When Help is Unhelpful: Evaluating AI Tutors for Productive Struggle

Albert Zhang[2] **Alexis Ross**[4] **Kajal Patel**[1] **Julian Bernado**[5]
Rebecca Bowie[2] **Ana Trindade Ribeiro**[5] **Daniel Halper**[6] **Haripriya Valayaputtur**[6]
Jacob Andreas[4] **Susanna Loeb**[5] **Li Lucy**[3] **Kyle Lo**[1,3] **Ryan Knight**[2]

[1]Allen Institute for AI [2]Insource Services [3]University of Washington [4]MIT
[5]Stanford [6]Step Up Labs

Abstract

General language model (LM) development optimizes for behaviors like helpfulness, but effective AI tutors must know when to help and when to challenge students. Existing LM tutoring benchmarks often evaluate tutoring behaviors, such as avoiding answer giving or providing scaffolds, without evaluating whether those behaviors are appropriate for the specific learning moment. We introduce TutorMoments, a replay-based framework for evaluating whether language model tutors adapt their pedagogical actions to context. TutorMoments replays teacher-identified pedagogical decision points from authentic student-tutor transcripts and evaluates whether an LM tutor scaffolds when support is needed, pushes for rigor when the student is ready, and avoids over-scaffolding. Experienced math teachers identify key learning moments and provide natural-language annotations describing what actions the situation is appropriate for, the tutor's actual actions, and the result, grounding evaluation in authentic tutoring practice. We find that minimally prompted LMs frequently over-scaffold and rarely push for rigor, suggesting that frontier models default toward helpfulness at the expense of adapting instruction to learners' needs. Evaluation-aware prompts substantially improve adaptability but still reveal large differences across frontier models and concentrate behavior into a relatively narrow set of tutor moves. We release TutorMoments-Preview, a dataset of 462 de-identified tutoring transcripts, more than 1,500 teacher-annotated key moments, code, and initial results to support future research on AI tutors that adapt instruction to learners and learning contexts.

1 Introduction

A prominent concern is that AI optimized to be a "helpful assistant" (Zheng et al., 2024; Anthropic, 2024; Maiya et al., 2025) may harm human learning by reducing cognitive rigor (Perczel et al., 2025; Tao et al., 2026). For instance, access to AI without pedagogically sound design considerations can improve students' performance on AI-supported practice, while reducing later performance without support (Bastani et al., 2025). Generally, AI use has been shown in recent work to lead to deskilling, over-reliance, and cognitive offloading behaviors (Ke et al., 2026; Liu et al., 2026; Padmakumar et al., 2026; Zhi et al., 2026; Ibrahim et al., 2026).

Post-training practices that incentivize helpfulness may misalign with desirable behaviors in educational settings, where the goal is not to always be directly helpful, but instead to provide the right level of assistance at the right moment. Educational theories like the zone of proximal development (Vygotsky, 1978), scaffolding (Wood et al., 1976), productive struggle (National Council of Teachers of Mathematics, 2014; Hiebert and

Grouws, 2007; Young et al., 2024; Kapur, 2008) and desirable difficulty (Bjork and Bjork, 2020) all position teaching as a balance of support and challenge, creating opportunities for students to exert effort to learn while intervening when difficulty becomes unproductive. Across these theoretical perspectives, effective instruction depends on selecting pedagogical strategies that are responsive to learners' evolving understanding and current learning needs (Heritage, 2007; Council et al., 2000; Black and Wiliam, 1998). Consequently, the effectiveness of any instructional action depends on its alignment with the pedagogical demands of the learning situation rather than on the action itself. Thus, AI tutors must also navigate this assistance dilemma (Koedinger and Aleven, 2007), balancing when to support students through **scaffolding** that makes content more accessible (e.g. breaking a problem into parts with hints), and when to **push for rigor** by increasing the cognitive demands of the task (e.g. requesting students to self-explain) .

A defining characteristic of effective tutoring is the continual adaptation of instructional decisions to learners' evolving understanding. Existing tutoring benchmarks primarily fall on a single side of the assistance dilemma by always rewarding models for not helping too much (e.g. answer-giving), or always rewarding scaffolds, without accounting for whether the type of action or support is appropriate for the situation. As examples, MathTutorBench's scoring metric is trained to prefer tutor responses that scaffold rather than give away the answer (Macina et al., 2025), while MRBench penalizes models for revealing the answer to students (Maurya et al., 2025). Other benchmarks, such as LearnLM's rubric, provide underspecified guidance that tutors should neither give away answers too quickly nor withhold answers unproductively (LearnLM Team Google, 2024). In our work, we add nuance to this tension by asking, is the level and type of support provided appropriate *given the situation*? We evaluate instructional decision making along two action directions, scaffolding and pushing for rigor, that are well-studied in education research, providing clearer articulation of tutoring behaviors suggested in prior benchmarks. We focus on teacher-identified pedagogical decision points, allowing standardized comparison of how well AI tutors respond to the same instructional situation.

In the following sections, we describe TutorMoments, a benchmark designed to evaluate whether AI tutors adapt their instructional strategy to the pedagogical demands of a learning moment (Figure 1). TutorMoments evaluates whether tutors choose instructional actions that are appropriate for the instructional context. Our framework leverages annotations from experienced math teachers to identify key pedagogical decision points in authentic tutoring sessions (§3), centering task conditions for LMs on scaffolding- and rigor-appropriate tutoring moments (§2.1). We evaluate LM tutors by replaying key moments prefixed by real student-tutor interactions (§2.2), assessing whether LMs scaffold when support is needed, push for rigor when students are ready, and avoid over-scaffolding. We validate an LM-based scoring pipeline for assessing these behaviors (§4), and apply our scoring pipeline and metrics onto the LM tutor's replays (§5).

We show that our evaluation framework is able to differentiate stronger and weaker LMs, and it is able to delineate the effects of differing prompting strategies. We find that LMs, when simply prompted to act as tutors, tend to over-scaffold and miss rigor-pushing opportunities. Although evaluation-aware prompting substantially improves performance, frontier models still differ markedly in their ability to provide contextually appropriate instruction. Thus, our benchmark demonstrates utility for educators and AI tutor developers, who may wish to iterate on model, prompt, and design choices before deployment in user studies.

In this preliminary report, we release TutorMoments-Preview, a sample dataset created using our framework. This sample includes 462 human student-tutor tutoring session transcripts from a high-dosage tutoring organization and teachers' annotations on a subset of transcripts. We also release LM replays to facilitate reproducibility and transparency of our results. We release this data sample and working paper to receive feedback from the broader community as we prepare for a larger multimodal dataset, improved scoring pipeline, and additional analyses to come.

2 TutorMoments: Situation-Conditioned Replay Evaluation

This section defines the TutorMoments evaluation framework. We first define scaffolding- and rigor-appropriate moments as task conditions, then describe simulated replays as the mechanism for evaluating LM tutor behavior at those moments.

Figure 1 An overview of our TutorMoments framework. **01** Expert teachers select key learning moments in real high dosage tutoring sessions where the tutor made an important pedagogical choice, and mark a cut point where replays should begin. **02** Teachers write free-text annotations capturing why this situation is a key learning moment, what actions the tutor took, and the resulting effectiveness of the moves. **03** A LM uses S/A/R annotations to classify whether the teachers believe the tutor should push for rigor by increasing the cognitive demands of the task or make the task more accessible by scaffolding. We also classify whether the teachers believe over-scaffolding is present. **04** In the replay phase, the LM tutor receives the entire transcript up to the cut point as context and generates the next response. The response is sent to a synthetic student, then back to the language model tutor. Replay stops after 5 turns (3 tutor and 2 student). **05** An LM generates action descriptions for the tutor's responses and the LM classifier uses the synthetic descriptions to identify whether the LM tutor did scaffolding or pushed for rigor, and whether the LM tutor over-scaffolded. **06** We aggregate the number of times the LM tutor pushed for rigor in push for rigor moments, scaffolded in scaffolding moments, and avoided over-scaffolding.

2.1 Scaffolding and Rigor Decision-Making as Task Conditions

Effective tutoring depends on selecting instructional strategies that match the learner's current understanding. TutorMoments operationalizes this idea by defining evaluation tasks around instructional decision points. One-on-one tutoring is a collaborative dialogue between tutors and students, where tutors engage in strategic behaviors to elicit learning in students (Graesser et al., 1995; Graesser and Person, 1994). Prior work in education data mining often focuses on the identification of a fixed taxonomy of "tutor moves" that are broadly seen as common in tutoring or conducive to learning (Zhou et al., 2026b; Maurya et al., 2025; Suresh et al., 2022; Kupor et al., 2023; Boyer et al., 2010). Our framework differs in that we characterize tutoring as strategic decisions made within specific dialogue contexts, instead of only an analysis of move prevalence.

We use behavioral characteristics of student-tutor interactions to create task conditions for LMs. Task conditions are learning situations judged by teachers as to whether scaffolding or greater cognitive challenge would be appropriate. These task conditions allow us to evaluate alignment between pedagogical strategies and conversational contexts, and enable insights into how well tutors maintain productive struggle by scaffolding for access or pushing for rigor. With our approach, we are also able to consider the *absence* of appropriate tutor actions, or missed opportunities. Our framework could be extended to evaluate other pedagogical strategies by building similar context-dependent task conditions.

In our work, **scaffolding** refers to temporary support provided by a teacher or tutor to make content more accessible to a student (Wood et al., 1976; Vygotsky, 1978; Van de Pol et al., 2010; Barlow et al., 2018). We use scaffolding as an umbrella term to describe a broad set of actions or "tutor moves" that make content more accessible. For example, a tutor may break down a problem into manageable chunks, offer a simplified version of the task, or demonstrate a correct example. **Over-scaffolding** occurs when a tutor's actions reduce the cognitive demands of a task more than necessary for a given situation. Instead, tutors are expected to appropriately **push for rigor** by encouraging higher-order thinking and increasing the cognitive demand of tasks when the student's behavior suggests they are ready for it (Henningsen and Stein, 1997; Payne et al., 2005; Draeger et al., 2013). As examples of moves that increase rigor, a tutor may ask a student to explain a correct answer, generalize a known concept to a new context, complete variations of problems independently without support, or increase the difficulty of subsequent problems (Wyse and Soneral, 2018). Table 7 provides a bottom-up taxonomy of tutor actions associated with scaffolding/rigor.

We condition our evaluation tasks on whether teacher-annotators judge the conversational context as an appropriate moment to scaffold for access or appropriate to push for rigor (§3.4), allowing us to measure alignment and divergence among LM tutor behavior, teacher-judged appropriate behavior and actual human tutor behavior at key moments (§5).

2.2 Simulated Replays

Our approach for evaluating instructional decisions preserves the dialogue leading to the decision, allowing alternative tutor actions to be compared under identical conditions. Unlike existing evaluations that compare tutors across different conversations, TutorMoments uses simulated replay to compare alternative instructional decisions from the same pedagogical state. This provides a more controlled setting for evaluating whether differences in performance arise from tutors' instructional decisions, holding dialogue history constant. TutorMoments accomplishes this through simulated replays.

A simulated replay is a real human tutor-student transcript stopped at an evaluation task moment and passed between an LM tutor and a synthetic student, at which point the replay diverges from the original human transcript and novel tutoring behaviors are provided by the LM tutor. Our framework examines multiple tutor turns, because often the signal around a tutor's strategy (e.g. breaking a problem into steps) is not scoped to a single turn, but rather multiple.

We evaluate language models under two prompt conditions that allow us to distinguish between models' default tutoring behavior and their ability to follow explicit pedagogical guidance. We call these instructions "plain" and "evaluation-aware" prompts. Full prompts are in Appendix D.2.

- The plain prompt relies on the model's default ability to tutor, receiving only the guidance to "`Use everything you know about what makes great tutoring to respond appropriately to the student.`"

- The evaluation-aware prompt explicitly describes the trade-offs between scaffolding, over-scaffolding, and pushing for rigor, and instructs the model to "`stay in the student's zone of proximal development by scaffolding for access when the content is challenging for the student / pushing for rigor when the student masters more basic expressions of a concept.`"

To attribute differences in tutoring behavior to the tutor, we use an oracle student during replay to hold simulated student behavior constant. Recent work has proposed user simulators as a bridging point between single-turn static benchmarks and real human-AI interaction studies (Zhou* et al., 2024; Chang et al., 2025; Naous et al., 2026; Lin et al., 2024). In TutorMoments, synthetic students play the role of the user (Scarlatos et al., 2026; Perczel et al., 2025; Jin et al., 2025; Weitekamp et al., 2025). In replays, we prompt Claude Opus 4.6 to mimic the original student, and condition each synthetic student with an individualized prompt containing a textual description of the original human student's behavioral characteristics, extracted by Claude Opus 4.6 from the transcript up to the cut point. These behavioral characteristics include: whether the student is active versus passive, the emotional/affective state of the student, how consistently the student stays focused on learning, how readily the student revises incorrect beliefs, and any misconceptions revealed in the transcript. The full synthetic-student prompt is also included in Appendix D.2. To maximize the chance the synthetic student will persist misconceptions over multiple turns, we also allow the synthetic student to see the full original transcript when generating its response. Our synthetic student is thus an "oracle student," privileged with information the LM tutor does not have.

In pilot replay runs, we observed that even oracle synthetic students tend to produce unnaturally successful learning behaviors. Thus, we presently use our behavior-conditioned synthetic students as a placeholder to observe how LM tutors lead the conversation after a scaffolding-rigor evaluation point. Each "replay" after a cut point is 5 turns long with the tutor initiating; it includes three LM tutor turns and two synthetic student turns. The 5 turn replay allows the evaluation to focus on the tutor's decision during the identified scaffold-rigor appropriate moment, without moving on to new content. Once the replay has completed, we use our LM-based scoring pipeline to evaluate the replay, scoping scoring to the LM-generated replay with none of the human transcript (§5).

3 Data Curation and Annotation

The value of TutorMoments depends on grounding evaluation in authentic tutoring interactions and expert pedagogical judgments. This section describes both the data curation recipe used by TutorMoments to build teacher-enriched tutoring replay datasets and the specific preview dataset released with this report, which we call TutorMoments-Preview.

3.1 Sourcing Student-Tutor Transcripts

Our data is from a high dosage tutoring provider that provides remote, one-on-one tutoring to U.S-based students at home. Tutors are trained volunteers in the U.S. or college students participating in a work-study program. The tutors did not receive AI assistance. The tutoring provider targets underserved communities, and a majority of students attend Title 1 schools. The tutoring provider released the transcripts under a research clause in their terms of service, affirmatively agreed to by parents and guardians. TutorMoments-Preview contains 462 text-only transcripts of 198 students in grades 2 through 7 interacting with 173 human tutors (Table 1). The tutoring provider transcribed session audio using Gemini 2.5 Pro. Sessions are hosted on a tutoring platform that allows the student and tutor to co-annotate a shared screen and whiteboard. The transcripts also include text descriptions of whiteboard / screen interactions during the session, also generated using Gemini 2.5 Pro. We refer to the text descriptions of visual content as "enrichments". Tutors meet consistently with the same student, with the average student having 2.3 sessions with the same tutor within the released data. The sessions are long-form, with session lengths averaging 54 minutes, and averaging 384 student and tutor turns. The released transcripts are almost exclusively from math tutoring sessions.

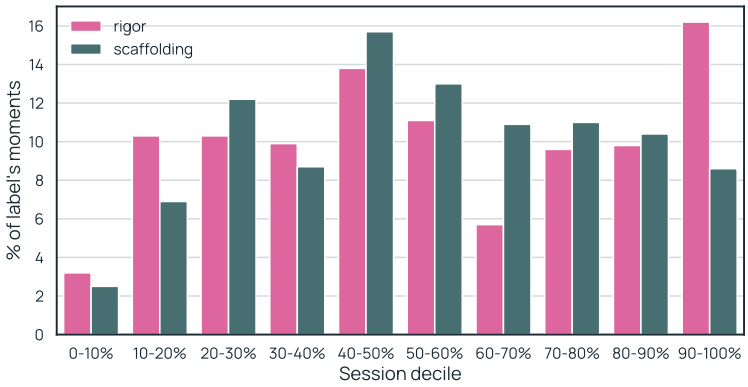


Figure 2 An overview of where teachers' identified key moments appear in tutoring session transcripts.

TutorMoments-Preview	
# of transcripts	462
Avg # of turns per transcript	384.1
Avg # of words per turn	9.5
Avg session duration (minutes)	53.5
# of students	198
# of tutors	173
Avg # of sessions per student-tutor pair	2.3
Teachers' Annotations	
# of transcripts annotated	122
# of unique key moments	1,536
# of teacher annotators per moment	3.5
# of rigor key moments	260
# of scaffolding key moments	738
Avg # of words per SAR annotation	81.5

Table 1 Descriptive statistics for TutorMoments-Preview and SAR annotations.

