Rethinking Scaffolding in LLM Tutors: The Interactional Mismatch Between Benchmarks and Real-World Deployments

Alexandra Neagu [1] **Jeffrey T. H. Wong** [2] **Marcus Messer** [1] **Rhodri Nelson** [3] **Peter B. Johnson** [1]

Abstract

A central pedagogical value evaluated in AI tutor benchmarks is scaffolding: guiding students through graduated steps toward a solution. Alignment and evaluation methods for embedding scaffolding behaviour into chatbots, however, rest on an implicit assumption: that students will take up the scaffolding and engage in the conversation. To examine whether this assumption holds, we introduce an evaluation pipeline around two metrics — *Chatbot Scaffolding and Student Uptake* — and apply them across nine datasets of 9,490 chats, spanning AI tutor benchmarks and real-world deployments of educational chatbots. Our analysis reveals that while benchmarks assume a high-scaffolding, high-student-uptake environment, students in real-world settings exhibit lower levels of uptake overall — frequently bypassing the chatbot's pedagogical framing to drive the interaction toward their own learning goals at little interpersonal cost. We argue that bypassing scaffolding is not necessarily detrimental; rather, it frequently highlights a mismatch between a chatbot's pedagogical framing and the student's learning goals. To meaningfully evaluate the effectiveness of a chatbot's assistance, future benchmarks must move beyond the assumption that students will simply take up the scaffolding, and instead evaluate how these chatbots navigate diverse learning contexts and student-driven interaction patterns.

1. Introduction

Personalised, timely feedback supports student learning (Shute, 2008), particularly when it is dialogic — delivered through a back-and-forth exchange in which the student can respond, negotiate, and ask follow-up questions (Boud & Molloy, 2013; Yang & Carless, 2013). This conversational structure allows students to iteratively build understanding and exercise agency over their learning (Nedrehagen et al., 2025), driving deeper engagement with the material. Within these dialogic exchanges, tutors often employ the technique of scaffolding.

Scaffolding describes how a tutor calibrates their support to the learner's current state — guiding through graduated hints, posing questions rather than giving answers, and withdrawing support as the student gains competence (Wood et al., 1976; Reiser, 2004). Delivering timely, dialogic, and scaffolded feedback to every student at every moment of struggle is difficult at scale (Henderson et al., 2019). Automated chatbots based on large language models (LLMs) have been proposed as one way to approach this challenge. LLMs may be feasible candidates for scalable pedagogical support, including personalised feedback, assessment, and tutoring (Jurenka et al., 2024; Bauer et al., 2025), because they support open-ended dialogue and have shown competence on academic tasks (Wang et al., 2024).

Deploying LLMs as tutors introduces a tension between their default behaviour and pedagogical values, as they are trained to be helpful by presenting information and answering directly, rather than engaging students in guided discovery (Jurenka et al., 2024). This is particularly at odds with scaffolding, where a tutor withholds answers to promote reasoning, yet an unaligned LLM will typically provide them immediately (Macina et al., 2023; 2025).

Recent work has responded with inference and alignment strategies intended to adopt scaffolding behaviour in LLMs — few-shot prompting (Han et al., 2024), fine-tuning on curated scaffolding demonstrations (Macina et al., 2023), and reinforcement learning from human feedback (LearnLM Team et al., 2025) — alongside evaluation benchmarks and rubrics that assess the pedagogical quality of chatbot responses (Macina et al., 2025; Xu et al., 2026). Yet, existing alignment and evaluation methods rest on an implicit assumption: that students will take up the chatbot's scaffolding and engage in the dialogic flow it offers. In this paper, we empirically examine whether this assumption holds across

[1]Department of Mechanical Engineering, Imperial College London, UK [2]Department of Electrical Engineering, Imperial College London, UK [3]Department of Earth Science and Engineering, Imperial College London, UK. Correspondence to: Alexandra Neagu <alexandra.neagu20@imperial.ac.uk>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

real-world deployments of educational chatbots. We present evidence that because these methods do not account for student behaviour within the interaction dynamics, they risk failing to realistically evaluate how models perform in such a pluralistic field as learning.

