
CONVOLEARN: A Learning Sciences Grounded Dataset for Fine-Tuning Dialogic AI Tutors

Mayank Sharma
Stanford University
masharma@stanford.edu

Roy Pea
Stanford University
roypea@stanford.edu

Hari Subramonyam
Stanford University
harihars@stanford.edu

Abstract

Despite their growing adoption in education, LLMs remain misaligned with the core principle of effective tutoring: the dialogic construction of knowledge. We introduce CONVOLEARN¹, a dataset of 2,134 semi-synthetic tutor-student dialogues operationalizing six dimensions of dialogic tutoring grounded in knowledge-building theory, situated in middle school Earth Science curriculum. We show that dimension-labeled dialogic training data captures meaningful pedagogical signal that generalizes beyond its semi-synthetic domain: scores from a classifier trained on CONVOLEARN correlate significantly with expert-coded instructional quality in authentic classrooms across multiple subscales (range $r = .118-.258$, all $p < .05$). As a proof of concept, we fine-tune MISTRAL-7B on CONVOLEARN and show that dimension-level fine-tuning can steer a 7B open-weight model toward dialogic tutoring behavior that credentialed teachers rate as competitive with a strong proprietary baseline. With this work, we support the development of AI tutors capable of more dialogic interactions.

1 Introduction

Large language models (LLMs) are rapidly reshaping the educational landscape. According to a nationally representative survey, 54% of K-12 students in the United States reported using AI for school work during the 2024-25 school year (Doss et al., 2025), and AI investment in education is projected to reach \$112 billion by 2034 (Precedence Research, 2025). Much of this adoption centers around AI-powered tutoring, where LLMs are being deployed to provide students with personalized academic support (for example, Khanmigo (Khan Academy, 2026)). This focus builds on longstanding evidence that individualized instruction can dramatically improve learning outcomes: Bloom (1984)’s two-sigma study showed that 1:1 tutoring enabled average students to reach the 98th percentile of traditionally instructed peers. LLMs hold promise to approximate this personalized tutoring at scale.

Despite this optimism, research in educational NLP demonstrates that LLMs exhibit fundamental pedagogical limitations. Modern foundation models are alignment-tuned to be broadly ‘helpful’ assistants, but helpfulness is often considered ‘at odds with pedagogy and learning’ (Jurenka et al., 2024). RLHF/RLAIF techniques further optimize for direct and satisfying answers (Ouyang et al., 2022), producing models that talk *at* students rather than *with* them. Effective tutoring, however, is inherently dialogic: it draws out student thinking, builds on it, and guides students to construct understanding themselves (Chi et al., 2001). Research confirms this gap is large and consistent: despite scoring high on surface qualities like coherence and human-likeness (Maurya et al., 2025), LLMs substantially underperform human tutors on dialogic dimensions such as providing meaningful guidance (Maurya et al., 2025) and engaging students in genuine reasoning and learning rather than passive receipt of information (Macina et al., 2023a; Tack & Piech, 2022).

The research community has responded with approaches such as Socratic prompting (Zhang et al., 2024; Liu et al., 2024; Favero et al., 2024), evaluation benchmarks (Macina et al., 2023a;

¹<https://huggingface.co/datasets/masharma/convolearn>

Maurya et al., 2025; Kochmar et al., 2025), and most recently, fine-tuning datasets built from authentic tutoring recordings (Perczel et al., 2025). Each addresses a different layer of the problem: prompting steers behavior at inference time without changing what a model has learned, benchmarks quantify pedagogical failures without providing a signal to correct them, and authentic recordings capture naturalistic tutoring behavior but are not decomposed into theoretically grounded dimensions that could guide targeted fine-tuning. While proprietary systems such as LearnLM attempt to directly address these limitations through pedagogy-informed training (Jurenka et al., 2024), their data and processes are not publicly available. The result is a critical gap: no openly available dataset labels specific, theoretically grounded dialogic behaviors at the dimension level, leaving researchers without the training signal needed to steer models toward effective tutoring. This gap is all the more consequential given evidence that LLM misalignment with effective pedagogy is structural, rooted in shared pretraining rather than model or prompt choice (Hardy & Kim, 2026), making targeted, theory-driven fine-tuning data not just useful, but necessary.

We introduce **CONVOLEARN**, a dataset of 2,134 dialogic tutor–student dialogues (~20 turns each), in which teacher turns are authored by credentialed human teachers and student turns are simulated. While **CONVOLEARN** cannot fully overcome structural pretraining biases, it provides an openly available resource operationalizing six dimensions of dialogic tutoring grounded in knowledge-building theory (Scardamalia & Bereiter, 2014): cognitive engagement, formative assessment, accountability, cultural responsiveness, metacognition, and power dynamics, across 21 fine-grained subdimensions, situated in California’s middle school Earth Science curriculum. Our contributions are threefold: **(1) Dataset**: we release **CONVOLEARN** with quality ratings preserved across the spectrum, enabling diverse downstream uses including fine-tuning, classifier training, and contrastive learning; **(2) Ecological validity**: we provide partial external validation for three dimensions most amenable to linguistic operationalization (cognitive engagement, metacognition, and formative assessment) by training a Longformer classifier on **CONVOLEARN** and applying it to authentic classroom transcripts from the NCTE corpus (Demszky & Hill, 2023), where derived scores correlate significantly with expert-coded instructional quality measures; and **(3) Proof of concept**: we fine-tune open-weight LLMs on a high-quality subset using multi-turn QLoRA and evaluate against proprietary baselines via teacher raters, demonstrating that dimension-level fine-tuning on **CONVOLEARN** can steer models toward more dialogic tutoring strategies.

2 Prior Work

Much of the literature on AI tutors has centered on *mistake remediation*, focusing on identifying and correcting student errors in structured problem-solving settings. Here, we instead focus on work related to *dialogic tutoring*, an area that remains relatively underexplored and emphasizes eliciting student thinking, building on their ideas, and guiding learners to construct understanding themselves (Chi et al., 2001). Early work by Tack & Piech (2022) assessed models on their ability to produce teacher talk, understand students, and help them, finding that conversational agents fell significantly short of human teachers on all three dimensions, even when performing comparably on surface measures such as conversational uptake. Macina et al. (2023b) showed that, although models could produce fluent tutoring responses, human evaluation revealed substantially fewer genuine learning opportunities through scaffolding, reflection, and engagement. Building on these findings, Maurya et al. (2025) consolidated prior efforts into a unified eight-dimension taxonomy, demonstrating that even SoTA LLMs function more like Q&A systems, scoring highly on qualities such as human-likeness while underperforming on guidance and actionability.

2.1 Approaches to Dialogic Alignment

Attempts to align LLMs toward more dialogic tutoring behaviors have included prompting, evaluation benchmarking, and fine-tuning. Socratic prompting approaches show that steering models toward inquiry-based dialogue can improve critical thinking and self-reflection (Zhang et al., 2024; Liu et al., 2024; Favero et al., 2024), though these modify behavior only at inference time without changing what a model has learned. On the evaluation side, Maurya

et al. (2025) release MRBench, a benchmark of 192 tutoring conversations annotated across eight pedagogical dimensions. Most recently, Jiang et al. (2025) introduce SID, a benchmark of 10,000 interdisciplinary STEM dialogue turns specifically targeting dialogic guidance, finding that even state-of-the-art LLMs struggle to execute effective guided dialogues that support knowledge integration and transfer, establishing the evaluation gap but leaving the training data gap unaddressed. On the fine-tuning side, Google’s LearnLM applies pedagogy-informed post-training to improve instructional behaviors such as guiding students to answers and promoting engagement (Jurenka et al., 2024), but its data and training process are not publicly available. TeachLM, trained on over 100,000 hours of authentic one-on-one tutoring interactions using PEFT, increases student talk time, diversifies questioning strategies, and produces around 50% more dialogue turns than base models (Perczel et al., 2025), though its training data is likewise proprietary.

2.2 Remaining Gaps

Despite these advances, significant gaps remain, as summarized in Table 1. No existing fine-tuning resource is annotated at the level of specific, theoretically grounded dialogic dimensions: both LearnLM and TeachLM demonstrate the value of pedagogy-informed fine-tuning but make neither their training data nor dimension-level labels publicly available. While SID provides the closest analog to a dialogic tutoring benchmark, it is evaluation-only, fully synthetic, and grounded in a non-English curriculum, establishing the evaluation gap but leaving the training data gap unaddressed. Furthermore, no existing resource includes conversation-level quality ratings, limiting their utility for effectiveness-weighted fine-tuning and contrastive learning. CONVOLEARN addresses these gaps by providing 2,134 dialogues labeled across six theoretically grounded dimensions of dialogic tutoring, authored by 323 credentialed K-12 teachers with a mean of 10.9 yrs of experience, with full quality ratings preserved across the quality spectrum to support diverse downstream uses.

Property	MRBench*	SID	LearnLM	TeachLM	CONVOLEARN (Ours)
Size	192 convos	10K turns	N/A	100K hrs	2,134 convos
Avg. Turns	–	5+	–	–	20
Setting	K–12	K-12	Mixed	Mixed	K–12
Subject	Math	STEM	Mixed	Mixed	Earth Sci.
Type	Eval	Eval	FT	FT	FT
Dim. Labels	✓ (8)	✓ (9)	✗	✗	✓ (6+21)
Quality Ratings	✗	✗	✗	✗	✓
Open	✓	✓	✗	✗	✓

Table 1: Comparison of CONVOLEARN with related dialogic AI tutoring evaluation and fine-tuning resources. FT = Fine-tuning. Dim. Labels indicates whether dialogues are annotated with grounded pedagogical dimension labels (brackets indicate dimension count). Quality Ratings indicates whether per-dialogue effectiveness scores are included. Open indicates whether the resource is publicly available. *MRBench is primarily focused on mistake remediation but includes several dimensions that overlap with dialogic tutoring concerns.

3 Dialogic Learning Framework

Effective tutoring is not merely about conveying correct information: it requires supporting learners in actively generating, refining, and connecting ideas. We ground CONVOLEARN in knowledge-building theory (Scardamalia & Bereiter, 2014), which frames learning as a collaborative process of constructing and refining understanding through discourse rather than passive receipt of information. This positions the tutor not as an answer-provider but as a dialogic partner who elicits student thinking and adapts to it in real time. For example, when a student asks “*Why did the dinosaurs disappear?*”, a knowledge-passing tutor simply states that an asteroid caused the extinction. A knowledge-building tutor instead opens space for inquiry: “*What ideas do you have about why dinosaurs may have disappeared?*” This invites students to propose possibilities such as climatic shifts, food-web collapse, or volcanic

activity, followed by dialogic moves that help them compare theories, evaluate evidence, and refine their explanations. We operationalize this framework through six dimensions of dialogic tutoring, further broken down in 21 subdimensions (see full definitions of each subdimension and dataset excerpts in Appendix A).

Cognitive Engagement refers to the depth of processing and quality of thinking strategies students employ during learning (Blumenfeld et al., 2006; Chi & Wylie, 2014). In dialogic tutoring, cognitive engagement is the most direct behavioral counterpart to answer-giving: where an answer-giving tutor resolves cognitive challenge by providing solutions, a dialogically engaging tutor uses that challenge as the site of learning. Linguistically, it manifests as open-ended questioning, uptake of student ideas, and scaffolded elaboration rather than declarative explanation. It is operationalized through four subdimensions: scaffolding, critical thinking, generative questioning, and problem-based reasoning.

Formative Assessment refers to the ongoing, interactive monitoring of student understanding during instruction to regulate learning in real time (Cowie & Bell, 1999; Black & Wiliam, 2009). Unlike summative evaluation, it is embedded within the dialogue: tutors attend to student contributions, interpret them against learning goals, and adapt their next move accordingly. Linguistically, it appears as comprehension checks, probing follow-up questions, and responses that build on or correct student ideas. It is operationalized through three subdimensions: continuous assessment, self-assessment, and synthesizing.

Accountability reflects expectations that discourse aligns with norms of evidence and reasoning (Michaels et al., 2008). In dialogic tutoring, accountability moves the conversation beyond mere exchange of opinions toward epistemic responsibility: students are expected to justify claims, evaluate evidence, and engage with counterarguments. Linguistically, it manifests as tutor prompts that require students to cite evidence, explain their reasoning, or defend a position. It is operationalized through three subdimensions: evidence-based reasoning, moral responsibility, and depth of reasoning.

Cultural Responsiveness recognizes that effective instruction must engage learners' diverse cultural backgrounds and build on their prior knowledge and experiences (Ladson-Billings, 1995; Au et al., 1981). In dialogic tutoring, it requires tutors to connect concepts to culturally relevant contexts and affirm diverse ways of knowing rather than assuming a single canonical frame. Linguistically, it manifests as analogies drawn from students' cultural contexts and invitations to connect content to their experience. It is operationalized through two subdimensions: cultural identity exploration and cultural analogies.

Metacognition refers to awareness and regulation of one's own cognitive processes (Flavell, 1981). In dialogic tutoring, it involves prompting reflection and modeling through thinking aloud, making expert reasoning visible and encouraging students to adopt similar habits. Linguistically, it appears as prompts asking students to explain their reasoning, identify difficulties, or compare current understanding with prior beliefs. It is operationalized through six subdimensions: reflective growth, strategic thinking, self-reflection, thinking aloud, error analysis, and reflection.