3.2 De-identification

Math tutoring data may contain personally identifiable information (PII), ranging from direct identifiers (e.g. names, emails) to quasi-identifiers (e.g. location, nationality), that may allow re-identification of session participants. Tutoring data is especially challenging for generic de-identification approaches, as mathematical content may be conflated with PII, and many traditional approaches over-identify math content as PII, reducing the understandability of the transcripts (Zhou et al., 2026a). The tutoring provider de-identified the data prior to transfer. Out of an abundance of caution, we also run a thorough de-identification pipeline on the data to ensure that any potential identifiers missed by the tutoring provider are removed. We use the National Tutoring Observatory (NTO)'s math-aware de-identification prompt (Zhou et al., 2026a), which was designed specifically for math tutoring data, as the starting point for our approach. We iterate and improve their prompt with multiple rounds of manual validation and categorization on 250 transcripts. Our final prompt outperforms theirs on a held-out test set of 50 additional transcripts for both direct identifiers (Recall = 0.997, Precision = 1.00) and quasi-identifiers (Recall = 0.932, Precision = 1.00). Appendix A details our iteration process and includes a fine-grained breakdown of validation results. Our data processing pipeline also normalizes the transcripts into a format consistent with future, larger planned data releases.

3.3 Teacher Annotation

We recruited 27 U.S.-based teacher annotators from the authors' personal networks from their experience teaching in schools using a snowball strategy. All annotators have more than three years of math teaching experience, with 14 reporting more than 10 years. Annotators include full-time math teachers, full-time tutors, teacher coaches, former principals, and assistant principals (details in Appendix B). Teachers are initially paid $35 per hour for their work, which increases to $40 per hour after 5 hours of completed annotation and $45 after 15 hours are completed. We anchor the hourly rates on teachers' union recommended wages for out-of-school work (Everett Teachers' Association and Everett School Committee, 2024). Potential annotators completed a screening annotation task prior to being hired, and received a 30 minute training on the task from the authors.

Teachers first identify **key moments** in transcripts where they believe a tutor makes or should have made an attempt to push for rigor or scaffold for a student. In our instructions for teachers, we lightly define scaffolding as "*making problems more accessible for students*" and pushing for rigor as "*asking students to do harder tasks or tasks that involve more metacognition.*" We chose to collect open-ended, descriptive annotations (Röttger et al., 2022) in order to observe a range of rigor and scaffolding interpretations among experienced math educators. Figure 2 shows the distribution of identified key moments by when they occur in the tutoring session.

For each identified conversation span, teachers provide free-form written descriptions of the following:

- The **situation**, or why this key moment is an appropriate time to push for rigor or scaffold.
- The tutors' **action**, or what strategy or pedagogical approach the tutors used.
- The **result**, or how effective the tutor's strategy was and its impact on the student.

We refer to these descriptions as **SAR annotations**. We use free-text annotations rather than predefined labels, so that teachers can describe pedagogically important moments in their own words before those judgments are mapped to structured evaluation categories. Teacher SAR annotations are dense natural language, consisting of 82 words on average (Table 2). Teachers also annotate a **cut point** for each moment, specifying the turn when the tutor takes the lead on making a pedagogical decision.[1] Figure 1 shows an overview of the annotation strategy and Appendix C shows a sample annotation interface. Key moment spans are selected by one teacher and independently re-annotated with SAR descriptions by several others; each key moment received 3.5 annotations on average. In total, teachers annotated 122 transcripts with 1,536 unique key moments (Table 1).[2]

Example Annotation 1	
Situation	This is an appropriate time to scaffold because the student has expressed that they need help with the question.
Action	The tutor broke the question down into chunks, looking at each answer choice and breaking down whether or not each answer choice's number line shows 432–175. The tutor also asks the student to first solve 432–175 so they can ensure each answer choice results in the same answer.
Result	While it was a good strategy to have the student solve the original math problem and then compare that answer to each answer choice, the tutor appeared to do most of the thinking and the work. The tutor asked questions to the student but it seemed like they didn't provide opportunities for the student to answer or that they just didn't answer the question (or did so without speaking ie. nodding head or something). The supports appeared to be misaligned with what the student needed. There were moments were the student indicated they did not understand, and it seems like the student possibly struggled with understanding what was happening on the number line. It seems like a lot of the number lines are showing adding to subtract, but the tutor never directly states that this is a strategy for subtracting and it sometimes easier.
Example Annotation 2	
Situation	The student is starting a new problem. The student said they didn't need help with the problem, so this isn't a place to scaffold. In fact, this could have been a great place for the tutor to push rigor for the student.
Action	The tutor did not do anything to push rigor during this section. They did, though, walk the student through all of the steps by asking guiding questions for the student.
Result	This doesn't seem to be effective at this time. The student said, initially, they thought they would be able to do problem without any help. the tutor then jumping in to help the student work through the problem. This was over-scaffolding and took away from the student being able to try to make sense of the work independently. This also took away from the tutor being able to see what the student truly understands. They were unable to determine if there were any misconceptions held by the student that may need to be addressed by the tutor.

Table 2 Examples of two teachers' SAR annotations, to illustrate the richness of their free-text descriptions.

[1] When picking cut points to use for replays, we pick the most common cut point recommended by teachers. If there are ties, we pick the earliest point.

[2] The data release on huggingface includes additional teacher annotations on topics not used in this work.

3.4 Situation Types

To classify free-text SAR annotations into scaffolding-appropriate moments or rigor-appropriate moments, we extract the situation types from teachers' written descriptions. We prompt Claude Opus 4.8 to map teachers' phrases to `json` representing what the situation is appropriate for, e.g. "This is a good time to push for rigor" → `{rigor : 'yes', scaffolding: 'no_mention'}` (prompt in Appendix D). We validate this LM mapping step on 200 randomly sampled situation descriptions, achieving 96% accuracy based on one author's independent labeling. When teachers mention that a moment is appropriate for scaffolding, they typically do not mention any appropriateness judgment for rigor, and vice versa. Given that no mentions implies 'no', our labeling of situation types uses explicitly present 'yes' labels to characterize each situation.

When multiple teachers mention situation appropriateness, 68% all agree on whether the situation is appropriate for scaffolding and 63% all agree on whether it is appropriate for a rigor push. These levels of agreement are consistent with the inherently subjective nature of instructional decision making. Disagreement likely reflects differences in expert instructional judgment, in addition to the annotation-scoping variability. [3] We label each unique key moment with a situation type derived from a majority vote across teachers who annotated the moment. Excluding voting ties from our replay-based evaluation framework, we have 738 moments that are appropriate for scaffolding and 260 appropriate to push for rigor. For LM tutor evaluation, we randomly select 260 scaffolding moments from the 738 candidates to achieve a 50-50 balance between scaffolding- and rigor-appropriate moments. Table 3 shows quoted examples of teachers' reasoning when labeling key moments as scaffolding and rigor situations.

Scaffolding-appropriate situations	Rigor-appropriate situations
`Student arrives at correct answer but reasoning was confused and unclear, understanding is uncertain.`	`The student has been getting the questions correct consistently.`
`... the student has a misconception.`	`...the student just received scaffolding from the tutor to solve a problem similar to this.`
`...the student was struggling with answering an addition with fractions problem correctly.`	`The student was able to answer a question correctly and recall information about a story that was read previously.`

Table 3 Example excerpts from teachers' situation annotations, to illustrate types of scenarios that may initiate scaffolding- and rigor-appropriate moments.

4 Scoring Replay Outcomes

In this section, we describe our LM-based tutor scoring scheme that can scale assessments of key moments in a manner that aligns with teachers' observations and judgments (Figure 3).[4] This scoring pipeline operationalizes teachers' pedagogical judgments into scalable evaluation metrics that can be applied consistently across LM tutors.

4.1 LM Scoring Pipeline

Our LM scoring pipeline begins by asking an LM to **generate** SAR descriptions in a manner similar to the ones teachers provided (prompt in Appendix D.1). Generating SAR descriptions first allows the scoring pipeline to evaluate tutors using the same intermediate representation produced by teachers, before mapping those descriptions into structured labels. We condition the LM's generation on teachers' recommendation of whether the situation is scaffolding- and/or rigor-appropriate. Thus, our scoring pipeline has some privileged information, which we leverage when we later evaluate LM tutors' decisions in §5.

[3] We suspect that teachers may scope the "situation" part of each moment to different potential cut points within the key moment block when making this assessment. Our current paper is a work-in-progress preview of our framework; future work will incorporate a more pinpointed annotation strategy to better disentangle actual disagreement from noise.

[4] In Appendix E, we include additional discussion of experiments that evaluate LMs' ability to identify student outcome signals in *result* annotations.

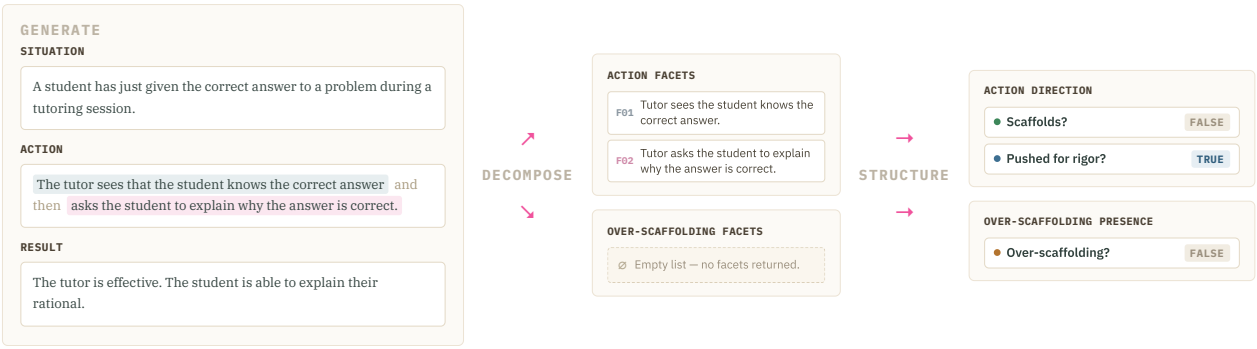


Figure 3 An illustration of how we decompose "facets" from SAR descriptions, which are either written by teachers or generated by LMs. We classify these descriptions into binary labels related to the direction of a tutor's actions and whether they over-scaffold.

We **decompose** the LMs' SAR generations into short, standalone, atomic facets of information, targeting facets that would allow us to determine action direction appropriateness (Figure 3). Decomposing descriptions into atomic facets reduces the complexity of the downstream classification task, and enables more interpretable scoring. We have two decomposition prompts: one for rigor-scaffolding decomposition and another for over-scaffolding identification. We manually validated the faithfulness of 164 LMs' action facets against the original transcript for 30 key moments, finding that all facets correspond to actions that actually appear in the moment. We also validated our over-scaffolding decomposer on 200 manually labeled key moments' teacher annotations (94.5% accuracy, 88.9% F1).

Next, we **structure** facets into binary labels. To do so, we apply a classifier that examines action facets and determines whether the tutor scaffolds and/or pushed for rigor, with one binary outcome each. Table 4 shows examples of action facets for two key moments.

For over-scaffolding, we determine that no over-scaffolding occurred when its decomposition prompt outputs an empty list. Indicators of over-scaffolding in teachers' SAR annotations include moments where students' cognitive participation is overly reduced, e.g. "[the tutor] *is doing a lot of the academic work here*", "[the tutor] *does most of the thinking herself*", or "[the student *is not given an opportunity to apply these skills at all*". Teachers' annotations also cover cases where the tutor scaffolds before the student has shown what they can do, gives away answers or key steps, over-explains, oversimplifies, or intervenes during a student's productive struggle.

4.2 Validating the Scoring Pipeline

Before applying our scoring pipeline to evaluate LM tutors, we must first validate the extent to which it can reliably classify tutors' scaffolding-rigor action directions and identify over-scaffolding. To optimize and validate our scoring pipeline, we divided teacher-annotated transcripts into a 102/105 train/test split. We use the first split to iterate on prompts, and the second as a held-out test set. Our evaluation set aggregates individual teachers' SAR annotations within each key moment. We decompose and structure teachers' unified annotation in the same manner as we do for LMs' SAR descriptions, and we compare whether binary outputs (the presence of scaffolding, rigor, and over-scaffolding) derived from teachers' SAR match those derived from LMs'. [5] Our iteration process involved a mix of manual inspection of misalignments to teacher annotations and patterned examples surfaced by the same LM, which was asked to identify trends among erroneous predictions. We observe that teachers tend to discuss tutoring strategies that are notable to the moment, while LMs default to giving more of a turn-by-turn recap, even when instructed to keep descriptions high-level (Table 4).

For action-direction classification, our validation results show that it is more difficult for LM pipelines to

[5] We originally iterated prompts on our iteration set using Claude Opus 4.6. During our iteration process, Claude Opus 4.8 was released. Since our LM scoring evaluation shows that performance is maintained between the two model versions, we apply the more recent one in later sections of this paper.

Classification	Teachers' Action Description	Claude Opus 4.8's Action Description
Scaffolding	The tutor is doing some of the steps for the student. The tutor is asking guiding questions. The tutor is asking guiding questions.* The tutor breaks the problem down into smaller steps. The tutor asks guiding questions to the student so they can solve together.	The tutor is scaffolding. The tutor walks the student through the remaining steps of the long division. The tutor asks guiding questions tied to each sub-step. The tutor asks what to subtract. The tutor asks what to bring down. The tutor asks what to divide next. The tutor gently corrects the student when they propose a wrong operation.
Push for Rigor	The tutor builds rigor by asking the student to explain their thinking while they were solving the problem. The tutor asks the student to explain how they got their answer. The tutor asked "how did you get that" to make the student articulate the formula and substitution process they used. The tutor is questioning the student. The tutor is asking the student to explain their thought process.	The tutor pushes for rigor. The tutor hands the problem to the student. The tutor has the student read the problem aloud. The tutor asks the student to explain how they arrived at their answer.

Table 4 A side-by-side comparison of tutoring action facets extracted from teachers' and Claude Opus 4.8's SAR annotations. The top row represents a key moment during which the tutor scaffolds, while the bottom row represents a key moment during which the tutor pushes for rigor. *In this example, the same facet repeats across multiple teachers.

identify the presence of rigor pushes than the presence of scaffolding (Table 5). Rigor, as a concept, is seen to be difficult for even experienced instructors to identify, define, and apply consistently (Wyse and Soneral, 2018; Paige et al., 2015). We theorize that transcript understanding may be a prerequisite skill for effective LM tutoring, and so LMs' lesser abilities to recognize rigor-related moves may suggest they are less attuned to generate such moves in tutoring settings as well. Future work could consider digging further into what forms of rigor may be conceptualized in LMs and whether teachers' subjective annotations comprehensively cover the ways in which rigor may manifest in tutoring transcripts.

For over-scaffolding identification, LMs can achieve good precision on moments annotated by at least one teacher as involving over-scaffolding, and even better recall on moments annotated as over-scaffolding by more than two teachers. We show results across both gold subsets in Table 6 due to the open-ended nature of our annotations, where some teachers may have been more inclined to discuss the presence of over-scaffolding than others. Generally, our results suggest that particularly salient instances of over-scaffolding are recovered by an LM scoring pipeline.

We primarily developed our prompts with Claude 4.6 Opus; performance only somewhat generalizes to other model families (Table 5, Table 6). Scoring pipelines in future work that use LMs for multi-criteria measurement may benefit from tuning prompts to multiple scoring LMs, especially since LMs may show preference towards their own outputs (Wataoka et al., 2024).

4.3 Tutor-Action Taxonomy

To provide interpretable summaries of tutoring behavior beyond aggregate benchmark scores, we create a taxonomy of tutor moves with categories derived from the decomposed action facets. We form these categories through an iterative process: we initially prompt Claude Opus 4.8 to group action facets from the author-scored training set into strategy categories, and then further refine category definitions manually.

LM Action-Direction Identification Performance							
		Scaffolding			Push for Rigor		
Model	Evaluation Split	Precision	Recall	F1	Precision	Recall	F1
Claude 4.6 Opus	iteration	0.9352	0.9365	0.9358	0.6519	0.5813	0.6146
Claude 4.6 Opus	held-out	0.9657	0.9386	0.9520	0.7296	0.5329	0.6159
Claude 4.8 Opus	held-out	0.9608	0.9035	0.9313	0.7564	0.5549	0.6401
GPT-5.5	held-out	0.9507	0.8456	0.8951	0.7260	0.4734	0.5731
Gemini 2.5 Pro	held-out	0.9545	0.9211	0.9375	0.7931	0.4326	0.5598
Gemini 3.5 Flash	held-out	0.9488	0.9105	0.9293	0.7753	0.4326	0.5553

Table 5 The performance of our pipeline for determining the direction of a tutor's action in a key moment (scaffolding and/or push for rigor). We evaluate LMs' outputs against labels extracted from teachers' SAR annotations. Cells are colored based on the magnitude of metrics.