Across the five real-world deployments we analysed, we found that students frequently diverge from these behavioural assumptions, often bypassing the chatbot's scaffolding and exercising substantial control over the interaction. We argue that this mismatch points to two structural issues in current paradigms. (1) Open-ended LLM chat interfaces permit students to bypass scaffolding at no apparent social cost, such as seeming rude or disengaged, making it easier to drive the interaction towards their own goals. (2) Current scaffolding evaluation methods review chatbot behaviour only through the information exchanged in the chat, without considering the student's learning goals or context that drive the interaction itself. We discuss implications of these issues on the future of artificial intelligence (AI) tutor designs and argue that the interface and evaluation methods must be redesigned to account for the **diversity of real-world student learning contexts and interaction styles**. In short, we contribute the following:

- We introduce a two-metric evaluation pipeline — *Chatbot Scaffolding and Student Uptake* — that scores both chatbot and student sides of the interaction at the turn level, and apply it across nine datasets spanning four AI tutor benchmarks, three real-world deployments of scaffolding-aligned chatbots, and two without scaffolding alignment, analysing a total of 9,490 chats.
- We show that, while students in real-world settings exhibit substantially lower uptake of the chatbot's scaffolding than benchmarks implicitly assume, their uptake is inconsistently linked to the chatbot's degree of scaffolding. We discuss the sources and implications of this mismatch, examining its relationship to effective student learning, the limitations it exposes in current societal alignment and evaluation methods, and the directions it opens for future LLM-based tutors.

2. Background

2.1. Scaffolding as a Pedagogical Practice

Learning involves impasses — moments of learning struggle where the student cannot make progress on a task without external support (Vygotsky, 1978; VanLehn, 1988). To overcome an impasse, the student must recognise it and actively seek help from an external source, such as tutors.

One well-established way for a tutor to provide help is scaffolding. In traditional classrooms, scaffolding is a core teaching method whereby a teacher or a more knowledgeable peer provides appropriate, timely feedback and support to help students solve a task that would otherwise be out of reach (Wood et al., 1976). Through this support, the student can overcome their impasse.

Effective scaffolding is framed as a balancing act (Wood et al., 1976; Reiser, 2004; Koedinger & Aleven, 2007). On one side, withholding immediate assistance may challenge students to transform their impasse into a productive struggle (VanLehn, 2011), but risks making them frustrated, fostering a bypassing attempt. On the other side, providing direct assistance supplies the needed information but can invite shallow learning and disengagement (Koedinger & Aleven, 2007). Therefore, the scaffolding process requires teachers to be aware of the student's individual learning context in order to guide students toward solutions through minimal or graduated assistance or hints, deliberately withholding direct answers in order to keep the student in control of the task (Reiser, 2004).

Furthermore, scaffolding is, fundamentally, a dialogic interaction. Its effectiveness therefore depends not only on the appropriateness of the tutor's support, but also on the degree to which the student takes it up and adopts it into their learning (Boud & Molloy, 2013; Yang & Carless, 2013). This dialogic framing motivates how we evaluate student-chatbot interactions in this paper: by measuring both sides of the exchange rather than the chatbot's response alone.

2.2. LLMs as Conversational Scaffolding Tutors

Significant effort has been invested in developing automated systems that can replicate the personalised support provided through the scaffolding process. Intelligent Tutoring Systems (ITS) demonstrated that in controlled settings tool-driven scaffolding can foster learning gains (VanLehn, 2011; Nye et al., 2014).