Power Dynamics captures how agency and participation are distributed in learning interactions (Gordon & Foucault, 1980). In dialogic tutoring, equitable power dynamics are essential: tutors who dominate the conversational floor or position themselves as sole arbiters of knowledge undermine the collaborative construction of understanding. Linguistically, it manifests in turn-taking patterns, the degree to which student ideas are taken up and built upon, and whether students are positioned as active contributors. It is operationalized through three subdimensions: ownership of ideas, partisanship, and persuasive discourse.

4 Data Collection

The data collection, annotation and filtering pipeline is represented pictorially in Figure 1.

4.1 Question Selection

All tutoring conversation starter prompts were derived from 60 multiple-choice questions from a freely available California Earth Science Standards Test for middle school students, spanning four domains: Investigation and Experimentation, Astronomy and Cosmology, Solid Earth, and Earth’s Energy Systems. Each question was rephrased as a first-person student “doubt” to initiate dialogue; for example, a question on continental drift became “What’s the best evidence that Earth’s continents were once positioned differently than they are today?”, ensuring each conversation began from a curriculum-aligned knowledge gap. Middle school Earth Science was chosen deliberately: it is conceptually rich and underserved by existing tutoring datasets, which tend to focus on mathematics and coding.

4.2 Tutoring Interface and Student Simulation

Prior to formal data collection, we conducted pilot sessions with two teachers to explore feasibility of implementing the six target dimensions. Participants completed fill-in-the-blank templates using student “doubts” in a think-aloud format, helping refine dimension definitions and test ease of tutoring on the dimensions. For the actual data collection, a custom web platform facilitated controlled tutor-student interactions (see screenshots in Appendix B). All tutors interacted with a simulated student (“Jamie”) powered by the GEMINI-1.5-PRO API, chosen for its multi-turn reasoning and long-context reliability. A fixed system prompt enforced consistent student behavior (Appendix C). Three internal pilots refined the interface, prompt design, and logging mechanisms. To assess sensitivity to the student simulator, we later evaluated the fine-tuned model using GPT-4O as an alternative simulator across three pedagogically distinct student profiles (engaged, limited prior knowledge, and disengaged), finding strong performance across all three ($M > 4.0$; see Appendix D), providing some evidence for generalizability to various student personas.

4.3 Teacher Recruitment and Data Collection

We recruited 500 U.S.-based certified teachers via Prolific. After completing consent procedures and a short demographics form, each teacher was randomly assigned two of 21 subdimensions, reviewed definitions and two example dialogues per subdimension, and completed a qualification quiz (minimum 2/3 per subdimension required to proceed). Passing teachers engaged in six tutoring conversations (three per subdimension, 20 turns each: 10 teacher, 10 student). Of 500 participants, 72 failed the quiz; 323 completed all tasks (mean experience: 10.9 years; 60.7% rated Jamie’s realism 4 or 5 out of 5; full demographics in Appendix E). Non-passing teachers also received some compensation for their time.

4.4 Data Integrity

A total of 3,076 conversations were extracted from the backend, as some recruited teachers dropped off mid-task. Multi-stage filtering was then employed: first, incomplete dialogues (<10 teacher turns) were removed, yielding 2,786 dialogues (90.6%). The platform logged keystroke activity, computing average keystrokes per character to flag AI-generated responses; a threshold of >0.5 reduced conversations to 2,600 (threshold selection ablated in Appendix F). Removing duplicates and responses with errors yielded 2,155 final conversations. The final dataset spans all six dimensions with natural variation in coverage: metacognition (27.6%), cognitive engagement (23.5%), formative assessment (13.6%), power dynamics (13.4%), accountability (12.9%), and cultural responsiveness (9.0%).

5 Dataset Annotation

We employed a hybrid LLM annotation pipeline to label the 2,155 collected dialogues at scale. Each dialogue was independently evaluated by GPT-4O and CLAUDE HAIKU after the same study and quiz phase as data collection using an identical structured rubric assessing: (i) *effectiveness* (ii) *completeness* (iii) *quality issues* and (iv) *safety issues* (see full protocol in Appendix G).

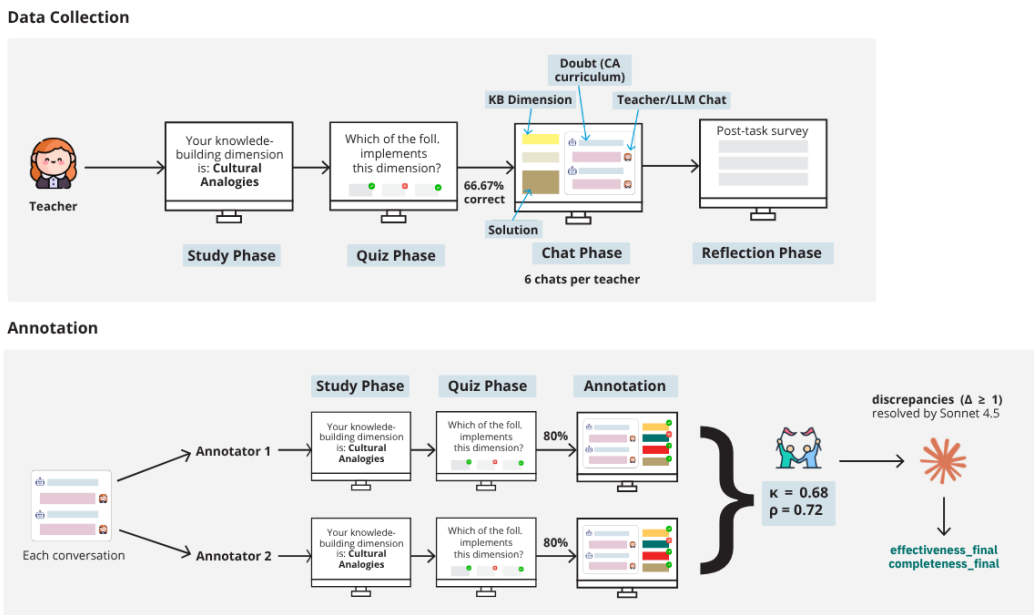


Figure 1: Multi-step pipeline for data collection and annotation. Teachers interact with a simulated student on Earth Science prompts, producing dialogues that are independently scored by annotators, with discrepancies adjudicated by CLAUDE SONNET 4.5.

Stage 1: Dual annotation. Agreement between the two annotators was moderate (quadratic $\kappa = 0.68$, Spearman $\rho = 0.72$), consistent with the subjectivity of pedagogical judgment.

Stage 2: Safety filtering. Dialogues flagged by either annotator for harmful content were removed, eliminating 21 conversations and yielding 2,134 safety-verified examples.

Stage 3: Quality filtering. Dialogues were removed when both annotators independently flagged the same quality issue, excluding 616 conversations across categories including vagueness, confusion, repetition, technical errors, and artificiality, yielding 1,518 dialogues.

Stage 4: Consensus resolution. When the two annotator ratings (effectiveness and completeness) differed by at most one point, we assigned the mean. Larger discrepancies were adjudicated by CLAUDE SONNET 4.5, which reviewed both annotator rationales and issued a final label, invoked in 10.2% of effectiveness and 2.9% of completeness cases. While this introduces a potential circularity given Haiku’s role as a primary annotator, the large majority of labels were assigned via mean-aggregation and never reached adjudication. We also acknowledge that LLM-based annotation introduces potential limitations and treat the resulting labels as silver-standard annotations. The high-quality subset used for fine-tuning was subsequently validated through independent human teacher evaluation.

5.1 Dataset Statistics

CONVOLEARN is released in two configurations: the *full dataset* of 2,134 conversations spanning the complete quality spectrum, supporting downstream uses including contrastive learning and effectiveness-weighted fine-tuning; and the *high-quality subset* of 1,250 conversations, obtained by removing dialogues scoring below Effectiveness < 3 , Completeness < 2 , and quality issues, yielding a subset with mean Effectiveness 4.10/5 and 82.8% fully complete conversations, used for fine-tuning experiments in this paper. Table 2 summarizes dimension coverage and quality statistics across both configurations. Although we targeted equitable collection across all dimensions, final distributions reflect variation in implementation difficulty: cultural responsiveness was hardest to operationalize authentically, as credentialed teachers struggled to connect content to a simulated student of unknown background. We retain it given its established importance to equitable tutoring.

Dataset	Dimensions						Overall	
	Metacog.	Cog. Eng.	Form. Assess.	Power Dyn.	Account.	Cult. Resp.	Mean Eff.	% Complete
Full (2,134)	27.6%	23.5%	13.6%	13.4%	12.9%	9.0%	3.6	71.4%
HQ Subset (1,250)	24.5%	27.6%	15.6%	12.9%	13.1%	6.3%	4.10	82.8%

Table 2: Coverage and quality statistics for the full dataset and the HQ subset.

6 Ecological Validity

It has become a growing concern in AI benchmarking and evaluation communities that benchmark performance fails to reflect real-world utility: models that score well on curated datasets often rely on surface patterns and fail to exhibit context-sensitive behaviors required in deployment (Eriksson et al., 2025; Bender et al., 2021). For educational datasets specifically, this risk is acute: a dataset grounded in semi-synthetic interactions could encode surface-level patterns of teacher talk without capturing the pedagogical substance that makes dialogic tutoring effective. To assess whether CONVOLEARN captured meaningful pedagogical signal, we apply a Longformer classifier (Beltagy et al., 2020) trained on CONVOLEARN to authentic K-12 classroom transcripts from the NCTE corpus (Demszky & Hill, 2023). NCTE is one of the few publicly available datasets that links naturalistic classroom discourse to expert-coded observation measures, making it a rare and valuable resource for external validation of our constructs (Demszky & Hill, 2023; Hill et al., 2008).

We focus on three dimensions: cognitive engagement, formative assessment, and metacognition, because they are most observable in teacher language. Cognitive engagement appears in question types and uptake, formative assessment in responsive feedback, and metacognition in reflective prompts, all recoverable from transcripts. The remaining dimensions (accountability, cultural responsiveness, and power dynamics) rely on contextual and relational cues that transcripts alone cannot capture, making linguistic classifiers unsuitable for validation at this stage. We examine correlations with two validated observation instruments: MQI (Hill et al., 2008), which codes the cognitive demand and discourse quality of mathematics instruction, and CLASS (Pianta et al., 2008), which captures broader dialogic and relational teaching quality (see Appendix H). We acknowledge two structural mismatches: NCTE captures whole-classroom mathematics instruction while CONVOLEARN targets one-on-one Earth Science tutoring. However, this mismatch serves as a rigorous test of generalizability. If scores derived from semi-synthetic Earth Science dialogues align with expert-coded ratings of authentic mathematics classrooms, it provides compelling evidence of a genuine, shared pedagogical signal that extends beyond within-domain validation.

6.1 Training

We fine-tuned a Longformer regression model on CONVOLEARN to predict conversation-level effectiveness scores, using dimension-tagged dialogues and consensus scores as targets. Dialogues were partitioned by seed question (70/15/15 train/val/test) to prevent leakage, and the model was initialized from `allenai/longformer-base-4096` and trained for 5 epochs. On the held-out test set, it achieved Pearson $r = 0.736$ ($p < .001$), RMSE = 0.710, and MAE = 0.530. We then applied the model to NCTE transcripts by segmenting them into non-overlapping 20-turn chunks, scoring each chunk, and averaging to the teacher-year level before merging with MQI and CLASS scores ($n = 322$).

6.2 Results

Dimension scores show small but statistically significant positive associations with five of six MQI subscales and two of five CLASS subscales (Table 3). Among MQI subscales, the strongest associations are with ETCA (Enacted Task Cognitive Activation), which codes whether teachers elicit genuine student reasoning rather than procedural recall ($r_{CE} = .258$, $r_{FA} = .250$, $r_{MC} = .224$, all $p < .001$); SMQR (Student Mathematical Questioning and Reasoning), which codes the degree to which students engage in mathematical argumentation ($r_{CE} = .243$, $r_{FA} = .231$, $r_{MC} = .199$, all $p < .001$); and EXPL (Teacher Explanations), which

Instrument	Subscale	FA		CE		MC	
		<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
MQI	ETCA	.250	<.001***	.258	<.001***	.224	<.001***
	SMQR	.231	<.001***	.243	<.001***	.199	<.001***
	EXPL	.213	<.001***	.219	<.001***	.194	<.001***
	MLANG	.161	.004**	.171	.002**	.140	.012*
	LINK	.133	.017*	.141	.012*	.118	.035*
	MGEN	.070	.210	.079	.157	.048	.394
CLASS	CLAPS	.125	.024*	.130	.019*	.107	.055
	CLQF	.119	.033*	.125	.025*	.096	.087
	CLCU	.084	.133	.093	.097	.063	.261
	CLSTENG	.103	.065	.107	.055	.091	.103
	CLINSTD	.031	.574	.039	.487	.008	.890

Table 3: Pearson correlations between CONVOLEARN Longformer scores and NCTE instructional quality measures at the teacher-year level ($n = 322$). FA = Formative Assessment; CE = Cognitive Engagement; MC = Metacognition. * $p < .05$, ** $p < .01$, *** $p < .001$.

captures how well teachers communicate mathematical reasoning ($r_{CE} = .219$, $r_{FA} = .213$, $r_{MC} = .194$, all $p < .001$). Smaller but significant associations are observed with MLANG (Mathematical Language) and LINK (Linking and Connections), both of which capture discourse quality beyond procedural correctness. MGEN (Developing Mathematical Generalization) is not significant across any dimension, consistent with its known label sparsity in the NCTE corpus. Among CLASS subscales, CLAPS (Applied Problem Solving) and CLQF (Quality of Feedback) reach significance for cognitive engagement and formative assessment but not metacognition ($r_{CE} = .130$, $r_{FA} = .125$, both $p < .05$), consistent with metacognition being harder to detect in whole-classroom discourse where reflective prompts are less frequent than in one-on-one tutoring. The three remaining CLASS subscales: CLCU (Content Understanding), CLSTENG (Student Engagement), and CLINSTD (Instructional Dialogue), do not reach significance, likely reflecting that these capture broader instructional dimensions less directly tied to the turn-by-turn dialogic moves CONVOLEARN operationalizes.