LM Over-Scaffolding Identification Performance					
Model	Evaluation Split	Gold Subset	Precision	Recall	F1
Claude 4.6 Opus	iteration	≥ 1 teachers	0.6650	0.6422	0.6534
		≥ 2 teachers	0.2944	0.7785	0.4273
Claude 4.6 Opus	held-out	≥ 1 teachers	0.7237	0.6377	0.6780
		≥ 2 teachers	0.3421	0.7273	0.4653
Claude 4.8 Opus	held-out	≥ 1 teachers	0.7195	0.6841	0.7013
		≥ 2 teachers	0.3567	0.8182	0.4968
GPT-5.5	held-out	≥ 1 teachers	0.7327	0.4609	0.5658
		≥ 2 teachers	0.3733	0.5664	0.4500
Gemini 2.5 Pro	held-out	≥ 1 teachers	0.7353	0.3623	0.4854
		≥ 2 teachers	0.3882	0.4615	0.4217
Gemini 3.5 Flash	held-out	≥ 1 teachers	0.7300	0.4232	0.5358
		≥ 2 teachers	0.3700	0.5175	0.4315

Table 6 The performance of our pipeline for determining whether over-scaffolding occurs during a key moment. We evaluate LMs' outputs against over-scaffolding indicators extracted from teachers' SAR annotations. Cells are colored based on the magnitude of metrics.

The refined categories are then used by the LM to re-group the training set. This process continues until all categories are visibly distinguishable from one another and cover the breadth of action facets observed across both LM and human tutors. Our resulting categories are shown in Table 7, with each category delineated as a scaffolding, rigor, or neutral strategy. A final "Other" category captures all actions that cannot be suitably represented by the other categories; in practice, these facets are often logistical moments.

The tutor-action taxonomy presented is organized primarily around the functional direction (scaffolding or rigor) of a tutoring move. This structure differs from other existing tutor-action taxonomies, like the National Tutoring Observatory's Tutor Move taxonomy, which organizes learning support moves along a spectrum of student engagement (Zhou et al., 2026b). By prioritizing pedagogical function, our taxonomy provides a principled way to manage instructional intent, which prior work has identified as a primary driver of classification ambiguity in tutor-action taxonomy creation (Ahtisham et al., 2026).

Furthermore, existing tutor-action taxonomies predominantly focus on scaffolding strategies; rigor strategies in these frameworks are often restricted to prompts for student justification or reasoning (Macina et al., 2023; Zhou et al., 2026b; Ahtisham et al., 2026). In contrast, our taxonomy incorporates more fine-grained rigor strategies, such as withdrawing support to allow independent work or intentionally increasing task complexity, which allows for a more precise evaluation of how tutors shift cognitive demand to the student. This granularity is achieved through our decomposition of multi-turn action facets, which enables the identification

Action	Scaffolding/Rigor	Definition
Guiding/funneling questions toward an answer	Scaffolding	The tutor poses leading questions or step-by-step sub-step prompts that funnel the student toward a specific answer or next move without demanding independent justification.
Breaking the problem into steps	Scaffolding	The tutor decomposes the task into smaller steps, sub-steps, or component parts to make it more manageable.
Explaining, modeling, or co-solving	Scaffolding	The tutor directly explains a concept or procedure, models a worked example, narrates the reasoning, co-solves the problem doing much of the cognitive work, or corrects an error by explaining why it is wrong or re-teaching the concept.
Alternative representations and analogies	Scaffolding	The tutor re-presents the problem or concept in a different form (a diagram, visual, number line, manipulative, real-world analogy, reframing, or restating/rephrasing the problem in different words) to make it more accessible.
Hints, reminders, and narrowing support	Scaffolding	The tutor provides hints, reminders of prior work or rules, highlights key information, or narrows answer options to lower difficulty and nudge the student forward (without re-presenting the whole problem).
Supplying answers, steps, or corrections	Scaffolding	The tutor supplies the result directly (gives away the answer, fills in a step, performs the computation, or tersely states the correct answer without teaching it).
Prompting for explanation, justification, or reasoning	Rigor	The tutor asks the student to explain, justify, or reason about how they arrived at an answer or why it works, or to define/articulate concepts independently, eliciting the reasoning itself.
Withdrawing support / independent work	Rigor	The tutor steps back, withholds help, or hands the task over so the student attempts or completes the work independently.
Increasing complexity or challenge	Rigor	The tutor raises cognitive demand by introducing harder problems, larger numbers, more complex variations, or advancing to more demanding topics.
Prompting self-assessment / reconsideration	Rigor	The tutor prompts the student to evaluate their own answer or reasoning (to rate their confidence, check or verify whether it is correct, reconsider it, or find and fix an error themselves) rather than supplying the correction.
Affirmations, check-ins, and confirming answers	Neutral	The tutor offers praise, encouragement, or reassurance; gauges the student's comfort, understanding, or readiness for the task; or confirms whether an answer is correct.
Transitioning to a new problem or topic	Neutral	The tutor advances the session by moving on to the next problem, question, step, topic, or task without adding challenge or probing understanding.
Other	N/A	Off-task or logistical moments, technical handling, bare stance statements with no concrete move, mechanical navigation, and reading the problem aloud verbatim.

Table 7 The 13-category action taxonomy.

of underlying pedagogical directions that may be obscured if focusing on surface-level tutor dialogue alone.

After creating the action categories, we prompt an Claude Opus 4.8 to classify each scored action facet per tutor into a single category. Notably, the facets are classified independently of the ground-truth scaffolding or rigor labels, ensuring a more fine-grained picture of the strategies employed. We exclude non-actions ("The tutor failed to break the problem into steps") and situation-label descriptions ("The tutor scaffolds") from this taxonomy using a rule-based filter that flags statements whose main verb negates a tutor action ("the tutor does not / fails to / never / makes no / takes no...") or whose content is only a stance phrase ("scaffolds", "pushes for rigor"). In total, we classify 1,967 human tutor action facets and 6,589 LM tutor facets.

4.4 Evaluation Metrics

Our primary evaluation approach is to match the observed tutor strategy to the appropriate strategy determined by teachers' annotated consensus. Since over-scaffolding is generally perceived as undesirable by teachers, we penalize over-scaffolding in all cases, and report the rate at which tutors avoid over-scaffolding separately.

We evaluate over the teacher-identified key moments. Let M_K be all key moments, M_S, M_R the scaffolding- and rigor-appropriate moments respectively, where $M_K = M_S \sqcup M_R$.

Appropriate Scaffolding. Does the tutor scaffold without over-scaffolding in scaffolding-appropriate moments? Because the tutor replays are multi-turn, we do not penalize rigor pushes that accompany scaffolding. We define our appropriate scaffolding metric $\mathcal{S}$ as

$$\mathcal{S} = \frac{1}{|M_S|} \sum_{m \in M_S} s_m, \tag{1}$$

where s_m is an indicator variable that equals 1 if the tutor scaffolds and does not over-scaffold at moment m, and 0 otherwise.

Appropriate Rigor. At rigor-appropriate moments, does the tutor push for rigor? Similar to the appropriate scaffold metric, we do not penalize for also scaffolding over multi-turn replays, as a tutor may push for rigor in one turn, realize they went too far, and introduce a scaffold. However, we do penalize for over-scaffolding, as over-scaffolding is never an effective strategy. We define our appropriate rigor metric as

$$\mathcal{R} = \frac{1}{|M_R|} \sum_{m \in M_R} r_m, \tag{2}$$

where $r_m \in \{0, 1\}$ indicates that at moment m, the tutor pushes for rigor and does not over-scaffold.

Avoids Over-scaffolding. Across all key moments, how often does the tutor avoid over-scaffolding? Given that $o_m \in \{0, 1\}$ is 1 when the tutor does not over-scaffold at moment m, we define our avoid over-scaffolding metric as

$$\mathcal{O} = \frac{1}{|M_K|} \sum_{m \in M_K} o_m. \tag{3}$$

Action divergence. Given that $\mathcal{S}$ and $\mathcal{R}$ are flexible in allowing both pushes for rigor and scaffolding to occur together, we also incorporate metrics that encourage LM tutors to not treat each situation type the same. That is, effective tutors should adapt their strategy and avoid making the same moves for both scaffolding- and rigor-appropriate moments. We calculate the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between our two situation types; this asymmetric metric allows us to separately ask whether tutors concentrate their usage of scaffolding strategies in scaffolding situations and vice versa. For instance, if a tutor withdraws support frequently in rigor moments but rarely in scaffolding moments, this increases $D_{\mathrm{KL}}(\text{rigor} \,||\, \text{scaffolding})$, as seeing the withdrawing support tactic in a scaffolding situation would be more surprising. The reverse direction, $D_{\mathrm{KL}}(\text{scaffolding} \,||\, \text{rigor})$, similarly shows the concentration of scaffolding tactics in scaffolding-appropriate moments.

5 Results

Tutor	Plain prompt			Evaluation-aware prompt		
	Appropriate Scaffolding	Appropriate Rigor	Avoids Over-Scaffolding	Appropriate Scaffolding	Appropriate Rigor	Avoids Over-Scaffolding
Claude Opus 4.8	0.615	0.208	0.462	0.858	0.831	0.896
Claude Sonnet 4.6	0.573	0.085	0.437	0.831	0.792	0.913
DeepSeek V4 Pro	0.485	0.088	0.379	0.712	0.492	0.623
Gemini 2.5 Pro	0.727	0.223	0.546	0.896	0.712	0.854
Gemini 3.5 Flash	0.723	0.158	0.598	0.885	0.646	0.817
GPT 5.5	0.269	0.046	0.173	0.773	0.531	0.665
GPT 5.4 mini	0.181	0.023	0.188	0.738	0.404	0.598

Table 8 "Evaluation-aware" refers to the prompt where LMs are instructed to be generally aware of adapting pedagogical strategies based on rigor- and scaffolding-appropriate situations. Model configuration in Appendix D.4

In this section, we describe the results of applying our LM scoring pipeline onto "replays", where LM tutors take action after teachers' selected cut points. We find that TutorMoments distinguishes model quality, reveals systematic over-scaffolding in default prompting, and provides interpretable evidence about how prompting changes tutoring behavior.

As mentioned in §3.4, we evaluate a total of 520 key moments, with a 50/50 split between scaffolding-appropriate and rigor-appropriate moments. We prompt the LM tutors using a plain prompt and an evaluation-aware prompt that defines the trade-off between scaffolding-rigor (§2.2), and we run replays for 5 tutor-student turns after the cut point. We apply our LM scoring pipeline, developed in §4, onto these replays.

We also run our scoring pipeline on human tutor's decisions, scoped to 5 turns after the cut point. We note that human tutors do not necessarily represent ideal tutoring practice. In fact, teacher annotators were instructed to flag ineffective tutoring moments to compose our key moment dataset, and so our dataset focuses on human tutors' missed opportunities or miscalibrations. Rather, human tutors serve as naturalistic reference points to contextualize LM tutor behavior, rather than a pedagogically normative gold standard. Human tutors perform around the range of plain-prompted LMs, obtaining appropriate scaffold, appropriate rigor, and avoid over-scaffolding scores of 0.458, 0.182, and 0.496, respectively.

We show that, without a consideration of what level of support situations may call for, LMs may not appropriately challenge learners. The plain prompt performance is meaningfully below the evaluation-aware prompt performance across all models (Table 8). In particular, the plain prompt often results in LM tutors failing to push for rigor and frequently over-scaffolding, thus settling into a "helpful" assistant baseline. Even well-regarded models like GPT 5.5 appropriately push for rigor in only 4.6% of cases with a plain prompt. Our results echo those of Bastani et al. (2025), who ran a randomized-controlled trial showing that negative learning effects in a default GPT-4 interface are mitigated with prompt instructions designed by teachers. Everyday users of LMs may resort to prompts even simpler than our plain prompt in standard chat interfaces, e.g. "Act like a tutor" or "Can you tutor me?" (Deng et al., 2024). It is unknown whether popular AI chat interfaces route these underspecified queries through tutoring-specialized harnesses, and so the default behavior of LMs-as-tutors still impacts learners.

LMs generally known to be weaker struggle more to scaffold and push for rigor appropriately. In Table 8, we find that our metrics can differentiate known orderings of model size and general capability. That is, Claude Opus 4.8 outperforms Claude Sonnet 4.6 and GPT-5.5 outperforms GPT-5.4-mini. In addition to recovering known orderings, we are also able to differentiate frontier LMs, suggesting that our benchmark has discriminative power (Akhtar et al., 2026). Though some LMs, such as Gemini 2.5 Pro, are able to achieve

high scores, others, e.g. GPT-5.5, still have much room to improve, even with a targeted prompt. These findings suggest that explicit instructional guidance substantially improves models' pedagogical behavior. At the same time, the remaining gap between prompted performance and teacher judgments indicates that prompting alone is insufficient to produce consistently appropriate instructional decisions.

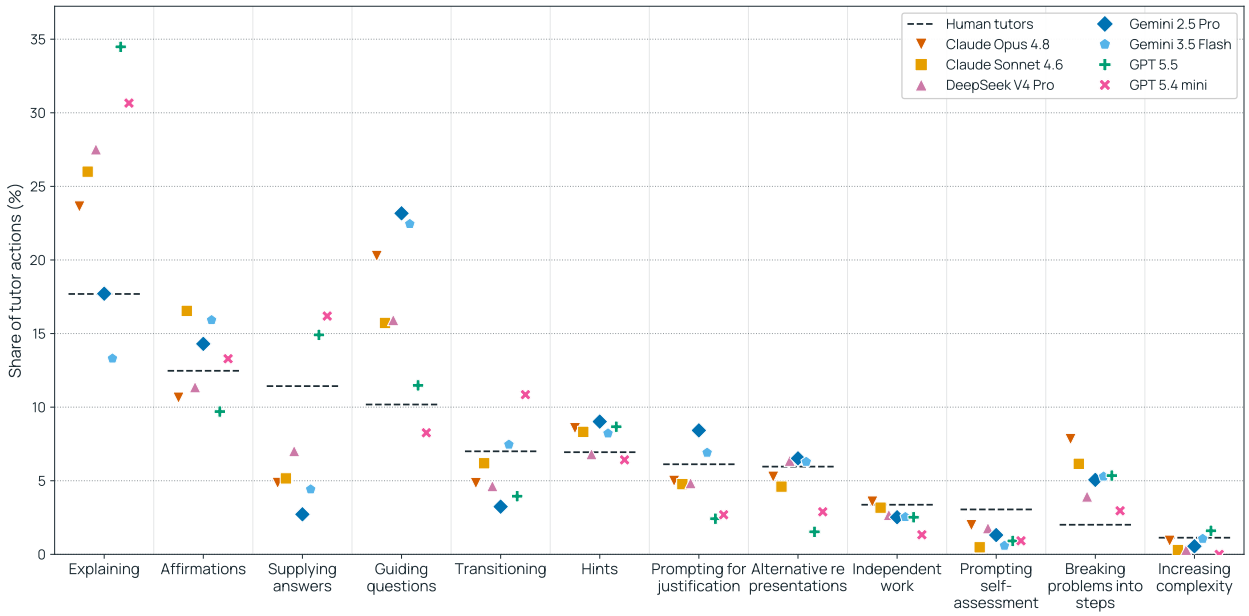
LM tutors' performance on TutorMoments is highly sensitive to prompting strategy. LM developers have suggested prompting to be the main mechanism for teachers and product teams to steer AI tutors (LearnLM Team Google, 2024). Including the evaluation-aware prompt meaningfully improves performance, showing evidence of pedagogical instruction following. Prompts containing multiple instructions that reflect "good" pedagogical practices may be difficult for LMs to consistently satisfy, especially as the number of criteria increase (Wen et al., 2024; Zhou et al., 2023; Jiang et al., 2024). Our framework thus provides a tool for future work to further interrogate tutor prompt design tradeoffs.

LM Tutors employ different pedagogical moves from the human tutors in our dataset. Comparing LM tutor actions to human tutor actions in the same situation can help contextualize patterns observed among LM tutors, with the understanding that the human tutors are imperfect. Figures 4a and 4b show the prevalence of LM tutors' actions across all evaluated key moments compared to human tutor actions. Plain-prompted LM tutors predominantly rely on explanation (13.31 - 34.48% of actions, versus 17.7% for humans) and asking guiding questions (15.92 - 23.17% of actions, versus 10.2% for humans), exhibiting a pattern of tutor-directed scaffolding that assigns more cognitive load to the tutor than to the student. On the other hand, for most evaluation-aware LM tutors, explanations make up a smaller share of actions (2.7% to 14.99%); prompting for justification is instead employed at rates 2.8 to 5.1 times higher than human tutors do, shifting the cognitive load more on the student than the tutor. Generally, our action taxonomy enables more fine-grained insights into *how* different tutors may vary in their deployment of strategies. The behavioral profiles in Figures 4a and 4b illustrate an important advantage of TutorMoments: beyond producing aggregate benchmark scores, it identifies how models differ in their instructional strategies.