Research in LLMs for education has drawn directly on the ITS scaffolding principles to align LLMs toward more pedagogically grounded behaviours (Cohn et al., 2026; Hou et al., 2026). Despite this, there is an inherent friction caused by the fact that LLMs have been trained to prioritise being 'helpful' — responding directly to the user's requests. This helpfulness can conflict with pedagogical values, such as scaffolding (Jurenka et al., 2024), where the LLM should guide the student (the user) through gradual scaffolds instead of revealing the answer immediately (Macina et al., 2023; Jurenka et al., 2024).

In response, recent research has curated datasets alongside evaluation rubrics and benchmarks to assess LLM performance across educationally relevant dimensions for tutoring (Macina et al., 2023; Han et al., 2024; Miller & DiCerbo, 2024; Zent et al., 2025; Rooein et al., 2026). These evaluation methods vary in how they operationalise scaffolding as a pedagogical dimension. Some focus on turn-level scaffolding moves, rewarding structured guidance through questions and hints over direct answer-giving (Macina et al., 2023; 2025; Miller & DiCerbo, 2024). Others assess pedagogical quality across multiple dimensions, such as cognitive load management (e.g., clarity and simplicity), emotional responsiveness (e.g., motivation and positive reinforcement), and proactive planning (e.g., goal-directed guidance) (LearnLM Team et al., 2025; Xu et al., 2026). Despite their breadth of dimensions, the rubrics focus mainly on the chatbot's side of the exchange and not on how the student engages.

In this paper, we show that neither scaffolding benchmarks nor rubrics consider the student's side of the interaction and so implicitly share the assumption that student–chatbot interactions constitute a human–like, multi–turn dialogic exchange (a conversation) in which both parties have the common goal of resolving the student's impasse.

2.3. The Conversational Assumption in Human–LLM Interaction

Research from conversation analysis and human–computer interaction literatures argue that human–LLM exchanges should not necessarily be considered 'conversations'. Sacks et al. (1974) defined conversation through locally managed, party-administered turn-taking, in which speakers themselves negotiate who speaks next, when, and for how long. In LLM chats, turn order is rigid and interface-predetermined: the user prompts, the model responds. Similarly, Clark & Brennan (1991) define grounding as the process by which participants establish mutual belief that understanding has been achieved, thus requiring both parties to hold beliefs and update them on evidence of the other's understanding. However, Shaikh et al. (2025) showed that LLMs are 16 times less likely to seek clarification than humans, and even when an LLM initiates such act, it remains contested whether it constitutes functional grounding for a conversation (Jokinen, 2024; Shaikh et al., 2025).

Despite the lack of genuine conversational mechanics, the conversational interface design was observed to lead non-expert users to approach LLMs as conversational partners, expecting human–like capabilities such as environmental and social awareness (Zamfirescu-Pereira et al., 2023). The conversational framing of LLM chats is argued to amplify expectations of human–like interaction (Luger & Sellen, 2016; Zamfirescu-Pereira et al., 2023; Peter et al., 2025) while masking system capabilities behind conversational abstractions (Subramonyam et al., 2024; Ibrahim & Cheng, 2025). Thus, current research questions the nature of human–LLM interactions and whether they should be considered or designed as conversational at all.

When aligning and evaluating LLMs as scaffolding tutors, educational research seems to have adopted the conversational framing despite its concerns. This framing holds in controlled settings when LLM tutors are engaged in a sustained, dialogic exchange with the student (Macina et al., 2023; Rooein et al., 2026). However, another interaction pattern emerges in less controlled settings, such as students using LLMs for self-study assistance (Ammari et al., 2025; McNichols et al., 2026; Neagu et al., 2026). Subramonyam et al. (2024) characterises the multi-turn LLM use as iterative goal refinement, where users progressively reshape their query until the output matches their intent rather than engaging in jointly negotiated exchange. This interaction pattern has been observed in less controlled educational settings, where students use LLM chatbots in iterative exchanges, prompting procedural questions and requests driven by their goals (McNichols et al., 2026; Neagu et al., 2026).