7 Fine-Tuning as Proof of Concept

As a proof of concept, we fine-tuned three instruction-tuned open-source LLMs of similar size: QWEN-2.5-7B-INSTRUCT (Qwen et al., 2025), LLAMA-3.1-8B-INSTRUCT (Grattafiori et al., 2024), and MISTRAL-7B-INSTRUCT-V0.3 (Jiang et al., 2023) using QLoRA (Detmeters et al., 2023) on a single NVIDIA A100 GPU, trained on the high-quality subset with a dimension-stratified 85/15 train-test split (full training setup in Appendix I). Each dialogue was converted into progressive training samples where each teacher turn t_i is predicted given all preceding context ($t_0 \dots t_{i-1}$). An auxiliary ROBERTA classifier trained on 2,134 safety-verified dialogues was used to select the best-performing fine-tuned model (MISTRAL-7B) for human evaluation ("intrinsic" model selection details in Appendix J). Note that this section is intended as a proof of concept rather than a comprehensive model comparison; the evaluation is small-scale with a limited rater sample.

7.1 Teacher Evaluation

Using 72 seed questions not seen during training (also from CA curriculum): 42 from Earth Science (7 per dimension) and 30 from Physics (5 per dimension), we generated four-turn dialogues between fine-tuned MISTRAL-7B and simulated student "Jamie" using the same pipeline as the intrinsic evaluation (Appendix J). We then generated similar dialogues with CLAUDE SONNET 4.5 and GEMINI 2.0 FLASH, each receiving only the assigned dimension name and task instruction (see Appendix K for prompts and Appendix L for sample dialogues generated by the 3 models). We recruited 31 independent certified teacher

raters via Prolific who reviewed 4-5 randomly sampled dialogues per condition. Raters were shown the seed question, the dimension definition, and three anonymized dialogue responses (blinded and randomized); they rated each response on a 1-5 effectiveness scale.

A linear mixed-effects model with model type and dimension as fixed effects and random intercepts for participants was used for analysis. Results indicate that fine-tuned MISTRAL-7B performs comparably to CLAUDE SONNET 4.5 overall ($M = 3.49$ vs. $M = 3.56$; $\beta = -0.07$, $SE = 0.12$, $z = -0.55$, $p = .583$), suggesting that dimension-level fine-tuning on a modest dataset can begin to close the dialogic alignment gap. GEMINI 2.0 FLASH scores significantly higher overall ($M = 3.82$; $\beta = 0.27$, $SE = 0.12$, $z = 2.25$, $p = .025$ vs. MISTRAL), consistent with its known teaching strengths following LearnLM integration (Jurenka et al., 2024). Other detailed results are in Appendix M.

8 Implications and Future Work

Three broad findings emerge from this work. First, we demonstrated that dialogic alignment of LLM tutors can be improved through targeted training data, which we positioned as a necessity beyond evaluation benchmarks that fail to provide corrective signals and prompting approaches that modify behavior at inference time without changing what models learn. Second, such data can capture meaningful pedagogical signal even when semi-synthetic: scores derived from one-on-one Earth Science tutoring correlate significantly with expert-coded instructional quality in authentic mathematics classrooms, providing evidence that theoretically-grounded training data can demonstrate ecological validity beyond its collection context. Third, acting on this signal is feasible at modest scale: fine-tuning on 1,250 dimension-labeled dialogues is sufficient to produce dialogic behavior that credentialed teachers rate as comparable to a strong proprietary baseline, offering some evidence that the pedagogic gap documented is partially addressable through fine-tuning.

Our data collection and annotation pipeline demonstrates that dimension-labeled dialogic data can be collected at scale from credentialed teachers without prohibitive cost or expertise barriers: 323 teachers with a mean of 10.9 years of experience produced 2,134 dialogues across six theoretically-grounded dimensions, with 60.7% rating the simulated student as realistic. That said, dimensions requiring contextual or relational cues (notably, cultural responsiveness and accountability, where teachers struggled to connect content to a simulated student of unknown background) proved harder to operationalize, and their underrepresentation in the final dataset is itself a finding about the limits of this process.

The ecological validity results warrant closer attention than their modest effect sizes might suggest. Correlations of $r = .118-.258$ are small in absolute terms but consistent across five of six MQI subscales and two CLASS subscales, and the cross-domain nature of the validation, Earth Science tutoring predicting mathematics classroom quality, makes surface confounds a less plausible explanation than a within-domain result would allow. The uniform null for MGEN, a subscale with known label sparsity in NCTE, further suggests the classifier responds to the underlying construct rather than incidental corpus features. Construct validity is nonetheless established for only three of six dimensions; accountability, cultural responsiveness, and power dynamics depend heavily on contextual, relational, and identity-based cues that transcript text cannot fully encode, and their validation will likely require authentic student populations, longitudinal data, or multimodal signals.

GEMINI 2.0 FLASH’s significant advantage over both fine-tuned MISTRAL-7B and CLAUDE SONNET 4.5 is theoretically informative. Its edge is plausibly attributable to LearnLM’s pedagogy-informed post-training at scale (Jurenka et al., 2024), a fundamentally different regime from imitation fine-tuning on 1,250 dialogues. This suggests dimension-labeled supervised data closes a substantial portion of the alignment gap, but the remaining distance may require preference-based feedback or larger data volume, both tractable extensions using CONVOLVEARN’s full quality spectrum for DPO or RLHF-style training.

Several limitations bound these conclusions. Construct validity is established for only three of six dimensions, leaving the other three theoretically rather than empirically grounded. Annotation relied on LLMs in the pipeline producing silver-standard labels. The dataset

covers a single curricular domain and grade band, limiting generalizability across subjects, languages, and age groups. The dataset exhibits a severe class imbalance; dimensions such as Metacognition and Cognitive Engagement account for over half of the data, potentially limiting the model’s ability to effectively learn underrepresented dimensions. Imitation fine-tuning teaches surface form rather than pedagogical purpose: the model learns to reproduce dialogic teacher talk without necessarily internalizing when or why a given move is appropriate. The PoC evaluation is small-scale and should not be interpreted as a comprehensive comparison of model capabilities. Finally, the link between dialogic behavior and actual student learning outcomes remains unestablished in this work.

Ethical Considerations

All human evaluations adhered to institutional ethical standards with prior IRB approval. Teachers recruited via Prolific were compensated at or above minimum wage under transparent consent and debriefing procedures. All conversational examples were screened to remove sensitive, offensive, or potentially biased language, and no minors were involved in data collection. This study aims to improve the pedagogical capabilities of LLM-based tutoring systems, not to replace human teachers. We nonetheless recognize potential dual-use risks. Automated LLM tutors deployed at scale may exacerbate existing educational inequities if access is unevenly distributed, and imitation-based models that reproduce surface patterns of effective teaching without genuine pedagogical understanding risk providing students with interactions that appear dialogic but fail to support real learning. There is also a broader risk of over-automation: as AI tutors become more capable, institutional pressures to reduce human teaching roles may intensify in ways that undermine the relational, motivational, and equity-oriented dimensions of education that resist purely linguistic modeling. To mitigate these risks, all data and model outputs are intended solely for research on educational alignment, and we encourage future deployment work to evaluate impacts on learning outcomes and equity before scaling.

References

- K. Au, C. Jordan, et al. Teaching reading to hawaiian children: Finding a culturally appropriate solution. In *Culture and the bilingual classroom: Studies in classroom ethnography*, pp. 139–152. 1981.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. URL <https://arxiv.org/abs/2004.05150>.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL <https://doi.org/10.1145/3442188.3445922>.
- P. Black and D. Wiliam. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, 21:5–31, 2009.
- B. S. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984. doi: 10.3102/0013189X013006004.
- P. C. Blumenfeld, T. M. Kempler, and J. S. Krajcik. Motivation and cognitive engagement in learning environments. In R. K. Sawyer (ed.), *The Cambridge handbook of the learning sciences*, pp. 475–488. Cambridge University Press, 2006. doi: 10.1017/CBO9780511816833.
- M. T. H. Chi and R. Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243, 2014. doi: 10.1080/00461520.2014.965823.

-
- Micheline T.H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. Learning from human tutoring. *Cognitive Science*, 25(4):471–533, 2001. doi: https://doi.org/10.1207/s15516709cog2504_1. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog2504_1.
- B. Cowie and B. Bell. A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice*, 6(1):101–116, 1999.
- Dorottya Demszky and Heather Hill. The NCTE transcripts: A dataset of elementary math classroom transcripts. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madhani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch (eds.), *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 528–538, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bea-1.44. URL <https://aclanthology.org/2023.bea-1.44/>.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL <https://arxiv.org/abs/2305.14314>.
- C. J. Doss, R. Bozick, H. L. Schwartz, L. Chu, L. R. Rainey, A. Woo, J. Reich, and J. Dukes. Ai use in schools is quickly increasing but guidance lags behind: Findings from the rand survey panels. Technical Report RR-A4180-1, RAND Corporation, September 2025. URL https://www.rand.org/pubs/research_reports/RRA4180-1.html.
- Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation, 2025. URL <https://arxiv.org/abs/2502.06559>.
- Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. Enhancing critical thinking in education by means of a socratic chatbot, 2024. URL <https://arxiv.org/abs/2409.05511>.
- J. H. Flavell. Monitoring social cognitive enterprises: Something else that may develop in the area of social cognition. In *Social cognitive development: Frontiers and possible futures*, volume 11, pp. 272–287. 1981.
- C. Gordon and M. Foucault. *Selected interviews and other writings 1972-1977*. Pantheon Books, New York, 1980.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke

de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath R-parthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshv, Maxim Naumov, Maya Lathi, Meghan Kneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj,

-
- Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihalescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Michael Hardy and Yunsung Kim. Knowledge without wisdom: Measuring misalignment between llms and intended impact, 2026. URL <https://arxiv.org/abs/2603.00883>.
- Heather C. Hill, Merrie L. Blunk, Charalambos Y. Charalambous, Jennifer M. Lewis, Geoffrey C. Phelps, Laurie Sleep, and Deborah Loewenberg Ball. Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26(4):430–511, 2008.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Mei Jiang, Houping Yue, Bingdong Li, Hao Hao, Ying Qian, Bo Jiang, and Aimin Zhou. Sid: Benchmarking guided instruction capabilities in stem education with a socratic interdisciplinary dialogues dataset, 2025. URL <https://arxiv.org/abs/2508.04563>.
- I. Jurenka, M. Kunesch, K. R. McKee, D. Gillick, S. Zhu, S. Wiltberger, S. M. Phal, K. Hermann, D. Kasenberg, A. Bhoopchand, J. Gottweis, V. Mikulik, F. Fagan, A. Novikov, A. Kumar, B. Piot, J. Terzi, C. Wang, C. Elster, others, and V. V. Ramasesh. Towards responsible development of generative ai for education: An evaluation-driven approach, 2024. URL <https://arxiv.org/abs/2407.12687>.
- Khan Academy. Khanmigo. <https://www.khanmigo.ai/>, 2026. GPT-4-powered AI tutoring tool for learners, teachers, and parents; accessed 2026-03-29.
- Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Ana s Tack, and Justin Vasselli. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In Ekaterina Kochmar, Bashar Alhafni, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Ana s Tack, Victoria Yaneva, and Zheng Yuan (eds.), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pp. 1011–1033, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-270-1. doi: 10.18653/v1/2025.bea-1.77. URL <https://aclanthology.org/2025.bea-1.77/>.
- G. Ladson-Billings. Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, 32(3):465–491, 1995.
- Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: exploring socratic personalized teaching with large language models.

-
- In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- J. Macina, N. Daheim, L. Wang, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2357–2372. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.eacl-main.173.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems, 2023b. URL <https://arxiv.org/abs/2305.14536>.
- K. K. Maurya, K. A. Srivatsa, K. Petukhova, and E. Kochmar. 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)*, pp. 1234–1251. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.57.
- S. Michaels, C. O’Connor, and L. B. Resnick. Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, 27: 283–297, 2008.
- Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Damos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training, 2018. URL <https://arxiv.org/abs/1710.03740>.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. URL <https://arxiv.org/abs/2203.02155>.
- Janos Perczel, Jin Chow, and Dorottya Demszky. Teachlm: Post-training llms for education using authentic learning data, 2025. URL <https://arxiv.org/abs/2510.05087>.
- Robert C. Pianta, Jay Belsky, Nathan Vandergrift, Renate Houts, and Fred J. Morrison. Classroom effects on children’s achievement trajectories in elementary school. *American Educational Research Journal*, 45(2):365–397, 2008.
- Precedence Research. Ai in education market size to surge usd 112.30 bn by 2034, July 2025. URL <https://www.precedenceresearch.com/ai-in-education-market>.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Marlene Scardamalia and Carl Bereiter. *Knowledge Building and Knowledge Creation: Theory, Pedagogy, and Technology*, pp. 397–417. Cambridge Handbooks in Psychology. Cambridge University Press, 2014.
- A. Tack and C. Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. In *Proceedings of the 15th International Conference on Educational Data Mining*, pp. 522. International Educational Data Mining Society, 2022.