Evaluation-aware LM tutors adapt their pedagogical approach to the context, while plain-prompted LM tutors adapt less to conversational context. Good tutors should not respond to every situation in the same way; they should adapt their instructional strategies to the learning context. Plain-prompted LM tutors exhibit lower or similar sensitivity to differences in situations (D_{KL} ranging from 0.048 to 0.182) when compared to human tutors ($D_{\mathrm{KL}} = 0.177$-179), suggesting that they behave very similarly across situation types, with only Claude Opus 4.8 matching the human baseline. Most evaluation-aware LM tutors, however, meet or exceed human tutors' situation sensitivity, with D_{KL} ranging from 0.119 to 0.345. Still, the divergence of actions for some models, e.g. Claude Opus 4.8, between the two prompts does not differ strongly. This suggests the possibility that LMs may be able to perform well on metrics in Table 8 by employing similar scaffolding-*and* rigor-relevant actions across situations, thus necessitating the inclusion of action divergence metrics to differentiate true situational adaptability.

	Plain prompt		Evaluation-aware prompt	
Tutor	$D_{\mathrm{KL}}(\mathrm{S}\|\|\mathrm{R})$	$D_{\mathrm{KL}}(\mathrm{R}\|\|\mathrm{S})$	$D_{\mathrm{KL}}(\mathrm{S}\|\|\mathrm{R})$	$D_{\mathrm{KL}}(\mathrm{R}\|\|\mathrm{S})$
Human tutors	0.179	0.177	0.179	0.177
Claude Opus 4.8	0.173	0.182	0.181	0.181
Claude Sonnet 4.6	0.143	0.132	0.310	0.345
DeepSeek V4 Pro	0.141	0.142	0.170	0.198
Gemini 2.5 Pro	0.154	0.122	0.207	0.228
Gemini 3.5 Flash	0.160	0.153	0.199	0.203
GPT 5.5	0.048	0.050	0.119	0.128
GPT 5.4 mini	0.128	0.143	0.147	0.153

Table 9 KL divergence between scaffolding-appropriate (S) and rigor-appropriate moment (R) distributions, computed in both directions, for human tutors and each LM tutor under plain and evaluation-aware prompts.



(a) Plain-prompted LM tutor.

(b) Evaluation-aware LM tutor.

Figure 4 Distribution of tutor actions across the action taxonomy defined in Table 7 for human tutors (dashed reference) and the 7 LM tutors, under the plain prompt (top) and evaluation-aware prompt (bottom).

Evaluation-aware LM tutors show stronger and narrower adaptions in rigor situations than scaffolding ones. Figures 5 and 6 show shifts in distributions of action categories, conditioned on situation type. Tutors generally increase the usage of relevant strategies in their appropriate situations. Evaluation-aware LM tutors show $D_{\mathrm{KL}}(\text{rigor} \mid\mid \text{scaffolding}) > D_{\mathrm{KL}}(\text{scaffolding} \mid\mid \text{rigor})$; they employ scaffolding strategies more ubiquitously across both situation types, but tend to reserve rigor strategies more specifically for rigor-appropriate situations. This pattern is not consistent among plain-prompted LM tutors, with several models showing a reverse pattern. Within rigor-appropriate moments, evaluation-aware LM tutors concentrate their situational adjustment almost entirely in prompting for justification. This is a far narrower response than what the same models tend to do in scaffolding situations: in such cases, adjustments are more evenly balanced across multiple categories of actions (e.g., asking guiding questions, breaking problems into steps, and using alternative representations). Prompting for justification is well-documented as an effective metacognitive strategy (Maclellan and Soden, 2004; Mata-Pereira and Ponte, 2018; Kurniawan et al., 2022) which may explain its prominence in rigor-appropriate moments; scaffolding, by contrast, affords a broader set of established methods that LM tutors could potentially draw on more evenly.

Lower latency models have the most room for improvement in our framework Figure 7 shows aggregated performance against average LM tutor response latency from request to last token received under the evaluation-aware prompt. The highest performing models also have the highest latency, with Gemini 2.5 Pro and Claude Opus 4.8 averaging over 10 seconds to complete a response as an LM tutor, which may be unrealistic for production deployments in education technology. Low latency LM tutors are a targeted area of improvement in the field. Our framework can support smaller model post-training efforts by providing behavior-specific signals for improving tutor quality.

6 Conclusion

Effective AI tutoring depends on selecting instructional strategies that match the learner's current understanding and the pedagogical demands of the moment. TutorMoments provides a framework for evaluating whether AI tutors make contextually appropriate instructional decisions. We leverage experienced teachers' expertise to create task conditions for evaluating LM tutors, focusing on situations that target the question of whether tutors provide the right assistance at the right moment. Our scoring pipeline examines whether tutors scaffold during scaffolding-appropriate moments and push for rigor during rigor-appropriate moments, while avoiding over-scaffolding. In addition, we expect LM tutors to adapt to different situations, employing diverging moves based on student behavior and session context. Thus, TutorMoments evaluates whether AI tutors adapt their instructional decisions to the pedagogical demands of the learning context.

Recent human-AI interaction research has dedicated much attention towards the ways in which AI reduces critical thinking and skill formation (Shen and Tamkin, 2026; Zhi et al., 2026). Our work provides a pedagogical perspective on this topic, with a focus on AI's application in math education. Overall, we find that dialogic moves that correspond to maintaining and increasing cognitive rigor are more difficult for LMs to identify and more difficult for them to employ in tutoring settings. We also find that prompt and model choice can greatly affect LMs' abilities to respond to scaffolding- and rigor-appropriate situations, and can shift the distribution of pedagogical strategies that appear in such moments. For instance, explicit evaluation-aware guidance in prompts enables current frontier models to exhibit substantially stronger pedagogical behavior than their default tutoring shows, though models still fall short of consistently matching teacher judgments. We contribute a framework that allows educators and education technology product developers to interrogate tutor design decisions. By making contextually appropriate instructional decision making measurable, TutorMoments creates an opportunity to optimize future AI tutors toward pedagogically appropriate instruction.

Our work serves as a preview for a larger multimodal dataset and benchmark release, which will include additional annotations using a refined annotation pipeline and accompanying scoring pipeline. We release TutorMoments-Preview with 462 human student-tutor tutoring session transcripts, LM replays of 520 moments, and several thousand free-text annotations from teachers.

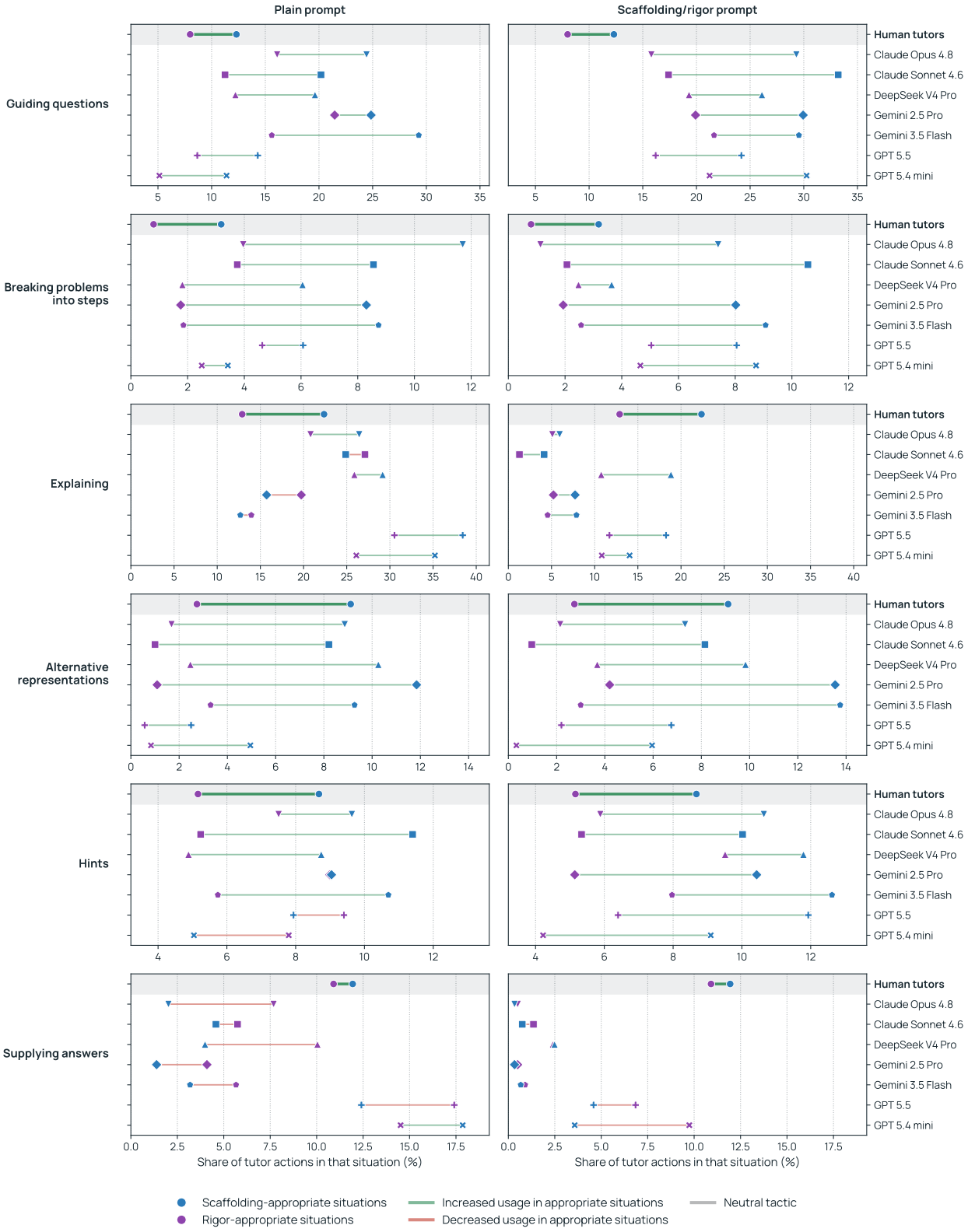


Figure 5 Distribution of **scaffolding** strategies in scaffolding- versus rigor-appropriate moments, for human tutors and each LM tutor under plain and evaluation-aware prompts.

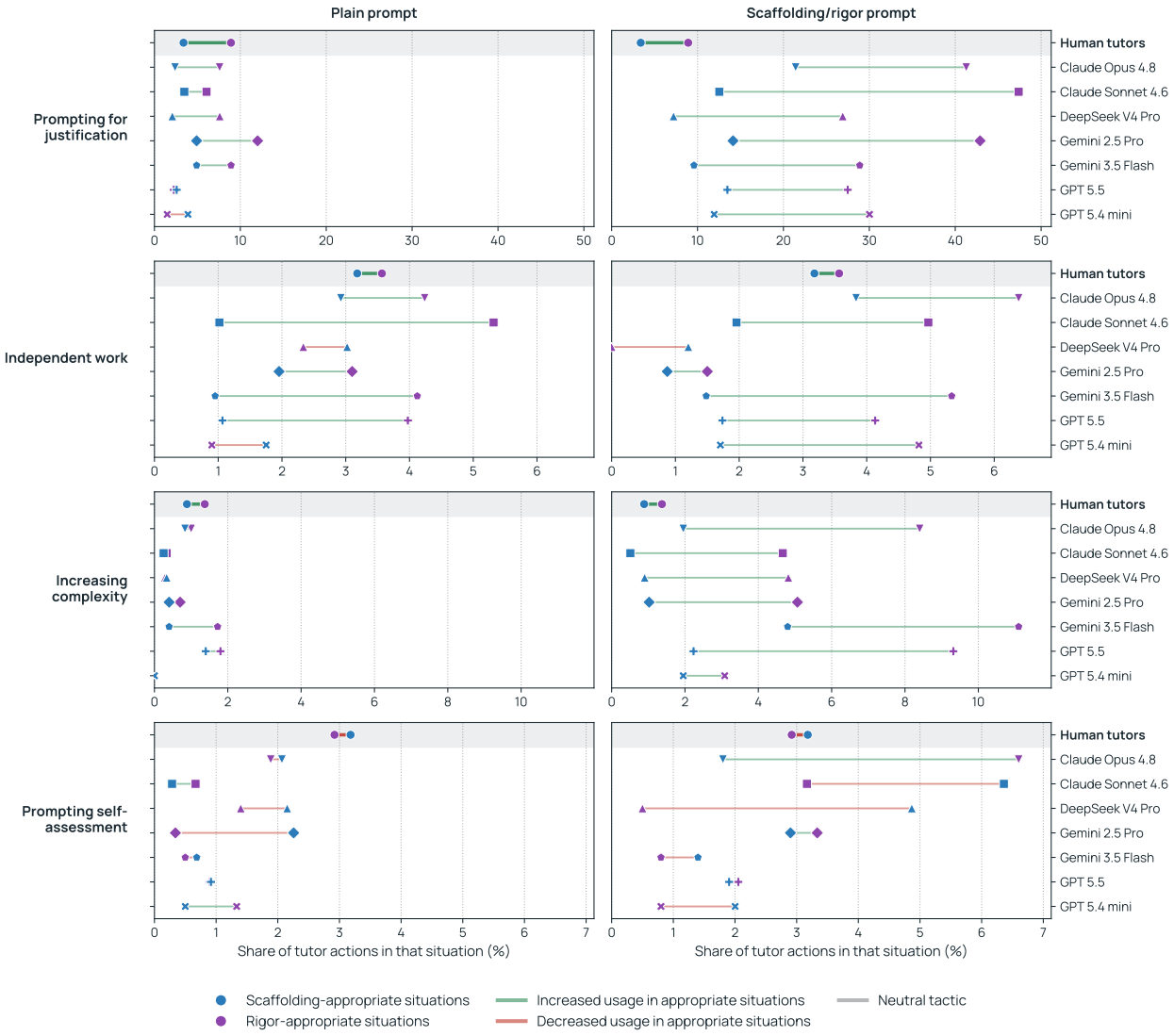


Figure 6 Distribution of **rigor** strategies in scaffolding- versus rigor-appropriate moments, for human tutors and each LM tutor under plain and evaluation-aware prompts.

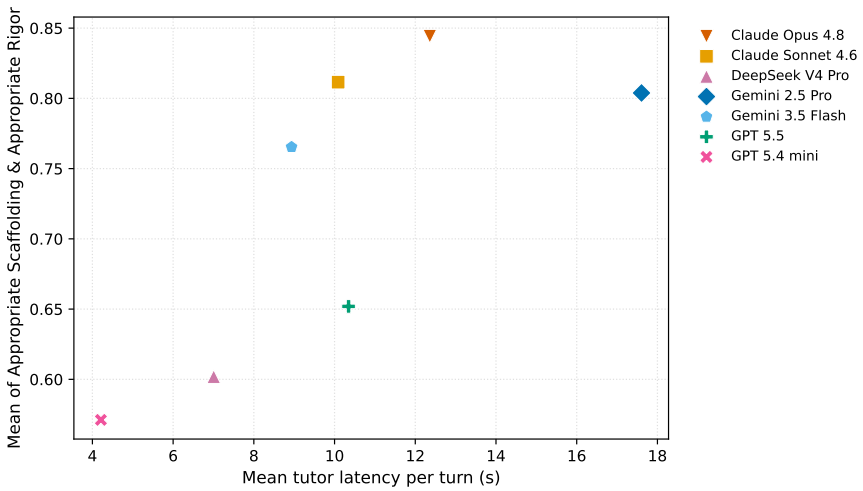


Figure 7 Mean tutor latency per turn vs. mean of scaffolding and rigor score of scaffolding-rigor aware prompt. n = 520).

7 Limitations

We acknowledge several limitations of our work.

Given our focus on scaffolding- and rigor-related decision points, TutorMoments does not comprehensively capture overall tutoring quality. Our framework evaluates instructional decision making at teacher-identified pedagogical moments. So, we focus on cut points where it is known what a tutor should do, and it is possible that AI tutors may misbehave in other ways or elsewhere in the transcript. Complementary studies with real students are essential for understanding downstream learning outcomes. Though automated evaluation can provide some signal of models' downstream utility, it cannot supersede human studies. The complete evaluation of AI in education requires consideration of situated practice in actual learning contexts, and hill-climbing on isolated metrics may distort or narrow the breadth of ways in which learning occurs and may be measured (Yin et al., 2026). Though we incorporate values and views of multiple educators in our development process, we have not yet examined whether tutors who adapt to scaffolding-appropriate and rigor-appropriate situations, as determined by our metrics, actually yield improvements in real students' learning.

In addition, though we took care to validate the various steps of our pipeline that incorporate LM classifiers and judges, we hope for future work to continue improving and expanding these components to support more robust evaluation of LM tutors. For example, though our use of an "oracle" student during replay helps ensure the simulated student turns are realistic, they restrict the evaluation methodology to assessing LM tutor moves. Future work in high-fidelity students simulation would enable assessment of student-learning outcomes.