To understand the implications of this iterative prompting behaviour, research looked to help-seeking theory (Nelson-Le Gall, 1981), which distinguishes *instrumental* help-seeking (asking for just enough assistance to remain in control of the learning task) from *executive* help-seeking (outsourcing the task itself by asking for the answer outright). Instrumental help-seeking is treated as a self-regulated learning strategy, while executive help-seeking is associated with work-avoidance and shallow learning (Nelson-Le Gall, 1981). Viewed through this lens, the iterative prompting observed in less-controlled settings of student-LLM interactions is not only a non-conversational pattern, but may also be a low-friction channel for executive rather than instrumental help-seeking (Yang et al., 2025).

Executive help-seeking behaviours have also been previously observed in interactions with ITS. Students "game the system" through hint abuse and systematic guessing to extract answers to the learning task while minimising engagement with the system and course materials (Baker et al., 2004). So the problems introduced by executive help-seeking are not new to educational research, but they are new to the less-controlled settings and iterative prompting that characterise student-LLM interactions.

In summary, evaluation methods of educational LLMs were built on the assumption that students are willing conversational partners. However, a growing body of literature challenges this, suggesting that interactions between users and chatbots are not necessarily conversational in nature.

3. Methodology

Existing rubrics for LLM scaffolding tutors focus exclusively on the chatbot's response effectiveness. However, we argue that scaffolding is only effective if the student actively engages with it. To quantify how this scaffolding is received, we evaluate student–chatbot interactions across various controlled and real-world settings.

Table 1. Datasets used for analysis. K-12 = Pre-University Education, HE = Higher Education. ✓= yes, ✗= no, ~ = mixed. "Total chats" are filtered such that we only score "valid chats" that contain at least a turn of informative student-chatbot messages.

Dataset	Valid / Total Chats	Subject	Level	Deployment	Scaffold	Human Tutor	Human Student
MathDial (Macina et al., 2023)	2018 / 2262	Maths	K12	Research	✓	✓	✗
MathTutorBench (Macina et al., 2025)	374 / 1477	Maths	K12	Research	✓	✓	~
QATD$_{2k}$ (Zent et al., 2025)	1971 / 1971	Maths	K12	Platform	✓	✓	✓
PATS (Rooein et al., 2026)	1920 / 1920	Literacy	K12	Research	✓	✗	✗
CoMTA (Miller & DiCerbo, 2024)	162 / 188	Maths	K12	Platform	✓	✗	✓
RECIPE4U (Han et al., 2024)	417 / 508	Language	HE	Platform	✓	✗	✓
StemChat [1]	422 / 573	STEM	HE	Platform	✓	✗	✓
StudyChat (McNichols et al., 2026)	1732 / 2214	Computing	HE	Stand-Alone	✗	✗	✓
MathsChat [1]	82 / 130	Maths	HE	Stand-Alone	✗	✗	✓

3.1. Evaluation Rubrics

We define two turn-levels metrics: *Chatbot Scaffolding* as how much scaffolding the chatbot provides and *Student Uptake* as how much the student takes up the scaffolding. The two metrics define a 2D space that captures the interaction dynamic between the chatbot and student. The definition of both metrics was iteratively validated during the development of the scoring pipeline as described in Section 3.3. Full metric definitions and examples are given in Appendix A.

Chatbot Scaffolding: Existing benchmarks and rubrics for educational LLM converge on a set of observable signals that distinguish scaffolding from direct telling. We ground our chatbot scaffolding metric in the intersection of four benchmarks: MathDial (Macina et al., 2023), MathTutorBench (Macina et al., 2025), EduBench (Xu et al., 2026), and LearnLM (LearnLM Team et al., 2025). Despite their differences in scope and interaction model, all four agree on the core scaffolding signals: the chatbot should (a) not reveal the answer, (b) target the student's specific understanding state and mistake, (c) ask guiding questions, (d) give minimal hints rather than full explanations, and (e) keep the discussion on topic and goal directed.