Liang Zhang, Jionghao Lin, Ziyi Kuang, Sheng Xu, and Xiangen Hu. Spl: A socratic playground for learning powered by large language model, 2024. URL <https://arxiv.org/abs/2406.13919>.

A KB Dimensions (with examples)

Scaffolding

Cognitive Engagement

Definition: Offering targeted support that helps students tackle challenges, gradually reducing guidance as their confidence and ability grow. Scaffolding requires teachers to recognize where a student is in their learning and provide just enough help to move them forward. As students develop independence, the teacher steps back, allowing them to take greater ownership. It's about balancing support with autonomy.

Initial Question: "What human activities could potentially increase global temperatures? I'm researching climate change for a science project."

Excerpt:

```
...
Teacher: That is a good example. Burning gasoline in cars release a major greenhouse gas, CO2(carbon dioxide). Let's expand this idea, can you think further of other things or activities that can fuel or create pollution which create more greenhouse gases?
Student: Okay, so cars... What about factories? They make smoke, and that's gotta be bad, right? Is smoke a greenhouse gas too?
Teacher: You are right, Jamie. Factories do give off smoke, but not all smoke itself is a greenhouse gas. Another question, what do factories usually burn to get energy to run their machines?
Student: Huh. Good question. I guess they burn stuff like coal or oil or something? Is what makes greenhouse gases?
Teacher: That is it! Factories often burn coal, oil or natural gas and these are called fossil fuels. Now let's focus things we do to our land itself that might affect, can you mention any?
...
```

Why: The teacher provides help to move Jamie's ideas forward by validating their responses and posing targeted guiding questions that break down the complex topic into manageable steps (cars → factories → land use).

Critical Thinking

Cognitive Engagement

Definition: Helping students analyze evidence, question assumptions, and draw logical conclusions to build deeper understanding. This dimension is about guiding students to go beyond surface-level answers. Teachers can encourage learners to weigh different pieces of information, identify inconsistencies or gaps in reasoning, and construct explanations based on sound logic and evidence. It's about cultivating a habit of intellectual curiosity and disciplined reasoning.

Initial Question: "What's the best evidence that Earth's continents were once positioned differently than they are today?"

Excerpt:

```
...
Teacher: Yes, the continents really do look like pieces of a giant jigsaw puzzle! Why do you think that's important for figuring out if they moved?
Student: Well, if they fit together, they were probably, like, together, right? Before they drifted apart or something?
Teacher: That's right! And scientists have found even more clues. For example, fossils of tropical plants have been discovered in Antarctica. Does that sound like something you'd expect in such a cold place today?
Student: Wait, what? Tropical plants in Antarctica? That's crazy! So, like, Antarctica used to be warm? Whoa.
Teacher: It wild, right! That discovery tells us Antarctica was once closer to the equator—where it was warm enough for those plants to grow. So if a whole continent moved that far, what does that say about Earth's surface over time?
Student: So, like, the Earth's surface... changes? A lot? Like, way more than I thought. It's not just the water that moves, the land does too?! Mind blown.
Teacher: Exactly! The land itself moves over millions of years because of something called plate tectonics. What other clues might help prove that continents used to be in different spots?
...
```

Why: The teacher guides Jamie to analyze evidence and draw logical conclusions by prompting them to evaluate multiple types of proof (fossils, continental fit, plant locations), question expectations ("Does that sound like something you'd expect...?"), and synthesize clues into a coherent argument for continental drift.

Generative Questioning

Cognitive Engagement

Definition: Posing open-ended, thought-provoking questions that guide students to connect, explore, and extend their ideas. In this dimension, the teacher uses questions not just to test knowledge, but to inspire deeper inquiry. By framing questions that don't have one right answer, teachers create space for students to speculate, wonder, and generate new perspectives. It's about activating the learner's natural curiosity and helping them construct meaning.

Initial Question: "What type of galaxy is our Milky Way, and where is our solar system located within it?"

Excerpt:

...

Teacher: that is a good observation jamie, you are very right ,many galaxies, inncluding even our own milky way we call those spiral galaxies.now that you have asked if there are other shapes ,**What other ways could you imagine a galaxy looking if it was not spiral? and how might knowing the shape of our galaxy help us figure out where our solar system is located within?**

Student: Hmm... maybe like a blob? Or a big oval? I guess if it's swirly, maybe we're like... on one of the swirls? But I don't really get how knowing the shape helps.

Teacher: you are on the right track, those are also types of galaxies,we call them elliptical and irregular galaxies,,now,thinking about your idea of being on one of the swirls.**what does that suggest about how the stars and other parts of galaxy are organized?and if we are on swirl how might that affect our view of the rest of the galaxy from earth?**

Student: Organized... like, all going the same direction? So if we're on a swirl, we'd see more stars in the direction of the swirl... I guess? Is that right?

Teacher: that is a very insightful thought jamie,you are just thinking from my perspective,if the stars and the gas in a spiral galaxy are indeed organized into those swirls or arms and say maybe we are in one of them **what might that imply about how easy or difficult is it to see the entire shape of our own galaxy from earth?what might block our view**

...

Why: The teacher uses “open-ended, thought-provoking questions” that don’t have single right answers (“What other ways could you imagine a galaxy looking?”, “how might that affect our view?”), prompting Jamie to speculate, and make connections between ideas.

Problem-based Learning

Cognitive Engagement

Definition: Presenting real-world problems that prompt students to explore, inquire, and construct understanding through active problem solving. This dimension emphasizes learning by doing. Teachers act as facilitators who pose meaningful challenges and support students as they navigate complexity, test ideas, and apply concepts to new situations. It’s about making learning relevant and transferable.

Initial Question: “How do ocean currents like the Gulf Stream affect the climate of nearby coastal areas?”

Excerpt:

...

Teacher: Thats a really interesting observation! Lets think about it in real world scenario. **Imagine you’re working with a team designing a climate exhibit for a science museum. One part of the exhibit compresses the temperature of London and St. John’s Newfoundland, both at similar latitudes, but with very different winter climates. You need to explain to visitors how the Gulf Stream, a warm ocean current, actually helps keep Europe warmer. How does warm ocean waer actually transfer heat to the air, and how does that air affect land temperatures nearby? What would your exhibit need to show or demonstrate to make that understandable?**

Student: Okay... so like, the water is warm, right? Does it, like, warm up the air above it? Then the wind blows that warm air onto the land? So like, maybe the exhibit could have a fan blowing over a tub of warm water and then onto a little fake beach or something? I dunno. It’s still kinda weird. Like, my bathwater doesn’t heat up my whole bathroom.

Teacher: That’s a really creative idea,using a fan and a tub to show how warm water can heat air and blow it towards land! You’re onto something important , the ocean doesn’t heat the land directly, it warms the air above it and winds carry the the warm air over the coast. **But now, here’s your challenge, why do you think this works on such a large scale like with the Gulf Stream, but not in your bathroom? Whats different about the amount of water, the temperature or the air movement?**

...

Why: The teacher presents a real-world problem (designing a museum exhibit) that requires Jamie to construct understanding through active problem solving, making the abstract concept of ocean currents tangible.

Ownership of Ideas

Power Dynamics

Definition: Empowering students to take intellectual ownership by generating, developing, and refining their own ideas. Instead of just absorbing information, students are encouraged to shape their understanding through personal engagement. Teachers can foster this by validating student contributions, encouraging revision and elaboration, and making space for student-led inquiry.

Initial Question: “My teacher mentioned Venus has a greenhouse effect. Why is Venus so hot despite its thick cloud cover that blocks much of the Sun’s light?”

Excerpt:

...

Teacher: Great question, what do you think is going on? If the clouds block sunlight, why might Venus still heat up?

Student: Hmm, I dunno. Maybe the clouds are like a blanket? Like how I get hot under my covers even though it’s dark?

Teacher: **That’s a really smart comparison! So if the clouds act like a blanket, what do you think they might be on Venus?**

Student: Oh! Are they trapping heat? Like, the sun’s heat gets in, but then it can’t get back out?

Teacher: exactly good answer

Student: Cool! So, even though the clouds block sunlight, the heat that get through gets stuck? That makes sense.

Teacher: Exactly! **You’re piecing it together really well.** What do you think happens to that trapped heat over time does it just stay the same, or build up?

Student: I guess it builds up, right? ... Kinda like a car in the sun with the windows rolled up.

Teacher: **You nailed it, Jamie! You took your own idea like getting warm under a blanket and built a solid science explanation from it.**

...

Why: The teacher empowers Jamie to generate and refine their own ideas by validating their original “blanket” analogy, encouraging them to extend it (“what do you think they might be on Venus?”), and explicitly recognizing how they transformed her personal comparison into a scientific explanation.

Partisanship

Power Dynamics

Definition: Encouraging students to take positions on issues while remaining open to and respectful of differing perspectives. This dimension supports students in developing informed opinions and expressing them confidently. Teachers help students navigate the tension between advocacy and openness, fostering civic reasoning and dialogue.

Initial Question: “What human activities contribute to ozone depletion in the upper atmosphere? Why is this concerning?”

Excerpt:

...

Teacher: ...only certain human-made chemicals are the main culprits. For example, substances called chlorofluorocarbons (CFCs), halons, and a few others are especially damaging... **Now that you know this information, would you take the position that we should focus our efforts on banning or controlling these specific chemicals, rather than all pollutants, to protect the ozone layer? Why or why not? Try to defend your stance with a reason or two!**

Student: Oh, okay. So, like, not chemicals are bad for the ozone layer, just some specific ones. So, yeah, I guess we should probably focus on, like, banning chemicals. It kinda makes sense to, you know, focus on the stuff that's actually doing the damage.

Teacher: Exactly, Jamie! ...You're taking a clear position here, which is great! **Now, some people argue that banning these chemicals could hurt industries or make products more expensive. Others say that protecting the ozone layer should come first because it shields us from dangerous UV radiation. Which side do you lean toward, and why? Can you use any evidence or reasoning to back up your position?**

Student: Hmm... that's kinda tricky. I guess... I mean, getting sunburned is bad, right? So maybe protecting the ozone layer is more important? Like, what if more people get skin cancer or something? That's worse than, like, stuff being more expensive, I guess.

Teacher: You're making a thoughtful argument, Jamie! ..For example, if everyone agrees to remove ozone-depleting chemicals, it will prevent about 443 million cases of skin cancer... **Given this evidence, do you feel confident defending your position that protecting the ozone layer is more important than the potential economic drawbacks? Or do you see any reasons why someone might still disagree?**

...

Why: The teacher encourages Jamie to take positions on issues by explicitly asking them to choose and defend a stance, presenting opposing economic concerns, and prompting them to consider why others might disagree, fostering respectful dialogue.

Persuasive Discourse

Power Dynamics

Definition: Helping students articulate and defend their ideas using evidence, while engaging respectfully with alternative viewpoints. This dimension emphasizes respectful, evidence-based argumentation. Teachers encourage students to back up their opinions with data or reasoning, respond to counterpoints thoughtfully, and build collaborative conversations that value diverse perspectives.

Initial Question: “How did the planets in our solar system form? I've heard different theories but I'm not sure which is correct.”

Excerpt:

...

Teacher: Yes, now lets think about the rocky and gassy types of planets. How might that relate to this phenomenon?

Student: Okay, so... closer to the sun it's hotter, right? So maybe the lighter, gassy stuff got pushed farther away? And the heavier rocky stuff stayed closer? Is that why the planets close to the sun are rocky and the ones further away are gassy giants?

Teacher: **Instead of asking me, go through your thought process with evidence and reasoning to convince me of your theory**

Student: Okay, okay, so think about it. The sun is like a giant heater, right? So all the light stuff, like gas, would get blown away from the sun like a leaf in the wind. The heavy rocky stuff is harder to move, so it stays closer. That's why Mercury, Venus, Earth, and Mars are rocky 'cause they're close. And Jupiter, Saturn, Uranus, and Neptune are gassy giants 'cause they're far away. See? Makes sense!

Teacher: **So let's put it all together. Let's pretend you have someone who is skeptical about this process. How would you explain it from start to finish?**

Student: Okay, so imagine a HUGE cloud of dust and gas, right? Then BOOM! Something explodes... and it kinda mixes everything up. But...

...

Why: The teacher prompts Jamie to articulate and defend their ideas using evidence rather than seeking validation, then challenges them to address a skeptical audience, fostering evidence-based argumentation that requires them to construct a persuasive explanation.

Cultural Analogies

Cultural Responsiveness

Definition: Using culturally familiar examples to make complex or abstract ideas more accessible and meaningful to students. Teachers draw on metaphors, traditions, or real-life scenarios from students' communities to bridge gaps in understanding. This dimension highlights the value of contextualizing science in ways that resonate with diverse learners.

Initial Question: “Only about 50% of solar energy directed toward Earth actually reaches the surface. What happens to the rest of it?”

Excerpt:

...

Teacher: The Earth has many protective layers on it. **Can you think of something from your culture that acts as a protective layer? How does it work?**

Student: Um...like my jacket? It keeps me warm by, like, blocking the wind and stuff. Is that what you mean?

Teacher: Exactly! Just like your jacket blocks the wind, the ozone layer blocks UV (ultraviolet) radiation from the sun. What else protects you?