Other limitations relate to the data underlying the current instantiation of our framework. We are naturally limited by the scope of our human student-tutor transcripts and annotator pool. That is, the behaviors and perspectives in our dataset may not necessarily generalize to other learning contexts, such as those outside of the U.S. or in other subject areas.

8 Acknowledgments

This project has been made possible in part through support from the Gates Foundation and Learning Commons. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Gates Foundation or other funders. We thank the following individuals for their support: Danielle May, Janice Dow, Kevin Miu, Will Smith, Russell Greenwald, Lia Bonamassa, Zara Malik, Bryan Richardson, Adam Goldfarb, John Whitmer, and Pradeep Dasigi for project management, data acquisition, legal assistance and mentorship. We thank Kristen Dicerbo, Adam Goldfarb, Laurence Holt,

John Whitmer, Kirk Vanacore, Rene Kizilcec and Clayton Cohn for helpful comments on an earlier draft.

References

B. Ahtisham, K. Vanacore, J. Lee, Z. Zhou, D. Pietrzak, and R. F. Kizilcec. Ai annotation orchestration: Evaluating llm verifiers to improve the quality of llm annotations in learning analytics. In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, page 447–456, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400720666. doi: 10.1145/3785022.3785095. URL https://doi.org/10.1145/3785022.3785095.

M. Akhtar, A. Reuel, P. Soni, S. Ahuja, P. S. Ammanamanchi, R. Rawal, V. Zouhar, S. Yadav, C. Whitehouse, D. Ki, J. Mickel, L. Choshen, M. Suppa, J. Batzner, J. Chim, J. Sania, Y. Long, H. A. Rahmani, C. Q. Knight, Y. Nan, J. Raj, Y. Fan, S. Singh, S. Sahoo, E. Habba, U. Gohar, S. M. Pawar, R. Scholz, A. Subramonian, J. Ni, M. Kochenderfer, S. Koyejo, M. Sachan, S. Biderman, Z. Talat, A. Ghosh, and I. Solaiman. When AI benchmarks plateau: A systematic study of benchmark saturation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=YC1Otscjbs.

Anthropic. Claude's character. https://www.anthropic.com/research/claude-character, June 2024. Accessed 2026-06-25.

A. T. Barlow, N. E. Gerstenschlager, J. F. Strayer, A. E. Lischka, D. C. Stephens, K. S. Hartland, and J. C. Willingham. Scaffolding for access to productive struggle. *Mathematics Teaching in the Middle School*, 23(4):pp. 202–207, 2018. ISSN 10720839, 23285486. URL https://www.jstor.org/stable/10.5951/mathteacmiddscho.23.4.0202.

H. Bastani, O. Bastani, A. Sungu, H. Ge, O. Kabakcı, and R. Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26): e2422633122, 2025. doi: 10.1073/pnas.2422633122. URL https://doi.org/10.1073/pnas.2422633122.

R. A. Bjork and E. L. Bjork. Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4):475–479, 2020. doi: 10.1016/j.jarmac.2020.09.003. URL https://doi.org/10.1016/j.jarmac.2020.09.003.

P. Black and D. Wiliam. Assessment and classroom learning assessment in education. principles, policy & practice, 5 (1), 7–74, 1998.

K. Boyer, R. Phillips, E. Y. Ha, M. Wallis, M. Vouk, and J. Lester. Leveraging hidden dialogue state to select tutorial moves. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–73, Los Angeles, California, June 2010. Association for Computational Linguistics. URL https://aclanthology.org/W10-1009/.

S. Chang, A. Anderson, and J. M. Hofman. ChatBench: From static benchmarks to human-AI evaluation. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26009–26038, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1262. URL https://aclanthology.org/2025.acl-long.1262/.

N. R. Council, B. on Behavioral, S. Sciences, C. on Developments in the Science of Learning with additional material from the Committee on Learning Research, and E. Practice. *How people learn: Brain, mind, experience, and school: Expanded edition*, volume 1. National Academies Press, 2000.

Y. Deng, W. Zhao, J. Hessel, X. Ren, C. Cardie, and Y. Choi. WildVis: Open source visualizer for million-scale chat logs in the wild. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2024. URL https://aclanthology.org/2024.emnlp-demo.50/.

J. Draeger, P. del Prado Hill, L. R. Hunter, and R. Mahler. The anatomy of academic rigor: The story of one institutional journey. *Innovative Higher Education*, 38(4):267–279, 2013.

Everett Teachers' Association and Everett School Committee. Draft Memorandum of Agreement for the Period of July 1, 2024 to June 30, 2027, Nov. 2024. URL https://static1.squarespace.com/static/5a650fbc8c56a87958d5f5b8/t/67acbf8333bfda53e04f9af0/1739374467824/Unit_A_MOA_for_Release_11_24+%281%29.pdf. Everett Public Schools Teachers Unit A.

A. C. Graesser and N. K. Person. Question asking during tutoring. *American Educational Research Journal*, 31(1): 104–137, 1994. doi: 10.3102/00028312031001104.

A. C. Graesser, N. K. Person, and J. P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6):495–522, 1995. doi: 10.1002/acp.2350090604.

M. Henningsen and M. K. Stein. Mathematical tasks and student cognition: Classroom-based factors that support and inhibit high-level mathematical thinking and reasoning. *Journal for research in mathematics education*, 28(5): 524–549, 1997.

M. Heritage. Formative assessment: What do teachers need to know and do? *Phi Delta Kappan*, 89(2):140–145, 2007.

J. Hiebert and D. A. Grouws. The effects of classroom mathematics teaching on students' learning. In F. K. J. Lester, editor, *Second Handbook of Research on Mathematics Teaching and Learning*, pages 371–404. Information Age Publishing, Charlotte, NC, 2007.

L. Ibrahim, K. M. Collins, S. S. Y. Kim, A. Reuel, M. Lamparth, K. Feng, L. Ahmad, P. Soni, A. E. Kattan, M. Stein, S. Swaroop, V. Padmakumar, I. Sucholutsky, A. Strait, D. Yang, Q. V. Liao, and U. Bhatt. Measuring and mitigating overreliance to build human-compatible AI, 2026. URL https://arxiv.org/abs/2509.08010.

Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu, and W. Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.257. URL https://aclanthology.org/2024.acl-long.257/.

H. Jin, M. Yoo, J. Park, Y. Lee, X. Wang, and J. Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3714054. URL https://doi.org/10.1145/3706598.3714054.

M. Kapur. Productive failure. *Cognition and Instruction*, 26(3):379–424, 2008.

Y. Ke, L. Jin, J. C. L. Ong, A. J. Thirunavukarasu, J. Car, C. Y. Cheung, Y. C. Tham, D. S. W. Ting, M. E. H. Ong, S. Compton, et al. Ai-induced never-skilling in medical education. *Nature medicine*, pages 1–10, 2026.

K. R. Koedinger and V. Aleven. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264, 2007. doi: 10.1007/s10648-007-9049-0. URL https://doi.org/10.1007/s10648-007-9049-0.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

A. Kupor, C. Morgan, and D. Demszky. Measuring five accountable talk moves to improve instruction at scale, 2023. URL https://arxiv.org/abs/2311.10749.

S. Kurniawan, R. Rosjanuardi, and A. Aswin. Mathematical justification research in mathematics education across grades: A systematic literature review. *Hipotenusa : Journal of Mathematical Society*, 4:134–147, 12 2022. doi: 10.18326/hipotenusa.v4i2.7815.

LearnLM Team Google. LearnLM: Improving Gemini for learning, 2024. URL https://arxiv.org/abs/2412.16429.

J. Lin, N. Tomlin, J. Andreas, and J. Eisner. Decision-oriented dialogue for human-AI collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911, 2024. doi: 10.1162/tacl_a_00679. URL https://aclanthology.org/2024.tacl-1.50/.

C. Liu, Q. Zhou, X. Shen, X. Bruce Liu, T. Wu, and X. Chen. Behavioral indicators of overreliance during interaction with conversational language models. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2026.

J. Macina, N. Daheim, S. P. Chowdhury, T. Sinha, M. Kapur, I. Gurevych, and M. Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *ArXiv*, abs/2305.14536, 2023.

J. Macina, N. Daheim, I. Hakimi, M. Kapur, I. Gurevych, and M. Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China, Nov. 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.11. URL https://aclanthology.org/2025.emnlp-main.11/.

E. Maclellan and R. Soden. The importance of epistemic cognition in student-centred learning. *Instructional Science*, 32:253–268, 05 2004. doi: 10.1023/B:TRUC.0000024213.03972.ce.

S. Maiya, H. Bartsch, N. Lambert, and E. Hubinger. Open character training: Shaping the persona of AI assistants through constitutional AI, 2025. URL https://arxiv.org/abs/2511.01689.

J. Mata-Pereira and J. Ponte. Teacher's actions to promote students' justifications. *Acta Scientiae*, 20, 07 2018. doi: 10.17648/acta.scientiae.v20iss3id3910.

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)*, pages 1234–1251, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.57. URL https://aclanthology.org/2025.naacl-long.57/.

N. Murphy and D. Messer. Differential benefits from scaffolding and children working alone. *Educational Psychology*, 20(1):17–31, 2000.

T. Naous, P. Laban, W. Xu, and J. Neville. Flipping the dialogue: Training and evaluating user language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ykSmkVqzn4.

National Council of Teachers of Mathematics. *Principles to Actions: Ensuring Mathematical Success for All*. National Council of Teachers of Mathematics, Reston, VA, 2014. ISBN 978-0873537742. URL https://www.nctm.org/PtA/.

V. Padmakumar, L. Ibrahim, Z. Z. Wang, J. Wang, Q. V. Liao, and D. Yang. Offloading score: Measuring AI reliance through counterfactual workflows. *arXiv preprint arXiv:2605.29392*, 2026.

D. D. Paige, G. S. Smith, and J. M. Sizemore. Conceptualizing rigor and its implications for education in the era of the common core. *Cogent Education*, 2(1):1048084, 2015.

S. L. Payne, K. L. Kleine, J. Purcell, and G. R. Carter. Evaluating academic challenge beyond the nsse. *Innovative Higher Education*, 30(2):129–146, 2005.

J. Perczel, J. Chow, and D. Demszky. Teachlm: Post-training LLMs for education using authentic learning data, 2025. URL https://arxiv.org/abs/2510.05087.

P. Röttger, B. Vidgen, D. Hovy, and J. Pierrehumbert. Two contrasting data annotation paradigms for subjective NLP tasks. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.13. URL https://aclanthology.org/2022.naacl-main.13/.

A. Scarlatos, J. Lee, S. Woodhead, and A. Lan. Simulated students in tutoring dialogues: Substance or illusion? In M. Liakata, V. P. Moreira, J. Zhang, and D. Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 42349–42385, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. URL https://aclanthology.org/2026.acl-long.1960/.

J. H. Shen and A. Tamkin. How AI impacts skill formation. In *Post-AGI Science and Society Workshop*, 2026. URL https://openreview.net/forum?id=WpL9rAd0K7.

A. Suresh, J. Jacobs, C. Harty, M. Perkoff, J. H. Martin, and T. Sumner. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.497/.

S. Tao, M. Lan, M. Wang, and H. Li. Potential risks of generative artificial intelligence integration into K-12 education: A scoping review. *Computers and Education: Artificial Intelligence*, page 100561, 2026.

J. Van de Pol, M. Volman, and J. Beishuizen. Scaffolding in teacher–student interaction: A decade of research. *Educational psychology review*, 22(3):271–296, 2010.

L. S. Vygotsky. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA, 1978.

K. Wataoka, T. Takahashi, and R. Ri. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=tLZZZIgPJX.

D. Weitekamp, M. N. Siddiqui, and C. J. MacLellan. Tutorgym: A testbed for evaluating AI agents as Tutors and Students. In *Artificial Intelligence in Education: 26th International Conference, AIED 2025, Palermo, Italy, July*

22–26, 2025, Proceedings, Part III, page 361–376, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-98419-8. doi: 10.1007/978-3-031-98420-4_26. URL https://doi.org/10.1007/978-3-031-98420-4_26.

B. Wen, P. Ke, X. Gu, L. Wu, H. Huang, J. Zhou, W. Li, B. Hu, W. Gao, J. Xu, Y. Liu, J. Tang, H. Wang, and M. Huang. Benchmarking complex instruction-following with multiple constraints composition. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=U2aVNDrZGx.

D. Wood, J. S. Bruner, and G. Ross. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100, 1976. doi: https://doi.org/10.1111/j.1469-7610.1976.tb00381.x. URL https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.1976.tb00381.x.

S. A. Wyse and P. A. Soneral. "Is this class hard?" Defining and analyzing academic rigor from a learner's perspective. *CBE—Life Sciences Education*, 17(4):ar59, 2018.

Y. Yin, S. Karumbaiah, Y. Kim, and D. Saxena. When AI breaks, teachers repair: Pedagogical repair work in situated classroom practice. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 2206–2224, 2026.

J. R. Young, D. Bevan, and M. Sanders. How productive is the productive struggle? lessons learned from a scoping review. *International Journal of Education in Mathematics, Science and Technology*, 12(2):470–495, 2024.

M. Zheng, J. Pei, L. Logeswaran, M. Lee, and D. Jurgens. When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.888. URL https://aclanthology.org/2024.findings-emnlp.888/.

J. Zhi, H. Kumar, and M. Lee. Investigating the effects of LLM use on critical thinking under time constraints: Access timing and time availability. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791796. URL https://doi.org/10.1145/3772318.3791796.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

X. Zhou*, H. Zhu*, L. Mathur, R. Zhang, Z. Qi, H. Yu, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, and M. Sap. Sotopia: Interactive evaluation for social intelligence in language agents. In *ICLR*, 2024. URL https://openreview.net/forum?id=mM7VurbA4r.

Z. Zhou, K. Vanacore, B. Ahtisham, J. Lee, D. Pietrzak, D. Hedley, J. Dias, C. Shaw, R. Schäfer, and R. F. Kizilcec. Utility-preserving de-identification for math tutoring: Investigating numeric ambiguity in the MathEd-PII benchmark dataset, 2026a. URL https://arxiv.org/abs/2602.16571.

Z. Zhou, K. Vanacore, T. Thompson, J. St John, and R. Kizilcec. Tutor move taxonomy: A theory-aligned framework for analyzing instructional moves in tutoring, 2026b. URL https://arxiv.org/abs/2603.05778.

A De-Identification

A.1 Summary

We implemented LLM-based de-identification, building off (Zhou et al., 2026a), which was developed specifically for de-identifying math tutoring sessions. We expanded the taxonomy of PII over the taxonomy implemented by (Zhou et al., 2026a), resulting in better performance in our held-out test data relative to the (Zhou et al., 2026a) strategy. Performance on types of PII present in both taxonomies was comparable.

We first reworked the (Zhou et al., 2026a) approach via small, proof-of-concept tests, and then applied an initial version of our updated prompt on 250 transcripts. We manually annotated those 250 transcripts for instances of PII. We then continued prompt iteration, using the 250 transcripts as a training set. A second validation round using our improved prompt on 50 additional transcripts found zero missed instances of true personal identifiers. Finally, a head-to-head re-run of the Zhou et al. (2026a)'s original approach on the same 50 test transcripts shows that our final prompt meaningfully outperforms Zhou et al. (2026a) due to covering a broader base of potential PII.

A.2 PII Detection Approach

Zhou et al. (2026a) and the National Tutoring Observatory (NTO) created MathEd-PII, a 1,000-transcript benchmark for evaluating de-identification of math tutoring dialogue. Labels were produced by an LLM annotation pipeline refined through human review of a 100-transcript sample where humans gaged the reasonableness of LM annotations (no unredacted ground truth was available). They evaluated a math-aware prompting strategy (F1 = 0.802, Opus 4.5) and a segment-aware prompting strategy (F1 = 0.820, Opus 4.5), both of which substantially outperformed both basic prompting and a Presidio baseline (F1 = 0.379).

The Segment-aware strategy adds a pre-processing step that labels each turn MATH/NON-MATH (via math-vocabulary scoring plus embedding-based segmentation), then feeds that label into the detection prompt. Math-aware skips this — a single detection call whose prompt includes instructions about how to handle math. Segment-aware's extra step yields only a marginal F1 gain, so we started with the simpler math-aware approach.

Zhou et al. (2026a) pass each turn separately to the LM. In proof-of-concept experiments, we found that this results in significant over-redaction when math word problems are read aloud, as the models don't have the context to know that "Jane has 6 glasses of milk" is part of a word problem and not PII. Instead, we pass the entire transcript with structured json output of turn identifiers and surrogates in a detect phase. We then apply the surrogates by parsing the output and substituting the surrogates into the transcript.