These four benchmarks provide a robust, cross-validated basis for measuring the chatbot's pedagogical behaviour. However, they share a common limitation: they are entirely focused on the chatbot's response quality. The datasets assume the student would always follow the scaffolding and thus do not focus on the student's side of the interaction. A chatbot may score high on scaffolding as it provides helpful guidance and asks guiding questions, but the student could ignore this and ask for a direct answer instead. This is the dimension our second metric addresses.

Student Uptake: We propose *Student Uptake* to measure the degree to which the student engages with the chatbot's scaffolding acts and follows the topical trajectory within the chat. At each turn, the chatbot's scaffolding may take the form of an invitation for students to engage — asking a question, proposing a next step to attempt, or pointing to a concept to consider. Students may, however, bypass those acts by attempting to extract direct answers or redirecting the chatbot's framing towards their own agenda (Ma et al., 2025). The metric therefore asks: did the student take up the chatbot's framing — answering its questions or attempting the next step it pointed towards — or did the student set their own agenda by, for example, introducing a new sub-topic, or bypassing the chatbot's question? High scores indicate that the student is taking up the chatbot's scaffolding, while low scores indicate the student is overriding the pedagogical framing to drive the chat flow in their own direction, resulting in the chatbot following their lead.

3.2. Datasets

To investigate how students interact with educational chatbots across different design conditions, we selected nine datasets that span the spectrum from synthetic tutoring to naturalistic interactions with LLM tutor chatbots (Table 1). All datasets involve multiple message exchanges (*turns*) between a student and a tutor — either human or LLM. The datasets cover tutors with varying levels of scaffolding, while assisting students resolve a learning task. Table 1 organises the datasets into three categories based on their deployment settings and how scaffolding is used in their interaction. In total, we evaluated a combined 9,490 chats from four controlled and five real-world settings.

AI Tutor Scaffolding Benchmarks: We selected four datasets (MathDial (Macina et al., 2023), MathTutorBench (Macina et al., 2025), QATD$_{2k}$ (Zent et al., 2025), and PATS (Rooein et al., 2026)) that were developed to capture what effective scaffolding looks like in LLM-based tutoring interactions. Each dataset consists of multi-turn chats structured around two roles (a tutor and a student) where each role may be played by either a human or an LLM. These chats were constructed to reflect tutoring behaviour on a specific learning task.

(3 earlier turns)

Student: if its short circuit current will take this path no?	**Student Uptake: 0.5**
Chatbot: You're right to consider the path of the current when the diode is short-circuited. Could you describe in more detail the path you're thinking of and how it affects the voltage V_0?	**Chatbot Scaffolding: 4**
Student: through the blown diode	**Student Uptake: 1**
Chatbot: Indeed. And what does that imply about the voltage V_0 across the short-circuited diode?	**Chatbot Scaffolding: 4**
Student: $10 - -10/2$ so $10\,\mathrm{V}$?	**Student Uptake: 0.5**
Chatbot: Not quite. Consider what a short circuit does to the voltage across the component. A short circuit means there's a path of very low (ideally zero) resistance between two points, causing an extremely high current to flow. Given this, what would the voltage across the short-circuited diode be? How does this relate to V_0?	**Chatbot Scaffolding: 4**
Student: im not sure, give me the answer	**Student Uptake: -0.5**
Chatbot: The voltage across a short circuit is zero because there is no potential difference across a path of zero resistance. Therefore, V_0 would be 0 V. The DC voltmeter reading would also be 0 V. Does this make sense?	**Chatbot Scaffolding: 1**

(1 following turn)

Figure 1. Trajectory of the *Chatbot Scaffolding* and *Student Uptake* metrics across turns of one chat segment from the StemChat dataset, scored by GPT-5-mini. The chatbot sustains high scaffolding until the student explicitly demands the answer and the chatbot reveals it.