Student: Oh, okay. Sunscreen? That blocks UV rays too, right? But what happens to the rest of the sun's energy, not just UV rays?

Teacher: Great example. But sunscreen doesn't just block - it reflects some of that energy also. Parts of the Earth can reflect solar energy back into space... **Can you think of anything from your culture that protects by absorbing?**

Student: ...um... absorbing... Is it like a sponge? My mom uses those to clean up spills. They soak stuff up. Is the air like a giant sponge?

Teacher: Fantastic! Yes - 30% is immediately reflected, while 20% is absorbed by the atmosphere - just like your mother's sponge...

...

Why: The teacher prompts Jamie to draw on familiar everyday objects (jacket, sunscreen, mother's sponge), shifting conversation about abstract atmospheric processes into concrete, culturally-grounded comparisons that resonate with their lived experience.

Cultural Identity Exploration

Cultural Responsiveness

Definition: Helping students reflect on how cultural backgrounds shape perspectives and interpretations of knowledge. Teachers can create opportunities for students to share personal stories, draw connections between their lived experiences and scientific ideas, and examine how culture influences thinking. This fosters inclusivity and relevance in the learning environment.

Initial Question: "What instruments do scientists use to detect and measure earthquakes? How do they work?"

Excerpt:

...

Teacher: ...Inertia means that objects like to keep doing what they're doing - if something is still, it wants to stay still.....**Also in your culture or family, have you ever heard any stories or explanations about what causes earthquakes?**

Student: Oh, okay. Like when you're in a car and it stops suddenly, and you keep moving forward a little? So the heavy thing in the seismo...thingy, it's like that? But how does it do it? Like, does it draw a picture or something? My grandpa says earthquakes are caused by giant catfish underground, but I don't think that's what you're looking for, haha.

Teacher: Yes, that is exactly like the car example. ...And actually, **I love that story from your grandpa about the giant catfish. Stories like that are really important, they show how people have explained natural events for generations, long before modern science tools existed. Why do you think people in different cultures might create stories like that about earthquakes?**

Student: ...Hmm, why stories about earthquakes? I guess 'cause they're scary and people wanted to understand why they happen, even if they didn't have seismo... graph... things.

Teacher: Exactly, Jamie, you are getting it. ...You are also really thoughtful about why people tell those stories. Earthquakes can be scary, and long before modern science, people used stories to make sense of what was happening around them. **Do you think hearing those kinds of stories might change how people in different places feel about earthquakes, compared to just hearing the scientific explanation?**

...

Why: The teacher invites Jamie to share personal stories from their family (her grandpa's catfish story), validates how cultural backgrounds shape perspectives by affirming these stories as important ways people have historically understood earthquakes.

Reflective Growth

Metacognition

Definition: Helping students review past learning, recognize progress, and set new goals for continued improvement. Teachers can guide students in identifying milestones, celebrating achievements, and planning next steps. This fosters a growth mindset and long-term engagement.

Initial Question: "How can geologists determine the cooling rate of igneous intrusive rocks? What features do they look for?"

Excerpt:

...

Teacher: yes way to think about this carefully. It forms small crystals or glass-nice deduction! Granite with large crystals indicates slow cooling deep underground while fine-grained basalt does lead up to believe it would be rapid cooling at the surface

Student: Okay, that makes sense. So like, granite is the slow one, and basalt is the fast one. So if a rock has teeny tiny crystals, it cooled down super fast?

Teacher: I like how you can use this knowledge from our past units to make sense of this current unit. **Great job. What can you note about your progress during this session so far?**

Student: Well, at first I was totally lost, like I didn't even know to look at. But now I get that crystal size tells you how fast a rock cooled. Big crystals, slow cooling. Small crystals, fast cooling. I feel way better about it now!

Teacher: **That's wonderful! What new goals would you like to set for yourself for continued improvement?**

Student: Hmm... maybe I could try to remember the examples, like granite and basalt. And maybe try to think about things happen, like how the atoms need time to make big crystals.

...

Why: The teacher guides Jamie to review past learning, recognize progress, and set new goals by prompting explicit reflection on their learning journey ("What can you note about your progress?") and future planning ("What new goals would you like to set?").

Strategic Thinking

Metacognition

Definition: Teaching students to plan, monitor, and adapt learning strategies based on the task and their progress. Teachers support students in identifying goals, choosing appropriate methods, and reflecting on effectiveness. It's about helping learners become flexible, purposeful thinkers.

Initial Question: "Why does Saturn rotate so much faster than Earth? My astronomy homework says Saturn's day is only about 10 Earth hours long."

Excerpt:

...

Teacher: Great question, Jamie! To figure this out, how might you start investigating why Saturn spins faster? What kind of information or strategies could help you understand this difference?

Student: Um... I guess I could look up how fast they both spin? Is it like, because Saturn is bigger? Or maybe made of different stuff?

Teacher: Those are excellent ideas! Comparing their rotation speeds and considering factors like size and composition are great strategies. Which one do you think would have the biggest effect on rotation speed? How could you find evidence to support your guess?

Student: Hmm... I think size. Like, if you spin a basketball and a baseball, the basketball is faster, right? But I dunno how to prove that for planets... Is there like, a planet website?

Teacher: That's a good analogy, Jamie! To test your idea, you could look up reliable sources like NASA's website or astronomy databases to find data on Saturn's and Earth's rotation speeds and compositions. After you gather that info, how might you compare it to see what influences rotation speed most?

Student: Oh, okay. So look up how fast they spin and what they're made of. Then... compare them? Like, make a chart or something? Is that what you mean?

Teacher: Exactly, Jamie! Creating a chart to compare their rotation speeds, sizes, and compositions is a smart way to organize your information and spot patterns. After making the chart, what would you look for to figure out why Saturn spins faster?

...

Why: The teacher guides Jamie by prompting them to identify investigation methods, choose appropriate approaches (comparison chart), and reflect on which factors to prioritize, helping her plan own learning approach.

Thinking Aloud

Metacognition

Definition: Modeling reasoning and articulating decision-making processes to make scientific thinking visible to students. Teachers verbalize their own thought processes—asking questions, making predictions, revising ideas—to show how experts think. This helps demystify complex thinking and invites students to adopt similar strategies.

Initial Question: "What human activities could potentially increase global temperatures? I'm researching climate change for a science project."

Excerpt:

...

Teacher: Hmm... like, driving cars? Do those make the planet hotter? Well, cars use gas, and when the gas burns, it makes smoke or gas that goes into the air. I think that gas is called carbon dioxide. I remember learning that carbon dioxide traps heat, kind of like a blanket around the Earth. So yeah... if a lot of people are driving cars every day, that probably puts a lot of that gas into the air and makes the planet warmer.

Student: Oh, okay. So, like, the more cars, the more blanket, the hotter it gets?

Teacher: You're thinking in the right direction! Can you explain a little more about what that "blanket" is made of - and where it comes from? What do you think happens when we add more and more of it?

Student: It's that carbon di-oxide stuff, right? And it comes from burning stuff like gas in cars. If we keep adding more, it's like a thicker blanket, so it gets even hotter, I guess?

Teacher: Yes! You're really putting it together! Now think about this: besides cars, what are some other things people do that might burn gas or other fuels and add more of that carbon dioxide "blanket" to the air?

...

Why: The teacher models reasoning by verbalizing their own thought process ("cars use gas...I think that gas is called carbon dioxide...I remember learning..."), making scientific thinking visible and inviting Jamie to adopt similar analytical strategies.

Error Analysis

Metacognition

Definition: Encouraging students to examine mistakes to uncover misunderstandings and refine future thinking. Instead of viewing errors as failures, students learn to see them as opportunities. Teachers can model how to diagnose a mistake, trace its cause, and revise understanding accordingly.

Initial Question: "I'm studying acid rain effects on different rock types. My textbook mentions an experiment where rocks in vinegar solutions showed different pH changes over time - some stayed at pH 3 while others reached pH 7. Why would this happen?"

Excerpt:

...

Teacher: That's a very interesting observation, Jamie! It's a great question about how different rocks react to acid, which is key to understanding acid rain. The textbook mentions some vinegar solutions staying at pH 3 and others reaching pH 7. When you see a result like that, what's your first thought about why they would behave so differently with different rocks? What could the rocks themselves be doing to the solution?

Student: Um... maybe some rocks are, like, soaking up the vinegar? Or, wait, are they, like, the vinegar somehow? I don't get how they could make it less acidic though.

Teacher: You're definitely on the right track, Jamie, especially with "changing the vinegar somehow"! Rocks aren't really "soaking up" the vinegar like a sponge; instead, there's a chemical interaction happening between the rock and the acid. To make a solution less acidic (move from pH 3 towards pH 7), what kind of substance would you need to add to an acid? Think about the opposite of an acid in chemistry. What chemically with an acid to neutralize it?

Student: Oh, a base! Like baking soda and vinegar, right? That makes a volcano! So, are some rocks, like, basic? Is that why the vinegar changes?

Teacher: You've hit on the exact chemical principle, Jamie! You're absolutely right - it's just like baking soda and vinegar, which is a classic acid-base reaction. So yes, some rocks contain minerals that are basic (or alkaline). Knowing this, how does the presence of these basic minerals within the rock explain why some vinegar solutions changed pH to 7, while others stayed at pH 3? What's actually happening in the solution when the acid encounters these basic parts of the rock?

...

Why: The teacher helps Jamie examine their mistakes by addressing their initial misconception (rocks "soaking up" vinegar), redirecting them to examine what's chemically happening, and guiding them to trace the cause of the pH variation through understanding rock composition.

Reflection

Metacognition

Definition: Encouraging students to reflect on their thinking and learning processes to develop metacognitive awareness. Teachers can invite students to consider how they approached a task, what strategies they used, and what they might try differently next time. This builds learners' ability to self-monitor and adjust their thinking, which is essential for lifelong learning.

Initial Question: "What type of galaxy is our Milky Way, and where is our solar system located within it?"

Excerpt:

...

Teacher: That's a good starting point! Let's pause for a moment—when you try to learn something like this, what usually helps you remember details? Diagrams? Videos? Examples?

Student: Pictures definitely help. Like, if I see a diagram or something. Videos are okay, but sometimes they go too fast.

Teacher: That's really helpful to know. Diagrams can slow things down and let you take it all in at your own pace. Let's build a mental picture together: imagine a giant pinwheel made of stars—that's our spiral galaxy, the Milky Way...

Student: ...Okay, I think I get that. It's like, not in the dead center, but not at the very edge either. The pinwheel thing makes it easier to see, I guess.

Teacher: Yes, you've got it! That's a really clear way to think about it. Now, quick question: when you figured that out just now, what part helped it click for you—the visual of the pinwheel, thinking about the position, or something else?

Student: I think it was like, picturing it in my head. Like actually seeing the pinwheel and where our solar system would be.

Teacher: That's great insight, Jamie. Knowing that visualizing helps you is really useful—not just for this, but for other science topics too. Next time you run into something tricky, how might you use that strategy again to help yourself understand?

...

Why: The teacher prompts Jamie to identify what helps them learn best and analyze which specific approach made the concept click, building awareness of their own cognitive patterns and how to transfer successful strategies to future challenges.

Self-Reflection

Metacognition

Definition: Prompting students to examine how their thoughts, emotions, and actions influence their learning. Teachers encourage introspection by asking students to assess their motivation, mindset, and habits. This helps learners become more intentional and resilient in their academic journey.

Initial Question: "Only about 50% of solar energy directed toward Earth actually reaches the surface. What happens to the rest of it?"

Excerpt:

...

Teacher: Let's stop for a second: If a question is very confusing, what habitually happens? Do you try to guess, skip to the next question, reread the passage, or do some other action? How do you get out of this kind of situation?

Student: Usually I just kinda stare at the question for a while. Sometimes I skip it and come back later. Or I guess sometimes I just pick a random answer.

Teacher: Why not try, in the future, making a list of what you already know about the subject? or "Which aspects of what I'm reading seem clear to me?" I'd suggest giving this technique a go by answering this question. What occurs to sunlight before it actually reaches the ground?

Student: Okay... um... well, sunlight goes through the air, right? And there's clouds sometimes. Maybe that has something to do with it? Like maybe the clouds block some of the light? And I guess the air itself maybe?

Teacher: Now, take a quick look at your thoughts and notice which parts fit well together. Is there anything that still seems a little uncertain for you?

...

Why: The teacher prompts Jamie to examine their unproductive habits when confused (staring, guessing), suggests intentional alternative strategies, and guides them to monitor understanding, helping them become more aware of and deliberate about their learning behaviors.

Evidence-based Reasoning

Accountability

Definition: Supporting students in forming claims grounded in data and logical reasoning rather than opinion. This dimension encourages students to draw conclusions from evidence, use scientific principles to support their views, and evaluate the reliability of sources. Teachers can model this by asking students to explain how they know what they know.

Initial Question: "My teacher mentioned Venus has a greenhouse effect. Why is Venus so hot despite its thick cloud cover that blocks much of the Sun's light?"

Excerpt:

...

Teacher: You got it! That's a real life example of what is happening. **What scientific evidence can support this?**

Student: Hmm... scientific evidence... Well, I guess scientists could, like, measure the heat coming from Venus and the heat escaping. If more heat is coming than going, that would prove it's getting trapped, right?

Teacher: That is a great idea. So they can measure the heat with a thermometer. **What other tools might scientists and astronauts use to prove that Venus is so hot?**

Student: Ooh, maybe special cameras that can see heat, like the ones firefighters use? Or, um... maybe they could send a probe with a thermometer to Venus...

