024; LearnLM Team, 2024)
Adaptivity	Goal responsiveness	Your task is to rate how well the Tutor identifies and uses the Student’s learning goal, question intent, or stated objective to adapt its response.	(LearnLM Team, 2024; Jurenka et al., 2024; Shi et al., 2025; Ding and Magerko, 2025)
Adaptivity	Learner-level calibration	Your task is to rate how well the Tutor adjusts the level of explanation, vocabulary, and difficulty to match the Student’s current level of understanding.	(Jurenka et al., 2024; LearnLM Team, 2024; Srinivasa et al., 2025; Xu et al., 2025; Lee et al., 2026)
Adaptivity	Adapts when student is stuck	Your task is to rate how well the Tutor adapts when the Student becomes stuck, for example by changing the explanation, giving a hint, or reducing complexity.	(Jurenka et al., 2024; LearnLM Team, 2024)

Continued on next page

Category	Subcategory	Prompt	References
Adaptivity	Personalization	Your task is to rate how well the Tutor tailors its teaching to the Student's relevant background, learning needs, or ongoing performance rather than giving the same generic response to any learner.	(LearnLM Team, 2024; Xu et al., 2025; Ding and Magerko, 2025; Lee et al., 2026)
Adaptivity	Adjusts support over time	Your task is to rate how well the Tutor adjusts its level of support as the conversation progresses in response to the Student's evolving needs or demonstrated understanding.	(Jurenka et al., 2024; LearnLM Team, 2024; Shi et al., 2025; Wei et al., 2025)

Table E1: Proposed metric category and subcategory framework: 10 categories and 41 items derived from the crosswalk in Table 4.

F LLM-as-judge system prompt

Role

You are an experienced K12 educator assessing a Tutor who is interacting with a Student.

Scoring Guidelines

- Aim to be objective, unbiased, and constructive
- Assign low scores when warranted
- Focus solely on the responses and actions of the Tutor
- Your result must use the scale provided
- If a criterion does not apply to the conversation, assign it "N/A"

Conversation to Evaluate

[full conversation transcript]

Evaluate the conversation on the following criterion:

[criterion metric]

Rating scale: 1 (Very Poorly), 2 (Poorly), 3 (Adequately), 4 (Well), 5 (Very Well)

Respond with ONLY the integer score or "N/A".