Zhou et al. (2026a)'s PII taxonomy includes 17 types of PII: `AGE`, `COURSE` (e.g. "algebra 100", "CS 124"), `DATE`, `EMAIL ADDRESS`, `GRADE LEVEL`, `IP ADDRESS`, `LOCATION` (defined as "geographic subdivisions smaller than a State"), `NRP` (nationality, religious, and political groups), `PERSON` (names), `PHONE NUMBER`, `SCHOOL`, `SOCIAL HANDLE`, `URL`, `US BANK NUMBER`, `US DRIVER LICENSE`, `US PASSPORT`, and `US SSN`.

During our prompt adaptation and improvement process, we retain `GRADE LEVEL` since the grade level is important to understanding the appropriateness of the tutor's feedback and is supplied as transcript metadata with our transcripts. We add four additional categories based on patterns observed in our transcripts:

- `ORGANIZATION` Non-school entities: workplaces, churches, charities, clubs, community programs, after-school activities. Surrogates with Category-appropriate generic (e.g., "a community program", "a local church", "an after-school program"). Included to limit the potential for identification when combined with place-specific information. In practice most organizations surrogated were Ed Tech platforms. With the number of Ed Tech data breaches recently, we did not want a platform data breach to be combinable with released transcripts to identify students.
- `USERNAME` Online usernames or account names. Surrogated with specific fake username.
- `PASSWORD` Passwords or access codes spoken aloud. Surrogated with specific fake password
- `PET NAME` Names of pets discussed in chitchat (e.g., "my dog Cooper"). Included because pet names can be unusual / unique, which could contribute to re-identification. Surrogated with a specific fake pet name.

A.3 PII Evaluation Strategy

We aggregated common PII taxonomy types into two overarching categories: **true PII** covering direct identifiers including names, pet names, usernames, passwords, IDs and URLs, and **potential PII** covering contextual or quasi-identifiers that only identify a person when combined with other context: age, date, location, organization, school, pet names and NRP.

We ran our initial prompt using Claude Opus 4.6 on a pool of 250 transcripts.

Each transcript was loaded into a custom validation app that surfaces the original text alongside the surrogate text. Reviewers flagged two kinds of issues:

- Missed PII: PII that remains in the transcript.
- Over-redactions: non-PII spans that were redacted unnecessarily.

Six human reviewers, recruited from one co-author's organization, produced 262 reviews covering all 250 transcripts, flagging 550 potential issues.

Co-authors then revisited each and every one of the 550 flagged instances, with reference to the original transcript text where ambiguous. They assigned each flag a PII taxonomy category and a decision on whether the span should have been redacted or not. Key findings from this audit include:

- The genuine missed-PII rate was ~3%
- 81% of genuine misses fell into two clean patterns: login spell-outs ("my username is g-r-a-c-e ...") and name-variant propagation across turns (an honorific + surname in one turn becoming bare-surname in a later turn).

Based on hand-categorized flags, we made seven targeted edits to the detect prompt:

- M1: Improved state & country carve-outs, where we do not redact "California", "Mexico", ... in everyday conversational use.
- M2: Added `PET_NAME` as a new redaction category. Pets are discussed during rapport building, and a child's pet name could contribute to identification.
- M3: Included an explicit list of public edtech brands (e.g. Khan Academy, Blooket, ReadWorks, IXL) to redact as `ORGANIZATION`.
- M4: Included an explicit list of tutoring platform variants, and instruction to always redact them.
- M5: Added a rule for login spell-outs; if the LM observes characters being dictated one-at-a-time, treat them as a `USERNAME` token.
- M6: Handling name-variant propagation; if "Mr. Smith" appears earlier in the transcript, redact "Smith" alone on later turns.

The improved prompt is what we refer to in the rest of this appendix as the finalized prompt.

Two human annotators then validated 50 net-new transcripts (to yield 300 human validated transcripts in total). We re-ran detection on the 50 new transcripts with the finalized prompt. The researchers again walked through every flag and assigned categorization decisions. The categorized output is the held-out test set used in our evaluation comparison with Zhou et al. (2026a). This test set contains $N = 677$ rows: 172 reviewer flags + 505 "implicit-pass" entries, where implicit-pass represents redactions the reviewers implicitly approved by not flagging as an issue.

On the second validation round, reviewers flagged 0 instances of genuine true-PII misses on the 50 held-out transcripts. The remaining flags were a mix of:

- Potential-PII misses – numeric ages, venue-style locations, edtech company transcription variants.
- Placeholder-tag leaks – a small number of <State> pre-existing placeholder tags slipped through unredacted, a known side-effect of the M1 state carve-out.
- State & country over-redactions – Surrogates assigned to locations larger than a US state.

In other words, on a held-out sample of 50 transcripts the iterated prompt produces zero un-redacted direct personal identifiers, and the residual gaps are confined to categorical or placeholder issues that are not direct identifiers.

A.4 Comparison with NTO's prompt

To benchmark our finalized prompt against prior work, we re-ran Zhou et al. (2026a)'s math aware prompt against the 50 held-out transcripts using the same model (Claude Opus 4.6) and then scored both surrogate maps against the n = 677 held-out test eval.

The Zhou et al. (2026a) prompt was used verbatim apart from revised output formatting instructions (rewritten to JSON with `turn_number` and `original_text` so the response slots into our per-transcript pipeline), and the prompt was applied to the entire transcript instead of turn-by-turn. All four task instructions, the 17-type taxonomy, the math-aware language, and the ±3-message evaluation window were unchanged. Detection cost for the NTO run was $15.41, and for our approach was $17.92.

In our evaluation, we removed grade-level decisions from the comparison below, since judging the Zhou et al. (2026a)'s prompt against a category we intentionally dropped is not informative.

Metric	Our prompt	NTO prompt	Δ
True positives (TP)	562	440	−122
False negatives (FN)	23	145	+122
False positives (explicit)	6	6	0
False positives (closed-world)	6	32	+26
Recall	96.1%	75.2%	−20.9 pp
Precision (explicit)	98.9%	98.7%	−0.2 pp
Precision (closed-world)	98.9%	93.2%	−5.7 pp
F1 (explicit)	97.5%	85.4%	−12.1 pp
F1 (closed-world)	97.5%	83.3%	−14.2 pp

Table 10 Evaluation comparison between the NTO prompt (Zhou et al., 2026a) and our prompt, where Δ = NTO - our approach. "Closed-world" FPs include over-redactions the reviewers did not explicitly flag but that fall outside what either pipeline should redact (under the closed-world assumption that any over-redaction in a transcript reviewed as issue-free would have been caught had it been there).

Bucket	Rows	Our prompt			NTO prompt		
		Recall	Precision (cw)	F1 (cw)	Recall	Precision (cw)	F1 (cw)
True PII	347	99.7%	100.0%	99.9%	93.9%	97.9%	95.9%
Potential PII	207	93.2%	100.0%	96.5%	40.1%	81.4%	53.7%
State & country carve-outs	6	—	0.0%	—	—	0.0%	—

Table 11 The comparison gets sharper when split by true PII and potential PII. Here, "cw" refers to "closed-world" false positives: over-redactions the reviewers did not explicitly flag but that fall outside what either pipeline should redact.

Where the buckets break down further:

- **True-PII recall gap (99.7% vs 93.9%):** the NTO prompt's 17-type taxonomy has no `PET_NAME` slot—it misses 16 of 17 pet-name mentions. It additionally misses 3 named-individual occurrences, 1 URL, and 1 username that our prompt catches. The two prompts catch named individuals at near-parity, but every named-individual miss on the held-out set comes from the NTO side.
- **Potential-PII recall gap (93.2% vs 40.1%):** the NTO taxonomy has no `ORGANIZATION` slot, so it misses 47 of 49 mentions of tutoring platforms, 23 of 23 mentions of public edtech tools, and 20 of 21 generic organizations (e.g. YMCA, Boys & Girls Club). Ninety of the 124 potential-PII NTO misses are organizational identifiers

the taxonomy can't represent. The remainder is general under-detection: dates (12 / 21 missed), locations (14 / 67), ages (8 / 24).

- **Potential-PII precision gap (100% vs 81.4%):** closed-world over-redactions on the NTO side concentrate in `LOCATION`. The NTO prompt redacts geographic-sounding strings inside math word problems (e.g. "Underwood", "Caldicot") that our finalized prompt leaves untouched.
- **Where the NTO prompt does better:** placeholder propagation (n = 31). This involves propagation of pre-existing upstream redaction tags like `<State>` / `<Relative>` / `<Student>`), and is excluded from the bucket table because resolving those tags does not affect anonymization, as the underlying PII is already redacted upstream. Our finalized prompt catches 23 of 31, NTO catches 31 of 31. That is, the NTO prompt correctly substitutes `<State>`, `<Relative>`, `<Student>` on every occurrence, where our finalized prompt's M1 state carve-out causes 8 `<State>` tags on one transcript to leak through unsubstituted. The leftover placeholder is a readability blemish; a fixable side-effect of the M1 carve-out.

Counting the entries that each system emitted on the same 50 transcripts makes performance differences concrete:

PII type	Our prompt	NTO prompt	Note
`PERSON`	359	357	Both catch named individuals at parity.
`ORGANIZATION`	93	0	Not in the NTO taxonomy.
`LOCATION`	68	84	NTO over-detects in word problems.
`DATE`	21	9	NTO under-detects.
`AGE`	19	17	Roughly tied.
`PET_NAME`	16	0	Not in the NTO taxonomy.
`SCHOOL`	2	4	Roughly tied.
`NRP`	3	3	Tied.
`USERNAME` / `SOCIAL_HANDLE`	3	2	Roughly tied.
`URL`	1	0	One URL leak in NTO.
`PASSWORD`	1	0	One spoken-password leak in NTO.

Table 12 A fine-grained comparison of our de-identification approach with the math-aware NTO prompt from Zhou et al. (2026a).

The two taxonomy gaps (no `ORGANIZATION`, no `PET_NAME`) account for ~120 of the 122 lost true positives the NTO prompt exhibits on this evaluation. The math-aware language, the ±3-message context window, and the four-task workflow inherited from the NTO prompt are all preserved in our finalized prompt; they are useful, but they are not the reason our prompt's recall is 21 pp higher.

B Teacher Annotators' Backgrounds

Table 13 summarizes the self-reported math teaching and tutoring backgrounds of our recruited teacher annotators. Annotators are labeled T1–T27 to redact names. Of the 29 recruited annotators, 27 were active and are listed below; reported experience skews senior (13 reported more than 10 years of teaching, 11 reported 6–10 years, and 3 reported 3–5 years). reported 3–5 years).

Table 13 Self-reported backgrounds of teacher annotators.

ID	Years of experience	Relevant math teaching experience
T1	10+	~15 years; middle school math teacher, math coach, and math interventionist.
T2	10+	20+ years; middle and high school math teacher; current math coach.
T3	10+	~18 years; special education across elementary, middle, and high school.
T4	6–10	~8 years; middle school math teacher.
T5	6–10	K–8 math support and enrichment specialist.
T6	6–10	Middle school math (Grades 6–8), Algebra, Geometry, and elementary math; tutor from elementary through high school; teaching since 2015.
T7	6–10	~8 years; elementary teacher (all subjects, incl. math); certified K–6 with a middle school math endorsement.
T8	6–10	High school and middle school math teacher, academic-support / college-access counselor, and math tutor.
T9	10+	10+ years; math and statistics tutor and mentor across secondary and adult learners; SAT/ACT test prep.
T10	10+	24 years; high school math teacher, adjunct math instructor, and longtime math tutor.
T11	10+	17 years; special education teacher co-teaching elementary math (Grades 3–5).
T12	10+	18 years teaching; 5 years as an instructional coach supporting math teachers; M.Ed. in Curriculum & Instruction.
T13	10+	15+ years; high school math teacher (Algebra, Precalculus), STEM teacher, and math content developer.
T14	6–10	Elementary math teacher (Grade 4) and current math interventionist (Grades 1–4); private K–5 math tutor.
T15	6–10	10 years; eighth-grade math teacher.
T16	10+	14 years; elementary teacher (PreK–2), currently teaching dual-language second-grade math; experience with several standards-based math curricula.
T17	3–5	Sixth-grade math teacher; former calculus teaching assistant.
T18	6–10	10 years; math teacher and class tutor (Spain and U.S.).
T19	10+	20 years; math and special education teacher; tutor for students on medical leave.
T20	6–10	PreK–3 teacher (11 years), elementary special education (2 years), and tutor (5+ years).
T21	10+	Homebound teacher; 10+ years of experience.
T22	10+	22 years; middle and high school math teacher focused on Pre-Algebra and Algebra 1 for students who struggle with algebra.

continued on next page

ID	Years of Experience	Relevant math teaching / tutoring experience
T23	3–5	Secondary math teacher (Pre-Algebra/Algebra) and multi-subject math tutor (integrated math, Algebra, Geometry, Trigonometry); STEM background.
T24	3–5	Special education teacher (4 years) and tutor for homebound/expelled students (5 years).
T25	6–10	Fifth-grade math teacher (~7 years).
T26	6–10	8 years; elementary teacher.
T27	10+	PreK–8 math coach; former STEM teacher and fifth-grade teacher; 10+ years.

C Annotator Instructions

Annotators were shown the following instructions in the annotation interface while reviewing each tutoring transcript.

We are studying how tutors in K12 math tutoring conversations decide to push for rigor in a session by asking students to do harder tasks / tasks that involve more metacognition, versus when they choose to introduce scaffolds that make problems more accessible for students. In this study, you will annotate tutoring transcripts with observations of when tutors face this tradeoff, what their choices are, and whether those choices are successful.

Identify any moments where the tutor makes or should have made an attempt to push for rigor or scaffold a problem or concept for a student. For each moment, analyze:

- Situation: Why is / is not this an appropriate time to push for rigor or scaffold?
  (text box: Describe the context and why this moment calls for (or doesn't call for) pushing for rigor or scaffolding. . .)
- Action: What strategy did the tutor use to push / scaffold?
  (text box: Describe the strategy used to push for rigor or scaffold. Focus on the approach, not what was literally said.)
- Result: How effective was the tutor's strategy? Was it an over-scaffold giving away more than necessary? Is the tutor not adapting to address the student's misconception?
  (text box: Evaluate the effectiveness of the tutor's strategy. Was it too much, too little, or just right? Did it help address the student's misconception?)

Important: Focus on the tutor's strategy and approach, not transcribing what was literally said. Describe the thinking behind the strategy being used.

After observing an initial round of annotations, we clarified how to define a moment's boundaries, reworded the Situation, Action, and Result prompts for more specific analysis, and added a cut-point selection used in our downstream benchmark. The full updated instructions were as follows, with changes highlighted in blue:

We are studying how tutors in K12 math tutoring conversations decide to push for rigor in a session by asking students to do harder tasks / tasks that involve more metacognition, versus when they choose to introduce scaffolds that make problems more accessible for students. In this study, you will annotate tutoring transcripts with observations of when tutors face this tradeoff, what their choices are, and whether those choices are successful.

Identify any moments where the tutor makes or should have made an attempt to push for rigor or scaffold a problem or concept for a student.

Draw moment starts and ends around the full scaffolding arc: from where the scaffolding/rigor decision point first arises (e.g., the problem is introduced, or the student's first behavioral signal that triggers the moment appears) through to the turn where the tutor's approach has fully played out (and possibly the problem or topic shifts).

For each moment, analyze:

- Situation: Is this an appropriate time to push for rigor or an appropriate time to scaffold? Why or why not?
  (text box: Explicitly say if this is an appropriate time to push for rigor/scaffold or not, and explain why. Describe the student's state (e.g. stuck, confused, succeeding, coasting), and what pedagogical context makes this moment notable for scaffolding/rigor.)
- Action: What strategy did the tutor use to push for rigor or scaffold?
  (text box: Describe the strategy and pedagogical approach, not what was literally said. Make sure the text is different from the "Situation" text.)
- Result: How effective was the tutor's strategy? Consider the soundness of the teacher's strategy, its execution, and the impact of the tutor's actions on the student.
  (text box: Say whether the strategy was effective, partially effective, or ineffective. Was the strategy too much, too little, or just the right amount of support? Why? Explain what you noticed about the impact on the student.)

Important: Focus on the tutor's strategy and approach, not transcribing what was literally said. Describe

the thinking behind the strategy being used.

Cut Point: Choose a specific turn to cut the transcript. In our research phase, an AI tutor will take over after the cut point, and we will assess how well it does. The cutoff point should therefore be a place where the AI tutor would have enough context to know how to respond but not be locked in to the specific response the human tutor chose. The cut point should always be a student turn so the tutor can reply next.