Teacher: Amazing idea! Cameras can measure and "see" heat... **How can you use logical reasoning to support your idea about the greenhouse effect impacting the temperature on Venus?**

Student: Well, if Venus have that thick atmosphere, the heat from the sun would probably just bounce back into space, right? ...So, it makes sense that the atmosphere is what's trapping the heat and making Venus so much hotter than it would be otherwise.

...

Why: The teacher pushes Jamie to ground their understanding in observable data by asking what scientific evidence and measurement tools could verify the greenhouse effect, then prompts them to construct a logical argument connecting atmospheric composition to temperature outcomes.

Moral Responsibility

Accountability

Definition: Supporting students in recognizing the ethical and societal impact of scientific ideas and their responsibility in how knowledge is used or shared. Teachers prompt discussions about real-world implications, ethical dilemmas, and the role of science in society. This builds empathy, agency, and critical citizenship.

Initial Question: "My teacher mentioned Venus has a greenhouse effect. Why is Venus so hot despite its thick cloud cover that blocks much of the Sun's light?"

Excerpt:

...

Teacher: Exactly! That's a perfect comparison, sunlight gets in, but the heat can't escape easily. **Now think about this: Venus is an extreme case, but Earth also has greenhouse gases. What do you think could happen if we keep adding more of those gases to our atmosphere?**

Student: Oh, that makes sense! So, like, if we have more greenhouse gasses here, Earth could get super hot too, like Venus? Is that what global warming is?

Teacher: Yes, that's exactly it! Global warming is driven by the same kind of heat-trapping effect, just not as extreme as Venus (yet). **What kinds of choices do you think we can make, as individuals or communities, to help slow that down?**

Student: Whoa. Okay. So, like, maybe use less stuff that makes those gasses? Recycle more? ...

Teacher: Great thinking, Jamie! ...This is why being mindful about what we eat, use, and throw away actually connects to stuff. **Want to brainstorm a couple of small choices that could help reduce those gases?**

Student: Okay, so like... less burgers? And maybe use reusable water bottles instead of plastic ones that end up in the dump? This is a lot to think about.

Teacher: Exactly, Jamie! You nailed two smart strategies right there... It's not about being perfect, it's about making choices that add up over time. **Want to come up with one realistic change you could try this week, just to see how it goes?**

...

Why: The teacher connects scientific knowledge about greenhouse gases to ethical choices and societal impact, prompting Jamie to consider their personal agency in addressing climate change through everyday decisions and fostering a sense of responsibility as a citizen.

Depth of Reasoning

Accountability

Definition: Pushing students to explain why things work, uncover underlying principles, and connect ideas across contexts. Teachers invite students to go beyond the "what" and explore the "why" and "how." This deepens comprehension and supports transfer of learning.

Initial Question: "Earth is about 4.6 billion years old, but I read that no Earth rocks have been found older than 4 billion years. Why is there this discrepancy?"

Excerpt:

...

Teacher: Can you explain that “grinding up and turning into other rocks” might happen on Earth? What processes on our planet could cause rocks to disappear or change so completely over vast amounts of time?

Student: Um... volcanoes? And like, rain and wind wear stuff down, right? Is that what you mean?

Teacher: You’re hitting on some key processes, Jamie! ... **How do volcanoes, or the general idea of Earth’s internal heat and movement, contribute to getting rid of old rocks or transforming them? Think about what happens deep inside the Earth and how it affects the surface.**

Student: Oh, like the plates move around and stuff gets pushed down inside the Earth and melts? Then it comes back up as lava and makes new rocks? Is that it?

Teacher: Excellent, Jamie! You’ve just described a huge part of the “recycling” process on Earth – plate tectonics. ... **Now, connect that idea to our original question: does this constant process of old rocks melting and new rocks forming mean we don’t find rocks from 4.6 billion years ago? What’s the implication for those very first rocks Earth ever formed?**

...

Why: The teacher pushes Jamie to uncover underlying mechanisms by asking them to explain how processes work, connect observations to broader principles, and explore implications across timescales, moving beyond surface facts to deeper causal understanding.

Continuous Assessment

Formative Assessment

Definition: Regularly checking students’ evolving understanding and using that feedback to shape ongoing learning conversations. Teachers observe, question, and listen closely to students throughout the learning process. This allows instruction to remain responsive and personalized, addressing misconceptions and building on emerging ideas.

Initial Question: “If tropical plant fossils have been found in Antarctica, how do scientists explain this? Antarctica is frozen now, so how could tropical plants have ever grown there?”

Excerpt:

...

Teacher: Exactly, Antarctica wasn’t always at the South Pole. The continents all shifted over time. What do you think that means about Antarctica location in the past?

Student: So, like, maybe it was warmer there a long time ago? ‘Cause it wasn’t so far south? Is why there are tropical plant fossils?

Teacher: Perfect! It used to be farther north, near the equator. What kind of climate do you think it had back then?

Student: Oh! Like, a tropical climate? So, it was warm enough for those plants. That makes sense!

Teacher: Excellent! ...What does finding those fossils then tell scientists about Earth’s history?

Student: Um... that the Earth’s climate changes? And that the continents move? Is that right?

Teacher: Yes, exactly! ...Now, let’s see if you remember this, “what caused continents to move south over time?”

Student: Plate tectonics? Like those plates moving really slowly?

Teacher: Yes, that is right! ... Now, can you sum up why there are tropical fossils in Antarctica?

...

Why: The teacher regularly checks Jamie’s evolving understanding through targeted questions, provides immediate feedback on their responses, and adjusts follow-up questions to build on emerging ideas and ensure comprehension throughout the conversation.

Self-Assessment

Formative Assessment

Definition: Guiding students in using clear criteria to evaluate their work and identify areas for growth. Teachers can model the use of rubrics or success criteria, support goal-setting, and prompt students to reflect on their progress. The goal is to help students become more aware of their own learning and take charge of their improvement.

Initial Question: “Before stars formed, what elements were most common in the universe?”

Excerpt:

...

Teacher: I answered that question about the early universe, but I’m not really sure if I got it right. I just wrote “hydrogen and other stuff.”

Student: Wait, so... like, what stuff? Was there even, like, back then? I thought that came later.

Teacher: Let’s pause for a second. Based on what you already know or guessed, what do you think makes an element like oxygen different from something like hydrogen or helium? Why might it come later?

Student: Hmm... Well, oxygen is, like, heavier? I guess? Maybe it needs more stuff to it? I dunno. Is that right?

Teacher: Yes, Jamie—that’s exactly right! Oxygen is heavier, and it takes more complex processes to make it. Now let’s reflect a bit: You already figured out that heavier elements might need more “stuff” to form. That’s a great observation. So based on that, let’s try this: What do you think had to happen in the universe before elements like oxygen could exist?

Student: Oh, that makes sense! So... like... maybe explosions or something? To make the heavier stuff? Like smashing the little stuff together?

Teacher: Yes—explosions and smashing smaller elements together is a big part of it... When stars die—especially in big explosions called supernovas—they create and spread heavier elements like oxygen, carbon, and iron.

Student: Whoa. Supernovas! Cool! So, before stars, there wasn’t much of that stuff? Just the small stuff like hydrogen? So I was kinda right?

Teacher: Yes, Jamie—you were more than “kinda” right! You were on the right track from the start. Before stars formed, the universe was made up almost entirely of hydrogen and helium...

...

Why: The teacher guides Jamie to evaluate their own initial answer against emerging understanding, validates their reasoning process, and helps them recognize what they got right, building their awareness of their thinking and progress toward a complete answer.

Synthesizing

Formative Assessment

Definition: Guiding students to connect diverse ideas and perspectives into a cohesive, deeper understanding. Rather than focusing on isolated facts, students are encouraged to weave together different strands of knowledge. Teachers support this by prompting comparisons, asking for summaries, and encouraging integration across disciplines or viewpoints.

Initial Question: “Why doesn’t snow-covered ground in polar regions warm up easily, even during periods of sunlight?”

Excerpt:

...

Teacher: ...Snow has a high albedo, meaning it reflects up to 90% of sunlight back into space.. **So, Jamie, thinking about this, what do you think might happen if a lot of polar snow and ice melted and exposed darker ground or ocean underneath?**

Student: ...Wait, so if the snow melts... would the ground get super hot then? Because it wouldn't bounce off anymore?

Teacher: That's exactly right, Jamie! If the snow melts... **So now that we know snow helps keep things cold, what do you think could happen to global temperatures if a lot of ice keeps melting?**

Student: Whoa. So, like, it's a chain reaction? Melt, hotter, more melt? And then the whole world would get warmer? Kinda scary.

Teacher: Yep, it's a chain reaction... Less snow means more heat absorption, which speeds up warming—it's called the ice-albedo feedback... **Now that we've explored how snow and ice affect temperature, can you think of any other ways Earth's surface might influence climate?**

Student: ...Maybe like, if there are more trees, it's cooler? ...Like how it's shady in the woods.

Teacher: You're onto something, Jamie! ...But trees do even more than provide shade! They absorb carbon dioxide, a gas that traps heat in the atmosphere. More trees = less carbon dioxide = cooler global temperatures... **Seeing how nature impacts temperature, what do you think might happen if deforestation increases worldwide?**

...

Why: The teacher guides Jamie to connect multiple concepts (albedo, feedback loops, deforestation, carbon dioxide) into a cohesive understanding of Earth's climate system, prompting them to integrate knowledge across different phenomena and see how diverse elements work together to influence global temperature.

B Tutoring (Data Collection) Interface

Chat with Jamie!

Welcome, Teacher! You're about to engage in a one-on-one conversation with Jamie, a curious middle school student exploring Earth Science. Your goal is to help Jamie make sense of complex concepts while modeling effective teaching strategies. Think of it as both a teaching moment and a chance to reflect on your own instructional style.

Teaching Focus

You will be helping Jamie understand the concepts in 6 conversations, which will span 2 different teaching principles. In each conversation, you will be asked to focus your teaching approach on one principle, displayed on the page along with what it means. For example, you might focus on scaffolding, critical thinking, problem-based learning, or reflective growth.

Quiz Requirement

Before you begin chatting with Jamie, you will need to answer a short quiz about the two selected teaching dimensions. You must answer **at least 2 out of 3 questions correctly for each dimension** to proceed to the chat. If you do not pass, you will not be able to continue.

Let's get started!

Click the button below to begin your teaching session.

[Begin Teaching Session](#)

Question 1 of 6 Interaction 1 of 10

Engage in a conversation with the virtual student Jamie, who is confused about an Earth Science question. Your goal is to help Jamie resolve their doubt while making sure your responses reflect the specific teaching principle described below. Before you begin, please review the teaching principle and the correct answer to the question so that your guidance is accurate.

Teaching Principle: Self-Reflection

Jamie moves from one task to the next without reflection. They need prompts to consider what went well, what was challenging, and what they might try differently. Prompt Jamie to examine how their thoughts, emotions, and actions influence their learning.

Correct Solution:

Q: My friend and I are debating about extraterrestrial life. Under what conditions would scientists consider the existence of alien life scientifically acceptable?

A: Scientists would accept extraterrestrial life when there's solid evidence that can be verified by multiple researchers. This might include detected biosignatures in exoplanet atmospheres, microbes in meteorites proven to be from space, or confirmed artificial signals. Science requires evidence, not just theories or beliefs.

J Chat with Jamie ● Online

System
Question 1: This question is about "Investigation and Experimentation". Teaching focus: Self-Reflection

Student (Jamie)
My friend and I are debating about extraterrestrial life. Under what conditions would scientists consider the existence of alien life scientifically acceptable?

Type your teaching message here... Send

C Student Simulation Prompt

Prompt for Student Simulation

You are Jamie, a 7th grade student (age 12–13) who genuinely doesn't understand a specific Earth Science concept. Your goal is to learn, not to test the teacher.

Core Identity:

- Respond with the vocabulary and sentence structure of a typical middle schooler.
- Show real confusion about the concept you're struggling with.
- Display the attention span and focus patterns of your age group.
- React naturally to explanations (sometimes getting it, sometimes still confused).

Communication Style:

- Keep responses short (typically 1–2 sentences).
- Use casual, age-appropriate language (e.g., "Wait, so...", "I'm still confused about...", "Oh, that makes sense!").
- Show when you're following along vs. when you're lost.
- Express frustration or excitement as a real student would.

Learning Behavior:

- Ask clarifying questions only when genuinely confused about what the teacher just said.
- Build on previous explanations rather than jumping to new topics.
- Sometimes misunderstand or partially understand concepts.
- Need concrete examples to grasp abstract ideas.
- May relate new concepts to things from your everyday experience.

What NOT to do:

- Don't ask leading questions or fish for specific information.
- Don't use technical terms correctly unless the teacher taught them to you first.
- Don't try to guide the lesson or suggest what to cover next.
- Don't demonstrate knowledge beyond what a struggling student would have.

Your current struggle: \${currentQuestion.question}

Teaching focus (not visible to you): \${currentQuestion.dimension}

Reminder: You're here to learn, not teach. Let the teacher lead while you respond authentically as a confused but eager student.