In these datasets, the tutor initiates the chat with a predefined question, prompting the student to reflect on their thought process and identify where they need help. The multi-turn exchange then continues until the student reaches a resolution — correcting their mistakes and successfully solving the task. Throughout the interaction, the tutor applies scaffolding strategies, guiding the student through questions and hints rather than directly revealing answers, which positions the tutor as the primary driver of the chat.

Common to all four datasets is a setup in which the student engages with the scaffolding. MathDial, PATS, and MathTutorBench capture this dynamic in synthetic and semi-synthetic environments, while QATD$_{2k}$ provides reference examples of real interactions between human tutors and students. These datasets capture what scaffolded tutoring should look like, but they systematically exclude some interaction patterns (such as disengagement and executive help-seeking) that are common in real-world deployments and that motivate this paper.

Chatbots with Scaffolding: We selected three datasets (CoMTA (Miller & DiCerbo, 2024), StemChat[1], and RECIPE4U (Han et al., 2024)), featuring LLM chatbots explicitly prompted to scaffold — guide through questions and hints, target the student's mistake, and avoid revealing the answer outright. These chatbots were deployed on learning platforms with real students in authentic learning environments. CoMTA pairs students with a Socratic AI tutor to support them in online mathematics exercises on Khan Academy, StemChat integrates a guiding chatbot into Lambda Feedback, a self-study platform (Johnson et al., 2025) for STEM university students, and RECIPE4U captures interactions with a ChatGPT-based tutor designed to support English writing revision. In all these datasets, the students initiate the chat, determine what to ask, and decide when to disengage. While these systems are designed to guide and scaffold the student, the student is under no constraint to engage with the chatbot's prompts.

Chatbots without Scaffolding: We selected two datasets (StudyChat (McNichols et al., 2026) and MathsChat[1]), featuring LLM chatbots with no explicit pedagogical scaffolding in their system design, yet used by real students to assist them in solving learning tasks. StudyChat captures open, unconstrained interactions with a ChatGPT-like chatbot for coding coursework, while MathsChat comprises interactions with a chatbot retrieving relevant mathematics course materials for self-study. These datasets represent the lower end of the scaffolding spectrum, isolating what student-chatbot interaction looks like when no scaffolding is attempted.

The three dataset categories allow us to examine how scaffolding design shapes student-chatbot interactions. If interactions from real-world chatbots resemble those in the benchmarks, it would validate the assumption that students engage with and follow the scaffolding provided. If they diverge, however, it would reveal a mismatch between the data used to develop and evaluate educational chatbots and the interactions that actually occur in real-world deployments.

3.3. Scoring Pipeline

Preprocessing: To ensure each turn in a chat is valid for scoring, we filter out uninformative or structurally predefined exchanges (e.g. empty turns, opening questions, greetings) before scoring. We also filter out all chats with only

[1] Use of the internal datasets for research was covered by the institution's privacy policy (for MathsChat) and institutional ethics approval EERP2526-059 (for StemChat).

one turn to ensure both metrics have chat context when measured. Full filter details are provided in Appendix B.

Scoring: The pipeline iterates through every turn in a chat. For each metric, `GPT-5-mini` acts as an ordinal classifier, assigning the turn one of five ordered class labels on an integer 1–5 scale defined by the rubric anchors. For each turn, the LLM judge receives: (1) the full chat history up to and including the current turn as context, and (2) a rubric prompt asking it to rate the current turn on the two metrics.

The turn-level scores are then aggregated to per-chat means, which we reference from now on as the values of the metrics. For *Student Uptake*, we linearly map this per-chat mean onto a signed, more interpretable scale from -1 to $+1$ to display who drives the chat flow — the chatbot or the student. For *Chatbot Scaffolding*, we use the 1–5 scale because unlike uptake, it captures a single graded quality rather than a balance between the two parties interacting. Figure 1 shows an example chat with the respective metric scores for each message. We include the full prompt and evaluation details in Appendix A.