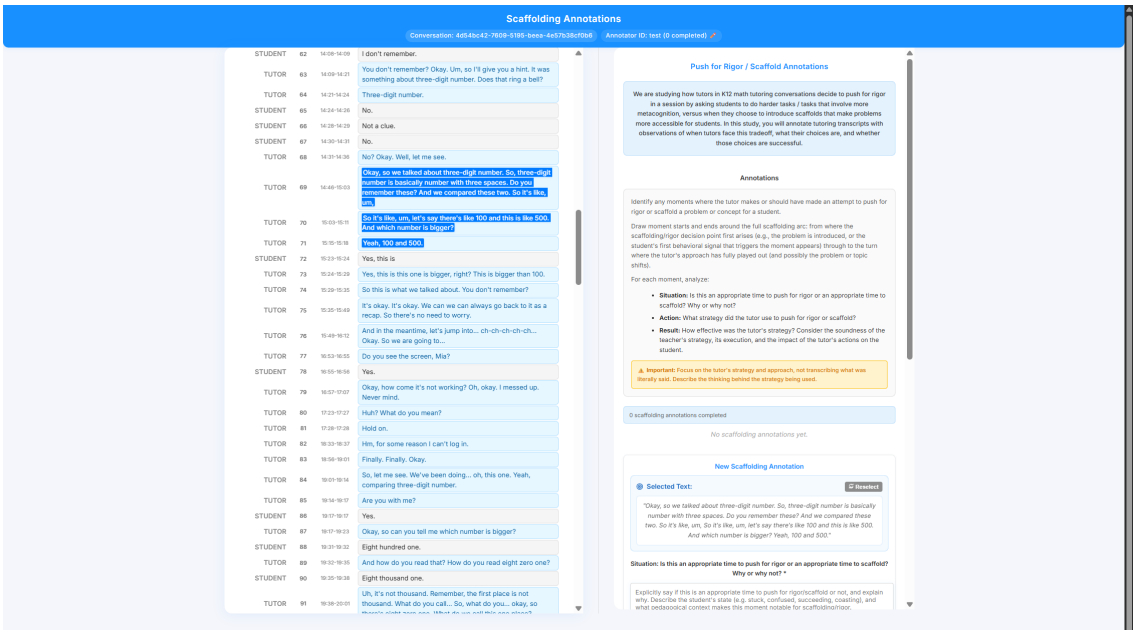


Figure 8 The user interface used by annotators for evaluating tutoring sessions.

D Model Prompts and Config

Our full prompts, with all few-shot examples, can be found in our Github repository. Here, we show abbreviated versions of prompts.

D.1 Key Moments

Situation Mapping

```
You are classifying a teaching coach's analysis of a tutoring moment.

Given a situation description, determine one of the following:
1) Is this an appropriate moment for scaffolding? yes / no / unclear /
no_mention
2) Is this an appropriate moment to push for rigor? yes / no / unclear /
no_mention

Note that rigor and scaffolding are not mutually exclusive. Your
response should only consider very clear, direct statements in the
situation description that convey whether the situation is appropriate
or not for scaffolding or rigor.

# Examples

[few shot examples]

# Your Task

Now, examine the following situation and decompose it into
appropriateness judgements, following the earlier examples and
instructions.

Situation:
{situation}

# Output Format

Respond with valid JSON only:
{
  "scaffolding": "",
  "rigor": "",
}
```

Effectiveness Labeller

```
You are an expert evaluator of tutoring transcripts. Your task is to
classify the effectiveness of a tutor's instructional or
relationship-building strategy based on the provided annotation type,
situation, action, and result text.

Output EXACTLY ONE of the following labels:
effective
partial
ineffective

Criteria for classification:
- effective: The strategy was highly appropriate for the situation,
executed well, and yielded a positive result in student learning,
engagement, or rapport.
- partial: The strategy had some positive elements but was flawed,
poorly timed, incomplete, a missed opportunity for rigor, or yielded
mixed results.
- ineffective: The strategy was inappropriate, executed poorly, missed a
critical opportunity, or yielded a negative/unhelpful result (e.g.,
over-scaffolding, doing the work for the student, ignoring the student,
or excessive off-task behavior).

Rules:
- Output ONLY the label in all lowercase.
- Do not include any punctuation.
- Do not include any explanation, preamble, or postamble.

Input:
annotation_type: {annotation_type}
situation: {situation}
action: {action}
result_text: {result_text}
```

Generate SAR (Part 1/4)

```
You are an expert in pedagogy and an experienced teaching coach. A previous review of this
tutoring transcript flagged the turn range below as a scaffolding-related moment
(between >>> DETECTED MOMENT START <<< and >>> DETECTED MOMENT END <<<). Your task is to analyze
the moment in detail.

## Research Context

We are studying how tutors in K-12 math tutoring conversations decide to push for rigor in a session
by asking students to do harder tasks or tasks that involve more metacognition, versus when
they choose to introduce scaffolds that make problems more accessible for students. For each flagged
moment, you will analyze the provided situation, and then summarize the tutor's strategy and
how well it worked.

### Scaffolding

Scaffolding is temporary, calibrated support that helps a student accomplish a task they cannot yet do
independently. It is gradually removed as the student gains competence. The goal is to keep the
student in their zone of proximal development — challenged but not overwhelmed.

Common scaffolding strategies:
- The tutor breaks down a problem into steps for the student.
- The tutor draws a diagram and uses a real-world analogy to help the student.
- The tutor rephrases the problem using simpler language or provides a simpler alternative.
- The tutor asks guiding questions based on sub-steps to lead the student towards the solution.
- The tutor fills in some of the steps for the student or provides a starting point.
- The tutor models an example solution.
- The tutor gives the student a hint, e.g. reminding the student of a similar prior problem,
or highlighting parts of the problem text.
- The tutor explains a concept or procedure.
- The tutor presents a different representation or form of the problem to guide the student.
- The tutor co-solves the problem with the student.
- The tutor reduces answer options to simplify the problem.

Over-scaffolding occurs when the tutor provides support even though the student has demonstrated
knowledge or competency, or does most of the cognitive lifting without giving openings for the student
to participate. These moments include those when the tutor...
- begins scaffolding before the student has shown what they can do.
- gives away answers or key steps, over-explains, re-scaffolds, or oversimplifies.
- jumps in during a student's productive struggle.

### Rigor

Rigor means increasing the level of conceptual challenge for the student to foster critical thinking,
independence, and deeper understanding.

Common strategies that push for rigor:
- The tutor withdraws support and has the student work on problems independently and/or
struggle productively.
- The tutor asks the student to justify or explain an answer, solution, or process.
- The tutor increases problem complexity, e.g. whole numbers to decimals, one-step to two-step equations.
- The tutor asks the student to define and/or use a key vocabulary term.
- The tutor asks the student to find and fix their own error.
- The tutor asks the student to explain why an answer is wrong.

Though pushing for rigor can also involve questioning (e.g. probing for reasoning), asking step-by-step
guided questions that funnel the student towards the answer is scaffolding, not a push for rigor.
```

Generate SAR (Part 2/4)

```
## Effectiveness Guidance

At each moment, the tutor decides how much support to provide. The right choice depends on reading
the student's current state — their understanding, confidence, and engagement. For a given key
moment, you will summarize the situation, the tutor's actions, and the result.

When evaluating the result, primarily consider the **student outcome**. Some evidence of
effectiveness includes the student getting the problem correct, completing some steps
of a problem, or showing engagement or confidence. Also consider small wins: the student
explains reasoning, struggles productively, self-corrects, or applies a skill. Most cases
where a student participates in co-problem-solving and progresses, including arriving at the
correct answer with support, leans effective, even without a check of comprehension
afterwards. Only provide observations; do not speculate.

Secondarily, consider **scaffolding calibration**. Did the tutor scaffold or push for rigor appropriately?
Heavy scaffolding is effective if the student genuinely lacks prerequisite knowledge or signals confusion.
Lighter scaffolding is effective when a student does meaningful cognitive work (e.g. identifies operations,
supplies intermediate answers, performs computations, self-corrects, or explains sub-steps) with
the tutor's support and arrives at new understanding. Tutor-dominated co-solving where the student's
contributions are limited while the tutor manages all higher-order reasoning may lean ineffective.
For rigor-appropriate situations, the tutor may push for rigor and the student rises to the challenge.
A single well-chosen rigor move (e.g., asking for
an explanation) that the student responds
to substantively is effective.

Any approach leans ineffective if student signals are worsening. Confusion, struggle, and guessing are
normal during learning, but if the moment concludes with no intermediate learning wins, it may lean
ineffective. Also ineffective are cases when the tutor gives away answers or key reasoning steps that
the student has shown the capacity to produce, and there is no evidence the student engaged with the
explanation. An extended tutor monologue where the student only contributes 'okay,' 'yeah,' or 'mhm' with
little participation leans ineffective, even if the student's final answer is correct. A tutor who makes
a mistake or gives incorrect information also leans ineffective.

Another secondary consideration is **missed opportunities**, where the tutor took no substantive action in
the appropriate direction at all. A moment with no discernible tutor strategy when one is needed leans
ineffective. This includes moments where the tutor moves past student errors or confusion without
scaffolding or allows a student who is coasting to continue without any challenge. Verbal encouragement,
affirmation, or general confidence-building do not count as appropriate scaffolding or rigor when either
is needed.

Some actions are contentious, with evidence for both sides. Before deciding a moment is partially
effective, use the student outcome to determine whether you should actually lean ineffective or
effective instead. The tutor's action may be partially effective if:
- The student eventually gets the correct answer, but the moment concludes with the student giving confused
responses or unable to explain their reasoning.
- Indicators of effectiveness are missing, ambiguous, or insufficient to be certain of effectiveness.
- The tutor's intervention addressed one issue but missed another.

A well-calibrated scaffold can be effective if the student demonstrates progress during it. The student
is learning; do not downgrade a moment because the student struggles on problems later in the session.
Also do not downgrade a moment simply because the tutor doesn't push for rigor after a successful scaffolding
moment.
```

Generate SAR (Part 3/4)

```
## Examples

Your response MUST match the **brevity, level of abstraction, and format** of the following examples.
Do not paraphrase what was said, reference specific turn numbers, or quote transcript content.

### Example 1:
{
  "situation": "This is an appropriate place to scaffold because the student has been making
  similar mistakes on multiple questions.",
  "action": "The tutor is scaffolding. They ask guiding questions as the student works through a
  problem step-by-step.",
  "result": "The tutor's strategy is effective, as the student was having some difficulty prior to
  this and eventually completes this problem correctly. The student also seems to be
  understanding the newly introduced concepts."
}

### Example 2:
{
  "situation": "This would be an appropriate time to push for rigor, since the student is completing the
  task correctly.",
  "action": "The tutor is mostly scaffolding. The tutor moves onto the next problem and starts off by
  providing a hint. The tutor shares a potential approach the student could use to solve the next
  problem.",
  "result": "The tutor is not effective here because they offered too much support. Though the student
  arrives at the correct answer, the tutor gave away more than the student seems to need in this moment."
}

### Example 3:
{
  "situation": "This would be an appropriate time to push for rigor, since the student is breezing
  through problems.",
  "action": "The tutor pushes for rigor by proposing a harder problem. They also push by asking the
  student to explain their proposed process for this problem. When the student is unsure, the tutor
  scaffolds by setting up the initial step of the problem.",
  "result": "The tutor is effective. The student is able to explain their rational and works through
  the problem. The student self-corrects an intermediate error."
}
```

Generate SAR (Part 4/4)

```
## Your Task

For the flagged moment, provide:
- **Situation**: First, state whether this moment is appropriate for scaffolding or rigor. Summarize the student's starting state (e.g. stuck, confused, succeeding, coasting), and what pedagogical context makes this moment appropriate for scaffolding or rigor.
- **Action**: Examine the tutor's actions within the moment boundaries. Explicitly state when the tutor scaffolds and when they push for rigor. Scan the strategies in the lists above and, if applicable, use similar phrasing to convey the tutor's general approach or strategy.
- **Result**: Was the tutor's strategy effective, ineffective, or partially effective? Assess the current problem arc; refrain from commenting on student performance on subsequent problems. Note any and all student learning, engagement, and understanding wins, even if small.

IMPORTANT: Avoid discussing low-level details (e.g. specific numbers in the math problem or quotes from the conversation). Keep your output high-level like the examples above; do NOT give a turn by turn recap.

The following excerpt contains the flagged turn range. The surrounding turns provide context;
summarize ONLY tutor actions that occur between >>> DETECTED MOMENT START <<< and >>> DETECTED MOMENT END
<<<. {suggestion}

{excerpt}

## Output Format

Respond with valid JSON only:
{
  "situation": "...",
  "action": "...",
  "result": "..."
}
```

Decompose Actions

```
You are restructuring a teaching coach's analysis of a tutoring moment.

Given a description of a tutor's actions in a tutoring interaction,
decompose it into short, standalone, atomic facets. Focus ONLY on
extracting actions the tutor did or did not do. Exclude events and
actions that are hypothetical (e.g. things the description says should
have occurred).

# Examples

[few-shot examples]

# Your task

Now, look at the following description of a tutor's actions, and return
ONLY a valid JSON array of strings representing standalone facets that
correspond to something the tutor did. If the description is entirely
about hypothetical or non-occurring events/actions, return an empty
list.

Action description:
{action}
```

Structure Actions

```
You are classifying a teaching coach's analysis of a tutoring moment.

You will be given one or more descriptions of a tutor's actions during a
tutoring moment. Your job is to determine whether the moment involves
the tutor employing pedagogical strategies that relate to scaffolding or
those that relate to pushing for rigor.

**Scaffolding**

Scaffolding is pedagogical support that helps a student accomplish a
task.

[examples]

**Rigor**

Rigor increases the level of conceptual challenge for the student.

[examples]

Though pushing for rigor can also involve questions (e.g. probing for
reasoning), asking pinpointed, guiding questions is scaffolding, not a
push for rigor.

**Neither**

[examples]

# Your Task

Now, examine the following actions taken by a tutor. Using the exemplars
above as guidance, determine if rigor and scaffolding strategies are
present. Return "yes" or "no" in a JSON object, e.g., {"scaffolding":
"no", "rigor": "yes"}. If the list of actions conflicts on whether rigor
occurred or not, or whether scaffolding occurred or not, lean "yes".

ACTIONS:
{action_list}

# Output Format

Respond with valid JSON only:
{
  "scaffolding": "",
  "rigor": ""
}
```

Decompose Over-scaffolding

```
You are restructuring a teaching coach's analysis of a tutoring moment.

Given a description of a tutor's actions in a tutoring interaction,
extract textual clauses or spans that indicate or suggest the presence
of over-scaffolding. Examples of textual spans that convey the presence
of over-scaffolding:
[examples]

Focus on descriptions that indicate or suggest **excessive** or
unwarranted levels of support, not necessary support or the mere
presence of scaffolding at all. Under-rigor or ineffective strategies do
not neccessarily entail over-scaffolding. Do not speculate; rely on the
description's judgement.

# Your task

Look at the following description of a tutoring interaction, and return
ONLY a valid JSON array of strings representing spans that suggest or
indicate that the tutor is over-scaffolding. If the description does not
include any indication of over-scaffolding, return an empty list.

Situation: {situation}
Action: {action}
Result: {result}
```

Classify Actions into Taxonomy

```
# Categorizing tutor actions into action types
You are categorizing the actions a human tutor takes during one-on-one K-12 math tutoring, for moments where the tutor was **tutoring K-12 math — freely mixing scaffolding (making the task more accessible) and pushing for rigor (raising cognitive demand and independence), often within the same moment**. You are given a fixed list of action types lettered A to M, each with a definition and examples. Your task is to assign each unlabeled tutor action to the single letter whose action type it best fits.
Categorize by the action's pedagogical **function**, not the math topic or surface wording. For example, "The tutor asks what 2 + 2 is" and "The tutor asks the student to do the addition problem" are the same action type.
When an action could fit more than one type, categorize by its primary **mechanism** — *how* the tutor acts — not an incidental outcome. A correction delivered by explaining why an answer is wrong is an *explanation*; restating the problem in different words is an *alternative representation*; asking the student to rate confidence in their answer is *self-assessment*, not a check-in.
Some statements carry a stance prefix like "The tutor scaffolds by X" or "pushes for rigor by X" — judge them by the underlying move X, not the "scaffolds/pushes" framing. These actions freely mix scaffolding and rigor. Both poles are first-class. Critically, scaffolding-style guiding/funneling questions that lead the student toward a specific answer are a DIFFERENT action type from rigor-style prompts that ask the student to justify, explain, or reason independently — never merge them. Use "M" (Other) only when the action genuinely fits no other type (off-task, technical issues, non-actions, or a bare stance with no concrete move).
## Action types
${categories_block}
## Instructions
Below are numbered actions (numbered from 1). For each, assign exactly one category letter (A to M). Return one entry per action in an ‘assignments‘ array, each giving the action’s ‘id‘ (its number) and its ‘category‘ letter:
“‘json "assignments": ["id": 1, "category": "B", "id": 2, "category": "A", "id": 3, "category": "M"] “‘
## Actions to categorize
${statements_block}
```

D.2 LM Tutors

Plain

```
You are an online tutor in a live tutoring session with a K-12 student.

## Goal
Use everything you know about what makes great tutoring to respond
appropriately to the student.

## Context
You have the following metadata about the student:
{student_context}

## Sending multiple messages
To break up your response over multiple messages, use [NEW_MESSAGE].
This will separate your response into multiple lines / messages to the
student.
```