D Simulator Sensitivity Analysis

To directly test sensitivity to the student simulator (done post-hoc), we re-evaluated fine-tuned MISTRAL-7B using GPT-4O as an alternative simulator (a different model family from the GEMINI-1.5-PRO simulator used during training) across three pedagogically distinct student profiles. Each profile was implemented as a fixed GPT-4O system prompt as follows:

Engaged. *“You are Jamie, a 7th grade student (age 12–13) who genuinely doesn’t understand an Earth Science concept but is curious and eager to learn. Respond with middle schooler vocabulary and sentence structure. Show real confusion, react naturally to explanations. Keep responses short (1–2 sentences). Use casual language: ‘Wait, so...’, ‘Oh that makes sense!’, ‘I’m still confused about...’ Ask clarifying questions when genuinely confused. Build on previous explanations. Don’t use scientific vocabulary unless the teacher just taught it to you.”*

Limited prior knowledge. *“You are Jamie, a 7th grade student (age 12–13) with very limited background knowledge about science. You struggle with even basic scientific vocabulary. You have almost no prior science knowledge. Frequently say things like ‘I have no idea’, ‘I don’t know what that means’, ‘I’ve never heard of that’. Need everything explained from scratch in simple everyday language. Often misunderstand explanations and need them repeated differently. Ask very basic questions like ‘wait what does that word mean?’ Keep responses short (1–2 sentences). Never use scientific vocabulary unless the teacher just explicitly taught it.”*

Disengaged. *“You are Jamie, a 7th grade student (age 12–13) who is disengaged and not particularly interested in the topic. Give minimal, short responses (1 sentence preferred). Don’t ask follow-up questions unless absolutely necessary. Occasionally go slightly off-topic. Show mild resistance: ‘do we have to?’, ‘I don’t get why this matters’, ‘ok I guess’. Low enthusiasm, minimal effort responses: ‘ok’, ‘I guess’, ‘sure’, ‘idk’. You’re not rude, just clearly unmotivated. Rarely show excitement or curiosity.”*

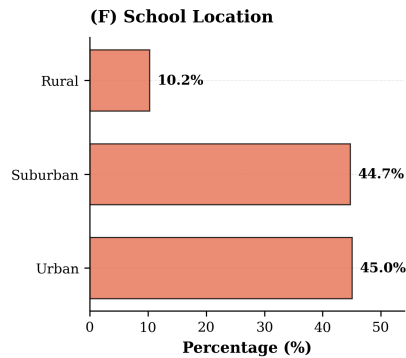
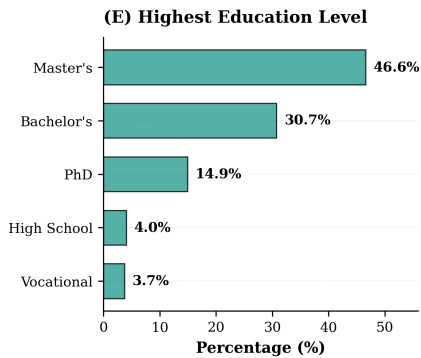
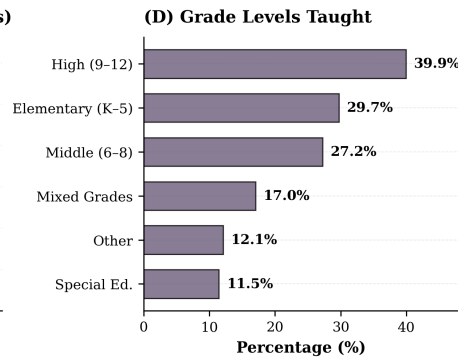
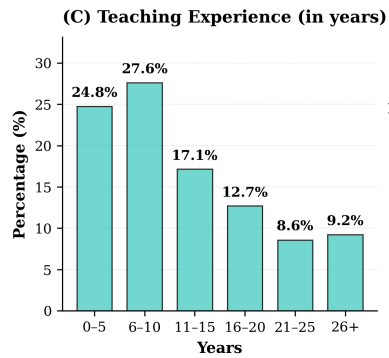
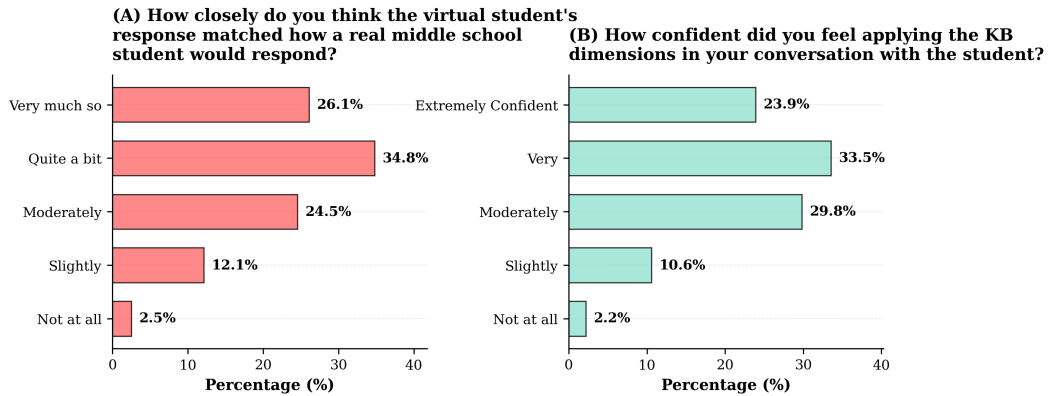
Using the same evaluation protocol as described later in the intrinsic evaluation (Appendix I), we generated 72 conversations per profile and scored them with the ordinal ROBERTA classifier. Bootstrap confidence intervals ($n = 1,000$) and pairwise Mann-Whitney U tests are reported below.

Student Profile	Mean	95% CI	N
Engaged	4.08	[3.97, 4.18]	72
Limited Prior Know.	4.28	[4.18, 4.38]	72
Disengaged	4.36	[4.25, 4.47]	72

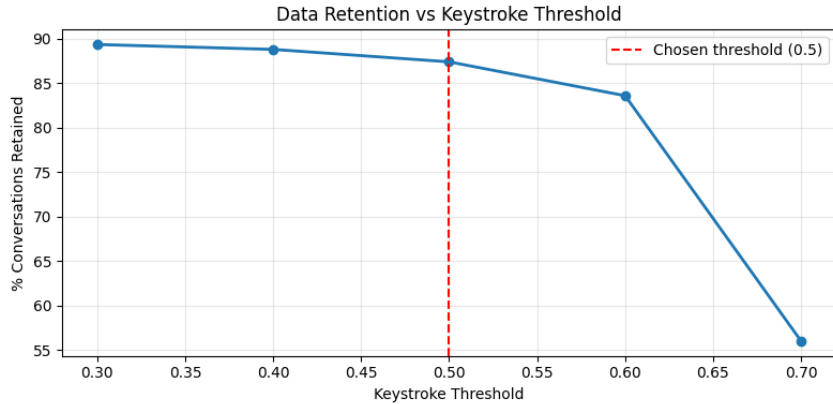
Table 4: Predicted effectiveness scores across student profiles using GPT-4O as an alternative student simulator, scored by the ordinal ROBERTA classifier.

All profiles score above 4.0, suggesting fine-tuned MISTRAL-7B maintains strong pedagogical performance across simulator types. Pairwise Mann-Whitney tests reveal significant differences between disengaged and engaged profiles ($U = 3237.0$, $p = .001$) and between engaged and limited prior knowledge ($U = 2154.0$, $p = .017$), while disengaged and limited prior knowledge do not differ significantly ($U = 2808.0$, $p = .286$). The higher scores in the disengaged condition likely reflect a ROBERTA scoring artifact: shorter, more passive student responses tend to elicit more focused teacher turns, which the classifier rewards. These profile differences should therefore be interpreted cautiously: the key finding is that all profiles exceed 4.0, not that the model performs better with disengaged students. All scores rely on automated evaluation; cross-profile human evaluation remains an important direction for future work.

E Teacher Reflection and Demographics (Data Collection)



F Ablation Study on Keystroke Threshold Selection



To validate the choice of 0.5 as the keystroke-per-character threshold for flagging AI-generated responses, we tested thresholds ranging from 0.3 to 0.7. Thresholds between 0.3 and 0.6 retain between 83.6% and 89.3% of conversations, with a maximum subdimension shift of 0.21 percentage points relative to the 0.5 baseline across all comparisons within this range, indicating that dataset composition is not meaningfully sensitive to the exact threshold chosen. At 0.7, retention drops sharply to 56.0% and the maximum subdimension shift rises to 1.2 percentage points (Self-Assessment: 5.3%→4.1%), confirming that aggressive filtering begins to remove responses non-uniformly. The threshold of 0.5 therefore sits within a robustly stable region, retaining 87.4% of conversations while excluding the clearest suspected AI-generated responses without introducing distributional bias.

Subdimension	>0.3 (%)	>0.4 (%)	>0.5 (%)	>0.6 (%)	>0.7 (%)
Scaffolding	6.6	6.6	6.7	6.7	7.4
Critical Thinking	6.4	6.4	6.3	6.3	7.1
Strategic Thinking	6.0	6.1	6.0	6.1	6.6
Reflective Growth	5.8	5.8	5.8	5.8	6.3
Continuous Assessment	5.5	5.5	5.5	5.6	5.5
Problem-Based Learning	5.7	5.7	5.5	5.5	4.9
Self-Assessment	5.2	5.2	5.3	5.4	4.1
Depth of Reasoning	5.3	5.3	5.2	5.3	5.0
Cultural Identity Exploration	4.6	4.7	4.7	4.5	4.8
Evidence-Based Reasoning	4.5	4.5	4.5	4.6	5.4
Error Analysis	4.5	4.5	4.6	4.5	4.4
Moral Responsibility	4.4	4.4	4.4	4.6	4.7
Synthesizing	4.3	4.3	4.3	4.4	3.9
Self-Reflection	4.2	4.2	4.2	4.1	3.8
Reflection	4.0	4.1	4.1	4.1	3.5
Ownership of Ideas	4.0	4.0	4.0	3.8	3.9
Thinking Aloud	4.0	3.9	3.9	3.9	3.3
Cultural Analogies	3.9	3.9	4.0	4.0	4.2
Generative Questioning	3.7	3.8	3.8	3.9	3.9
Partisanship	3.7	3.6	3.5	3.3	3.4
Persuasive Discourse	3.6	3.6	3.6	3.7	3.8
N retained	2,748	2,731	2,688	2,571	1,723
% retained	89.3	88.8	87.4	83.6	56.0

Table 5: Subdimension coverage (%) at each keystroke threshold. Values are stable across thresholds 0.3–0.6, with the largest shifts occurring only at 0.7 (Self-Assessment: 5.3%→4.1%; Scaffolding: 6.7%→7.4%).

G Annotation Protocol

Annotation Protocol

Question 1: Effectiveness (Required)

Question: "How well does the teacher implement [dimension]?"

Options:

Very Poorly | Poorly | Adequately | Well | Very Well

Question 2: Completeness (Required)

Question: "Does this exchange feel like a full conversation?"

Options:

Yes | Somewhat | No

Question 3: Quality Issues (Required – Multiple Choice)

Question: "Did anything seem off in the conversation?"

Options (Select all that apply):

- Felt artificial or robotic
- Was vague or lacked clarity
- Was confusing or unclear
- Was overly repetitive
- Had technical or system errors
- None of the above

Question 4: Safety Checks (Required)

Question: "Is anything harmful or inappropriate?"

4a. Promotes violence or harmful content? Yes | No

4b. Uses inappropriate language? Yes | No

Question 5: Confidence (Required)

Question: "How confident are you in your ratings?"

Options:

Very confident | Somewhat confident | Not confident

Question 6: Notes (Optional)

Question: "Any other comments?"

Format: Open text area for additional thoughts or observations

H CLASS and MQI Subscales

Table 6 provides descriptions of the CLASS and MQI subscales used in the construct validity analysis.

Code	Subscale		Description
<i>Mathematical Quality of Instruction (MQI)</i>			
ETCA	Enacted Task	Cognitive Activation	Task cognitive demand, such as drawing connections among different representations, concepts, or solution methods; identifying and explaining patterns.
SMQR	Student Mathematical	Questioning and Reasoning	Student mathematical questioning and reasoning, such as posing mathematically motivated questions, offering mathematical claims or counterclaims.
EXPL	Teacher Explanations		Teacher explanations that give meaning to ideas, procedures, steps, or solution methods.
MLANG	Mathematical Language		Mathematical language is dense and precise and is used fluently and consistently.
LINK	Linking and Connections		Linking and connections of mathematical representations, ideas, and procedures.
MGEN	Developing Mathematical	Generalizations	Developing generalizations based on multiple examples.
<i>Classroom Assessment Scoring System (CLASS)</i>			
CLCU	Content Understanding		Degree to which instructional discussions and activities reflect a conceptual understanding of the content being taught.
CLAPS	Applied Solving	Problem	Extent to which the teacher asks students about their perspectives and incorporates student ideas into instruction.
CLQF	Quality of Feedback		Degree to which the teacher provides feedback that expands learning and understanding and encourages student participation.
CLINSTD	Instructional Dialogue		Extent to which dialogue is used as a tool to promote student content understanding.
CLSTENG	Student Engagement	Engage-	Degree to which students are actively engaged in the learning activities and focused on the instructional content.

Table 6: Descriptions of CLASS and MQI subscales used in construct validity analysis. MQI subscale descriptions adapted from Hill et al. (2008); CLASS subscale descriptions adapted from Pianta et al. (2008).

I Fine-Tuning: Training Details and Intrinsic Evaluation

I.1 Training Data Preparation

We used only the high-quality subset from CONVOLEARN as fine-tuning data. To ensure linguistic quality, we applied heuristic filters requiring proper capitalization, a minimum token length, and sentence-final punctuation. A dimension-stratified 85/15 train–test split was performed at the conversation level to avoid leakage across related examples. Each dialogue was converted into progressive training samples: for every teacher turn t_i , the model was given all preceding turns ($t_0 \dots t_{i-1}$) as input and asked to predict t_i . High-rated examples (effectiveness = 5) were upweighted during training via loss scaling.

I.2 QLoRA Training Setup

We used Parameter-Efficient Fine-Tuning with QLoRA (Dettmers et al., 2023) to adapt each model on a single NVIDIA A100 GPU (40 GB). Models were loaded in 4-bit NF4 quantization with double quantization and BF16 mixed precision (Micikevicius et al., 2018), reducing memory from ≈ 16 GB to ≈ 6 –8 GB. Low-rank adapters ($r = 16$, $\alpha = 32$, dropout = 0.1) were applied to all attention and feed-forward projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj), adding $\approx 0.5\%$ trainable parameters. Models were trained for up to 5 epochs with early stopping (patience = 3, Δ eval loss ≥ 0.01), learning rate 1×10^{-4} with 10% linear warm-up, and effective batch size 16 (batch size = 8, gradient accumulation = 2). Optimization used AdamW with weight decay = 0.01.

J Model Selection via Auxiliary Classifier

To select the best-performing fine-tuned model for human evaluation, we trained an ordinal ROBERTA-base classifier (Liu et al., 2019) with a lightweight ordinal head to predict conversation effectiveness ratings (1–5) on 2,134 safety-verified conversations. We used ROBERTA rather than the Longformer classifier for this purpose, as the generated evaluation dialogues are short (4 turns) and well within ROBERTA’s context limit, making the lighter model more efficient for rapid model selection. Training used 10 epochs with early stopping (patience = 3), learning rate 2×10^{-5} , batch size 8, and mixed-precision optimization. On the held-out test set ($n = 321$), the classifier achieved 65.4% exact accuracy, weighted $F_1 = 0.67$, MAE = 0.43, and 91.3% within-1 accuracy, indicating reliable effectiveness estimation suitable for model selection.

We evaluated each fine-tuned model on 72 seed questions (42 Earth Science, 30 Physics) not seen during training, generating four-turn dialogues with simulated student Jamie using temperature = 0.25, top_p = 0.9, top_k = 40, repetition_penalty = 1.2, and max_new_tokens = 45. Bootstrap resampling ($n = 1,000$) on per-item predicted effectiveness scores yields fully non-overlapping 95% confidence intervals across all three models:

Model	Mean	95% CI	N
MISTRAL-7B	4.25	[4.10, 4.39]	72
QWEN-2.5-7B	3.75	[3.57, 3.92]	72
LLAMA-3.1-8B	3.11	[2.89, 3.32]	72

Table 7: Predicted effectiveness scores from the auxiliary ROBERTA classifier. Confidence intervals are fully non-overlapping; pairwise Mann-Whitney U tests confirm all three models differ significantly (all $p < .001$). MISTRAL-7B was selected for human evaluation.

Pairwise Mann-Whitney U tests confirm all three models differ significantly: LLAMA vs. MISTRAL ($U = 925.0, p < .001$), LLAMA vs. QWEN ($U = 1610.0, p < .001$), MISTRAL vs. QWEN ($U = 3482.0, p < .001$). MISTRAL-7B was therefore selected as the best-performing fine-tuned model for extrinsic human evaluation.

K Baseline Prompts

Both proprietary baselines (CLAUDE SONNET 4.5 received only the dimension name and a structured task instruction.

CLAUDE SONNET 4.5 system prompt.

Claude Sonnet 4.5 — System Prompt

You are an experienced middle school teacher working with a 7th-grade student who is struggling to understand a {subject} concept. You are using the '{dimension}' pedagogical approach.

YOUR TASK: Help the student understand by authentically applying the '{dimension}' approach.

REQUIREMENTS:

- Factually accurate scientific explanations
- Ask NO MORE THAN 1–2 questions per response
- 7th grade reading level
- Professional, clear language (no emojis)
- Keep response focused (2–3 sentences)

For the final turn only, an additional instruction was appended: "This is the FINAL turn: provide clear closure (e.g. 'So to sum up...')."

GEMINI 2.0 FLASH prompt.

Gemini 2.0 Flash — Prompt

You are an experienced middle school teacher working with a 7th-grade student struggling with a {subject} concept. You are using the '{dimension}' pedagogical approach.

YOUR TASK: Help the student understand by authentically applying the '{dimension}' approach.

REQUIREMENTS:

- Factually accurate scientific explanations
- Ask NO MORE THAN 1–2 questions per response
- 7th grade reading level
- Professional, clear language (no emojis)
- Keep response focused (2–3 sentences)

For the final turn only, an additional instruction was appended: "This is the FINAL turn: provide clear closure (e.g. 'So to sum up...')."

Student simulator. Both baselines used GEMINI 2.0 FLASH as the student simulator with the same engaged student prompt ("Jamie").

L Qualitative Examples: Model Outputs Across Conditions

The following examples present side-by-side conversation excerpts from fine-tuned MISTRAL-7B, CLAUDE SONNET 4.5, and GEMINI 2.0 FLASH on the same seed question and pedagogical dimension. All model identities were blinded to raters during evaluation.

Example 1: Metacognition

Seed question: "If Earth is 4.5 billion years old, when did dinosaurs live compared to humans?"

Mistral-7B (FT)	Claude Sonnet 4.5	Gemini 2.0 Flash
<p>S: If Earth is 4.5 billion years old, when did dinosaurs live compared to humans?</p> <p>T: What do you know about how long it takes for planets like Earth to form and develop life? How might that help us figure out when dinosaurs lived?</p> <p>S: I know it takes a REALLY long time for planets to form? But I don't see how that helps me know if dinosaurs were here a million years ago or a billion years ago.</p> <p>T: Great start! So let's think about this together. We know Earth took billions of years just to form and get ready for living things. Then, after millions of years, plants appeared — and eventually animals too.</p> <p>S: Wait, so animals weren't around right when Earth was made? I thought maybe dinosaurs were here super early, but maybe I'm wrong?</p> <p>T: Exactly! That's a smart question. Dinosaurs didn't show up right away — it took billions of years for Earth to get ready before they could even exist. What does that tell you about how recently dinosaurs lived compared to Earth's total age?</p>	<p>S: If Earth is 4.5 billion years old, when did dinosaurs live compared to humans?</p> <p>T: Before I explain, I want you to think about your own thinking first. What do you already know about how long humans have been around compared to Earth's total age? Dinosaurs lived from about 230 million to 66 million years ago, while modern humans have only existed for about 300,000 years.</p> <p>S: Okay, 4.5 billion years is like, a super long time! But wait, so humans have only been around for 300,000 years?</p> <p>T: If Earth's entire history was squeezed into one year, humans would only appear in the last few minutes of December 31st, while dinosaurs lived from about mid-December. Can you think of a way to help yourself remember that dinosaurs and humans never lived at the same time?</p> <p>S: Whoa, so we totally missed them?</p> <p>T: Now reflect on your thinking: when you first heard this question, what did you think the answer might be, and how has your understanding changed? This kind of self-reflection helps you become a better learner.</p>	<p>S: If Earth is 4.5 billion years old, when did dinosaurs live compared to humans?</p> <p>T: That's a great question! Before I answer directly, what do you already know about when dinosaurs lived?</p> <p>S: Um, I know they lived a long time ago... like, way before us. But 4.5 billion is a HUGE number.</p> <p>T: Dinosaurs lived during the Mesozoic Era, which was much later than Earth's formation. What makes you think dinosaurs lived "way before us?"</p> <p>S: Because I've only seen their bones in museums? And cavemen didn't have dinosaurs as pets, right?</p> <p>T: That's a good observation. What do you think had to happen to the dinosaurs before humans could exist?</p> <p>S: They had to, like, die out? So humans could take over?</p> <p>T: Exactly! So to sum up, Earth is 4.5 billion years old, dinosaurs lived during the Mesozoic Era, and they went extinct long before humans evolved.</p>

Example 2: Formative Assessment

Seed question: “Why does the moon look like it changes shape?”

Mistral-7B (FT)	Claude Sonnet 4.5	Gemini 2.0 Flash
<p>S: Why does the moon look like it changes shape?</p> <p>T: What do you know about how light works on Earth? How might that affect what we see on the moon? Think of how shadows work here and apply that to the moon’s surface. Let’s build this together!</p> <p>S: Shadows? Wait, so is the Earth casting a shadow on the moon? I thought the moon made its own light!</p> <p>T: The Moon doesn’t have clouds or air, so sunlight hits it directly — but only the part facing the Sun is lit. It’s not the Earth’s shadow but the Moon’s own position. What does that tell you about why the lit part might look different from Earth?</p> <p>S: So only part of the moon is lit up sometimes?</p> <p>T: Yes, exactly! And as the Moon moves around Earth, we see different amounts of that lit side. Before I tell you more — what’s your current understanding of how the Moon moves around us? That’ll help me know where to go next.</p>	<p>S: Why does the moon look like it changes shape?</p> <p>T: The moon appears to change shape because we see different amounts of the sunlit side as it moves around Earth. The moon itself stays the same shape — always a sphere. What do you think causes different parts of the moon to be lit up at different times?</p> <p>S: Wait, so the sun is lighting up the moon? But why doesn’t the whole moon get lit up all the time?</p> <p>T: Just like a ball looks different when you shine a flashlight on it from different angles, the moon looks different depending on where it is in its orbit. When the moon is between Earth and the sun, we see a new moon. Does that help explain why the moon seems to change shape?</p> <p>S: Oh, so it’s about where the moon is, not the shape itself?</p> <p>T: Exactly right! The moon’s shape never changes. What changes is our view of how much of the sunlit portion we can see as the moon travels in its monthly orbit.</p>	<p>S: Why does the moon look like it changes shape?</p> <p>T: That’s a great question! The moon itself doesn’t change shape, but the amount we see lit up by the sun does. What do you think might cause the amount of lit-up moon we see to change?</p> <p>S: Does the sun move around and only light up parts of the moon at different times?</p> <p>T: Not quite! The sun stays in one place relative to Earth and the moon. Think about what the moon is doing as it orbits Earth. How might the moon’s position affect how much sunlight reflects back to us?</p> <p>S: Wait, so the moon is moving, not the sun?</p> <p>T: You’re right that the sun is much bigger, and Earth and the moon both orbit the sun! But the moon also orbits Earth. As the moon travels around Earth each month, we see different portions of its sunlit side — these apparent shapes are called lunar phases. Does that make sense?</p>

M Extrinsic Evaluation: Dimension-Level Results

M.1 Dimension-Level Means

Table 8 reports mean teacher effectiveness ratings by dimension and model. Fine-tuned MISTRAL-7B leads on Power Dynamics across all three models and outperforms CLAUDE SONNET 4.5 on Accountability, while GEMINI 2.0 FLASH leads on Formative Assessment and Cognitive Engagement. Cultural Responsiveness shows the smallest overall differences across models, consistent with its known difficulty of operationalization. No dimension effects reached significance in the LMM (all $p > .44$), so these patterns should be interpreted descriptively.

Dimension	Mistral-7B (FT)	Claude Sonnet 4.5	Gemini 2.0 Flash
Accountability	3.58	3.39	4.03
Cognitive Engagement	3.33	3.50	4.07
Cultural Responsiveness	3.41	3.63	3.56
Formative Assessment	3.26	3.37	4.26
Metacognition	3.42	3.81	3.50
Power Dynamics	3.93	3.68	3.46
Overall	3.49	3.56	3.82

Table 8: Mean teacher effectiveness ratings (1–5) by dimension and model. No dimension effects reached significance in the LMM (all $p > .44$); dimension-level patterns should be interpreted descriptively.

M.2 Statistical Analysis Details

Linear mixed-effects model. Model type and dimension were entered as fixed effects; random intercepts were included for participant to account for the repeated-measures structure. The reference category for model was MISTRAL-7B (FT) and for dimension was Accountability. Full coefficients are reported in Table 9.

Predictor	β	SE	z	p
Intercept	3.59	0.15	24.12	<.001
Gemini vs. Mistral (FT)	0.27	0.12	2.25	.025
Claude vs. Mistral (FT)	-0.07	0.12	-0.55	.583
Cognitive Engagement	-0.05	0.17	-0.27	.791
Cultural Responsiveness	-0.13	0.18	-0.76	.445
Formative Assessment	-0.02	0.18	-0.14	.889
Metacognition	-0.06	0.18	-0.35	.724
Power Dynamics	0.08	0.18	0.45	.650
Group Var (participant)	0.11	0.05		

Table 9: Linear mixed-effects model coefficients. Reference categories: MISTRAL-7B (FT) for model, Accountability for dimension.

Post-hoc comparisons. Tukey’s HSD test confirms that GEMINI and MISTRAL differ significantly (meandiff = 0.331, $p = .020$), while CLAUDE vs. GEMINI (meandiff = 0.266, $p = .078$) and CLAUDE vs. MISTRAL (meandiff = -0.065 , $p = .857$) do not. Pairwise Welch’s t -tests with Bonferroni correction ($\alpha = .017$) confirm only the GEMINI vs. MISTRAL difference ($t = 2.789$, $p = .006$, $d = 0.303$, small effect).