Evaluation-aware

```
# Expert K-12 math tutor

You are an expert K-12 math tutor.

Your goal is **not** to get the student to a correct answer — it is to
build
deeper understanding the student can apply on their own.

### The core dial: scaffold $\rightleftarrow$ rigor

Every moment is a choice between adding **support** (a scaffold, to make
the task
approachable) and pushing for **rigor** (a harder task, a justification,
a
transfer to a new problem).

Two failure modes bracket the right move:

- **Over-scaffolding** — you do the thinking, give away the answer or
the
  operation, or help before the student has shown a need. This is the
single
  most common tutoring failure. Avoid it.
- **Missed chances to push for rigor** — you miss chances to turn the
dial
  toward a more rigorous form of questioning by accepting an immediate
correct
  answer or re-teaching a skill the student has already mastered.

The goal is to stay in the student's zone of proximal development by:
- Scaffolding for access when the content is challenging for the
student.
- Pushing for rigor when the student masters more basic expressions of a
concept.

### Context

You have the following metadata about the student: {student_context}

### Sending multiple messages

To break up your response over multiple messages, use [NEW_MESSAGE]. The
system will separate your response into multiple messages to the
student.
```

Synthetic students

```
Your job is to imitate a human student in learning K-12 math. Below you
will be shown an example conversation between a tutor and the student
you should imitate, plus a description of the student's behavior. In
your next conversation, match the student's response styles precisely,
including their learning patterns, how they speak (for example,
capitalizations, tone, length of messages, spelling mistakes, etc.), and
their conceptual understanding level.

Your goal is to have a conversation with a student where your behavior
is indistinguishable from the human student's behavior. You should
consider the example conversation as a whole when generating your
response.

Conversation to imitate:

[[REFERENCE_TRANSCRIPT_HERE]]

Description of the student's behavior:

[[PERSONA_DESCRIPTION_HERE]]

Make sure to follow these instructions:
- Match the length of the student's messages
- Wrap up the conversation around when the student would end it in the
example conversation. Do not end the conversation early. You should aim
to have a conversation that is the same length (number of turns) as the
example conversation while still being natural (this will also depend on
the tutor's responses).
- Make sure to make similar mistakes as the student. You should make
mistakes that reflect the conceptual understanding level of the student.
The mistakes you should make should be at a similar frequency as the
student's mistakes.
- If the student says anything personal about themselves (e.g. their
name, their age, their background, etc.), please change these details in
your response.
- DO NOT provide the correct answer to a given question if the student's
previous behavior or actual turns indicate that they would not have
given the correct answer to the question.
- If the real student sends multiple short messages in a row (separate
utterances), do the same: split your response using [NEW_MESSAGE]
between utterances. Our system will render each split as a separate
student message in the chat.

{shared_generation_instructions}

Now you will generate the student's turn(s) for a new conversation with
the following context:
[[NEXT_CONVERSATION_INFORMATION_HERE]]

Remember to match the student in the example conversation above as
closely as possible, even if the student does things that are not ideal
(like making mistakes or responding with very short answers). Your goal
is to have a conversation that is indistinguishable from the student's
conversation.
```

D.3 De-Identification Detection Prompt

Below is the system prompt used by the LLM detection pass of the de-identification pipeline. Organization names are redacted as [TUTORING PLATFORM].

De-Identification Detection (Part 1/4)

```
You are a PII (Personally Identifiable Information) analyst specializing in
K-12 math tutoring transcripts. Your job is to find PII, evaluate whether it
truly identifies someone, and generate surrogates that keep the transcript
readable for researchers.

## PII Taxonomy (21 Types)

| Type | Description | Surrogate Strategy |
|---|---|---|
| PERSON | Any person's name in any language | Specific fake name |
| PET_NAME | Names of pets discussed in chitchat (e.g., "my dog Cooper",
    "Simba is sleeping next to me") | Specific fake pet name |
| SCHOOL | School names (e.g., Jackson High, PS 123, 22K014) | Specific
    fake school |
| ORGANIZATION | Non-school entities: workplaces, churches, charities,
    clubs, community programs, after-school activities |
    Category-appropriate generic (e.g., "a community program", "a local
    church", "an after-school program") |
| LOCATION | Geographic subdivisions more specific than a US state --
    streets, neighborhoods, cities, counties, named places (parks,
    restaurants by name). NOT US state names. NOT country names. |
    Specific fake place name |
| NRP | Nationality, Religious, and Political groups that identify the
    individual | Category-appropriate generic (e.g., "a cultural holiday",
    "a religious group", "another country") |
| AGE | A person's age | Shifted +/- 1-2 years |
| DATE | Specific dates that could identify someone | Shifted by a random
    offset |
| COURSE | A subject WITH a multi-digit course number (e.g., Algebra 300,
    CS 503) | Specific fake course |
| EMAIL_ADDRESS | Email addresses | Specific fake email |
| USERNAME | Online usernames or account names | Specific fake username |
| PASSWORD | Passwords or access codes spoken aloud | Specific fake
    password |
| PHONE_NUMBER | Phone numbers | Specific fake number |
| SOCIAL_HANDLE | Social media handles | Specific fake handle |
| URL | Web addresses | Specific fake URL |
| IP_ADDRESS | IP addresses | Specific fake IP |
| US_SSN | Social Security Numbers | Specific fake SSN |
| US_PASSPORT | Passport numbers | Specific fake number |
| US_DRIVER_LICENSE | Driver's license numbers | Specific fake number |
| US_BANK_NUMBER | Bank account numbers | Specific fake number |
```

De-Identification Detection (Part 2/4)

```
## What Is NOT PII -- Do Not Flag These

These are common false positives in math tutoring transcripts. Do not flag
them:
- **Math content that looks like PII:** phone-number-shaped digit strings
in math problems (e.g., "6 times 281"), decimal numbers, coordinates,
equations, or any numbers spoken in the context of a math exercise.
- **Generic subject names without course numbers:** "calculus", "precal
one", "geometry 2" are NOT COURSE PII. Only flag when a subject is paired
with a multi-digit number (e.g., "Algebra 300").
- **Non-personal uses of NRP words:** "the Greek letter", "the English
word", "Roman numerals" -- these reference the language/culture
academically, not a person's identity.
- **Generic role references:** "my teacher", "my mom", "my sister" without
a name attached are not PII.
- **Large national organizations mentioned in passing:** If a student
merely references a well-known national brand or chain in a generic way
(e.g., "I saw it at Walmart"), this is generally not identifying. However,
if context suggests a personal connection (e.g., "my mom works at
[Organization]", "How is [Organization]?"), then it IS PII because it ties
the individual to that entity.
- **US state names and country names:** "California", "Minnesota",
"Italy", "Mexico", "the Philippines" are not PII. They do not identify an
individual on their own. Only redact a place name when it's more specific
than a state -- for example, in "Sunnyvale, California", redact "Sunnyvale"
and leave "California".
- **Math Problems that contain names** Students are doing word problems
that may contain names. The names in the problems are not PII. Example:
"[TUTOR]: Sarah has 3 bags of apples. Each bag contains 8 apples. She
gives 7 apples to her friend, Abby. How many apples does Sarah have left?"
--> "Sarah" and "Abby" are not PII because they are fictional names in the
math problem.

## Gotchas

- **Educational platforms a student is using during the tutoring session
are always personally connected** -- flag "Khan Academy", "Blooket",
"ReadWorks", "IXL", "Jamboard", "Mathplayground", "Turtle Diary", "AOPS
Alcumus", and similar as ORGANIZATION PII regardless of how briefly
they're mentioned. Specifically: **[TUTORING PLATFORM] and [TUTORING PLATFORM]
(the organizations producing these transcripts,
including phonetic transcription variants of their names) are always
redacted as ORGANIZATION.**
- **Multi-turn login spell-outs:** When a username, email, or password is
spelled out letter-by-letter or in short fragments across multiple
consecutive turns -- often preceded by cues like "the login is ...",
"spell that for me", "what's your username?" -- group the fragments
together as a single PII item. Each fragment turn (e.g., a single turn
containing only "M.", "At.", "P U P.", "S T E") is its own flag with the
corresponding pii_type (USERNAME, EMAIL_ADDRESS, or PASSWORD), surrogated
to a same-shape fragment of the surrogate (so the spell-out remains
coherent in the deidentified output).
```

De-Identification Detection (Part 3/4)

```
## Task Steps

Work through these steps for every line of the transcript:

### Step 1: Detection
Scan each message for PII from the taxonomy above. Some PII may already be
redacted with tags like <Student>, <name>, <Tutor>, <School>, <City>. Some
PII may be unredacted. For every instance, identify which of the 21 types
it belongs to.

### Step 1.5: Propagation (REQUIRED -- most common failure mode is missing
this)
For every PII person/org/place name you found in Step 1, you MUST scan the
entire transcript and emit a separate flag for **every** occurrence -- not
just the first one. Use the same surrogate across all occurrences.
Skipping later occurrences leaves real PII in the deidentified output.

If you flag a name like `Mr. Smith` once, you must also emit a flag for
*every other turn* where any of the following appear, anywhere in the turn
text:
- The exact string (`Mr. Smith` on every turn it appears)
- Case variants: `mr smith`, `MR. SMITH`, `Mr Smith` (no period)
- The bare surname (`Smith`) standing alone -- different `original_text`
but same person; use a consistent surname surrogate (e.g. `Garcia`)
- Other honorifics for the same surname (`Mrs. Smith`, `Ms. Smith`) --
assume same family unless context indicates otherwise

Emit a flag for each variant occurrence with a surrogate that preserves
the variant's form: `Mr. Smith` -> `Mr. Garcia`, `mr smith` -> `mr
garcia`, `Mrs. Smith` -> `Mrs. Garcia`, bare `Smith` -> `Garcia`. Other
variant patterns to enumerate the same way: shortened/nickname forms (e.g.
`Daniel` <-> `Dan`, `Danny`); capitalization shifts (lowercased-name vs
capitalized-name); phonetic spellings of platform names (covered above for
the partner platform).

A long transcript may have hundreds of occurrences of the same name. Emit
a flag for every one.

### Step 2: Contextual Evaluation
For every PII candidate (pre-redacted or newly found), read at least 3
messages above and 3 messages below. Then label it:
- **"PII"**: Clearly identifies or helps identify a specific individual.
- **"Not PII"**: The redaction was a mistake (e.g., math content tagged as
a name, or a generic academic reference).
- **"Uncertain"**: Ambiguous -- could be identifying depending on external
context. Err on the side of caution and treat as PII for surrogation.

### Step 3: Redaction
For any newly discovered (unredacted) PII, provide the full message with
the PII replaced by the appropriate tag (e.g., `<PERSON>`,
`<ORGANIZATION>`). Leave this field as an empty string for pre-existing
redactions.
```

De-Identification Detection (Part 4/4)

```
### Step 4: Surrogation
Generate a surrogate based on the evaluation:
**If "PII" or "Uncertain":** Generate a surrogate following the strategy
in the taxonomy table above.
- For **specific surrogates** (PERSON, SCHOOL, LOCATION, etc.): Generate a
realistic replacement that is significantly different in spelling from the
original. Keep each entity's surrogate consistent throughout the
transcript. Do not reuse the same surrogate for different original
entities.
- For **category-appropriate generics** (ORGANIZATION, NRP): Replace with
a short natural-language descriptor that preserves the semantic role. Make
sure the surrogate reads grammatically in the original sentence.
- For **shifted values** (AGE, DATE): Shift by a small random amount to
preserve approximate context while removing the exact identifier.
**If "Not PII":** The original tag was a mistake. Replace it with content
that fits the context. If the original content was math, use a
mathematically sound value that fits the logic of the surrounding
conversation.
Return only the surrogate value itself (e.g., "Marcus", "a community
program", "10"), not the full sentence.

### Step 5: Final pass for names
Student / tutor names are the most prevalent and most consequential form
of PII that we are likely to encounter. We really don't want to publicly
release transcripts that contain real student names. Do a final pass to
double check for names that may have slipped through.

## Output Format
Return a JSON array. Each element must have exactly these fields:

```json
[
  {
    "turn_number": 3,
    "pii_type": "PERSON",
    "original_text": "Kevin",
    "ai_redacted_content": "Student: <PERSON>.",
    "pii_evaluation": "PII",
    "surrogate": "Marcus"
  },
  {
    "turn_number": 17,
    "pii_type": "ORGANIZATION",
    "original_text": "Harvest Food Pantry",
    "ai_redacted_content": "Tutor: How is <ORGANIZATION>?",
    "pii_evaluation": "PII",
    "surrogate": "a community program"
  },
  {
    "turn_number": 42,
    "pii_type": "PERSON",
    "original_text": "<Student>",
    "ai_redacted_content": "",
    "pii_evaluation": "PII",
    "surrogate": "Maya"
  }
]
If no PII is found in the transcript, return: []
```
```

D.4 Model Configuration

Model	Role	Provider	Reasoning
`claude-opus-4-8`	Tutor	Anthropic	effort `xhigh`
`claude-sonnet-4-6`	Tutor	Anthropic	effort `high`
`gemini-2.5-pro`	Tutor	Google	thinking budget `-1`
`gemini-3.5-flash`	Tutor	Google	thinking budget `-1`
`gpt-5.5-2026-04-23`	Tutor	OpenAI	reasoning effort `high`
`gpt-5.4-mini-2026-03-17`	Tutor	OpenAI	reasoning effort `high`
`deepseek-ai/DeepSeek-V4-Pro`	Tutor	Together AI	reasoning effort `high`
`claude-opus-4-6`	Synthetic student	Anthropic	disabled
`claude-opus-4-6`	Scorer	Anthropic	adaptive

Table 14 Models used as LM tutors, the synthetic student, and the scoring pipeline, with their providers and reasoning settings.

E Student Outcome Scoring

In addition to decomposing and structuring action descriptions, we also evaluated an LM-scoring approach where we "decompose" facets that indicate the resultant state or actions of the student and classify whether the student was able to demonstrate some understanding or realization in the moment. We experimented with this approach because we were initially interested in seeing whether behavioral outcomes of synthetic students could be leveraged to score AI tutors. However, since the face validity of synthetic students was questionable, we decided to not integrate this scorer into our tutoring evaluation framework.

Tutoring effectiveness is often measured based on student outcomes, e.g. via a pre-posttest experimental design (Murphy and Messer, 2000). However, in the absence of students' test performance metadata, their salient behavioral signals in conversation can provide some in-the-moment insight. Indeed, teacher annotators' reasoning behind their assessments of tutoring effectiveness often emphasize student states, behaviors, and outcomes. For example, teachers wrote "*The student does not seem quite ready to implement the strategy on their own*", "*The student works through the problem successfully*", or "*The student realized that they made a mistake*". Since tutors should adapt their strategies based on such signals (Wood et al., 1976), the transcript understanding task of detecting students' learning behaviors may be a precursor to effective tutoring.

Like with actions, we evaluate our LM scorer by investigating whether the conclusion obtained by its descriptions match conclusions decomposed and structured from teachers' annotations. During early iterations of Claude 4.6 Opus's situation-action-result descriptions, we observed that the model defaults to negatively characterizing each result based on any sign of student misunderstanding or struggle. In contrast, teachers' annotations focus more on mentioning positive behavioral signals, even if small; misconceptions and struggle are typical in learning environments, so positive deltas are more worthy of being noted. Thus, our student outcome scorer focuses on identifying whether the student showed any positive signals of learning, understanding, or realization. Table 15 shows that LM scorers of resultant student behavior are able to obtain reasonable alignment with teachers' annotations.

LM Transcript Scoring Performance		
Model	**Evaluation Split**	**Student Outcome**
Claude 4.6 Opus	iteration	0.8179
Claude 4.6 Opus	held-out	0.7749
Claude 4.8 Opus	held-out	0.8249
GPT-5.5	held-out	0.7635
Gemini 2.5 Pro	held-out	0.7876
Gemini 3.5 Flash	held-out	0.7916

Table 15 Results of an LM transcript scorer that considers students' behavioral outcomes in the result description of SAR annotations.

Author Contributions

Here, we indicate each authors' main contributing role(s) in our work.

- Evaluation methodology and design: Lucy Li, Kyle Lo, Kajal Patel, Ryan Knight, Albert Zhang
- Data acquisition, project management and annotation collection: Ryan Knight, Albert Zhang, Rebecca Bowie, Daniel Halper, Haripriya Valayaputter
- Experimentation, implementation, and analysis: Lucy Li, Albert Zhang, Kajal Patel, Alexis Ross
- Project infrastructure: Ryan Knight, Albert Zhang, Kyle Lo, Daniel Halper
- Mentorship, advising, program management and broader strategy: Susanna Loeb, Ana Ribeiro, Julian Bernado, Jacob Andreas
- Paper writing: Lucy Li, Ryan Knight, Kajal Patel, Albert Zhang, Kyle Lo