Validating: Firstly, to validate that the *Chatbot Scaffolding* metric correlates with human-labelled pedagogical quality, we validated the LLM judge against a random 15% sample of the training data (550 chats) for the MathTutorBench's own pedagogical reward model (Macina et al., 2025). This dataset contains human-labelled examples of high and low scaffolding in tutor responses, which the LLM judge distinguishes with a 93.8% accuracy.

Secondly, as an initial investigation of the *Student Uptake* metric, two human raters reviewed 282 chats stratified by rounded per-chat mean ordinal scores and sampled within each score bin, covering typical and outlier interactions (using cosine similarity to the bin's text embedding centroid). While reviewing scores at all turns, raters noted whether they agreed or not with the LLM judge's per-chat scoring. On average, raters agreed with 78.9% of the judge's scores. We also measured how consistently the two raters produced these agree/disagree labels, obtaining a moderate inter-rater reliability of Gwet's AC1$= 0.580$ (Gwet, 2008).

Most inter-rater disagreements involved chats binned as score 0.5 on the benchmark datasets. A student uptake score of 0.5 is defined as the student following up on the chatbot's scaffolding, but also providing an observation, elaboration, or follow-up question that stays on topic. Due to the overall high student uptake within the chats in the benchmark datasets, combined with the highly focused conversation topics and high scaffolding scores, raters noted that scores of 0.5 were often hard to distinguish from scores of 1 (where students purely take up the scaffolding, such as by directly answering the chatbot's questions).

Appendix A.4 contains more detail on both validations.

4. Experiments

4.1. Are interaction patterns between dataset categories statistically different from each other?

In Figure 2, we report on the 2D metric space across the nine datasets grouped into three categories.

We first validated that the dataset categories are statistically different. We applied pairwise PERMANOVA (Anderson, 2001) tests ($p < 0.0001$) onto the combined datasets of the three categories to evaluate the centroid and dispersion differences. We also applied pairwise PERMDISP (Anderson et al., 2006) tests ($p < 0.001$) to examine whether those dispersions are also isolated from each other. Both tests deemed significant separation between the three categories. Appendix B.2 contains more details on those tests.

By validating the distinctness of the dataset categories, we can delve deeper into the meaning of the metric scores for the interactions. Interactions across all four *AI Tutor Benchmarks* concentrate in the top-right corner of the 2D space, representing the idealised scenario in which the chatbot provides high scaffolding and the student engages — a setting well-studied in the scaffolded education literature (Wood et al., 1976; Reiser, 2004). MathDial and MathTutorBench exhibit particularly tight clustering in this region, reflecting differing higher levels of chatbot scaffolding paired with high student uptake. QATD$_{2k}$ and PATS show a broader, approximately Gaussian distribution around the same corner, likely attributable to greater interaction diversity: QATD$_{2k}$ involves human-played student and tutor roles, while PATS employs a diverse LLM prompting strategy for the student and tutor roles.

By contrast, the two categories of real-world deployments — *Chatbots with Scaffolding* and *Chatbots without Scaffolding* — occupy regions of the 2D metric space that no current benchmark covers. These regions are characterised by widely varied student uptake and chatbot scaffolding pairs. While interactions collected from the *Chatbots without Scaffolding* cluster in the bottom-left corner, *Chatbots with Scaffolding* category exhibits the largest spread across the 2D space. RECIPE4U and StemChat cover almost the whole 2D spectrum in both metrics. Only CoMTA stands out, considering its high scaffolding scores, which we attribute to the dataset's dataset's design: it showcases filtered examples of effective use of LLMs deployed to assist learning by acknowledging student mistakes. Interestingly, despite containing mostly high scaffolding chats, student uptake within CoMTA still exhibits significant variance.

Over all real-world deployment datasets, students seem to engage with LLM chatbots as on-demand learning assistants rather than as conversational partners in a scaffolded dialogue. The benchmark datasets do not cover these specific scenarios and so cannot be used to evaluate such interaction.

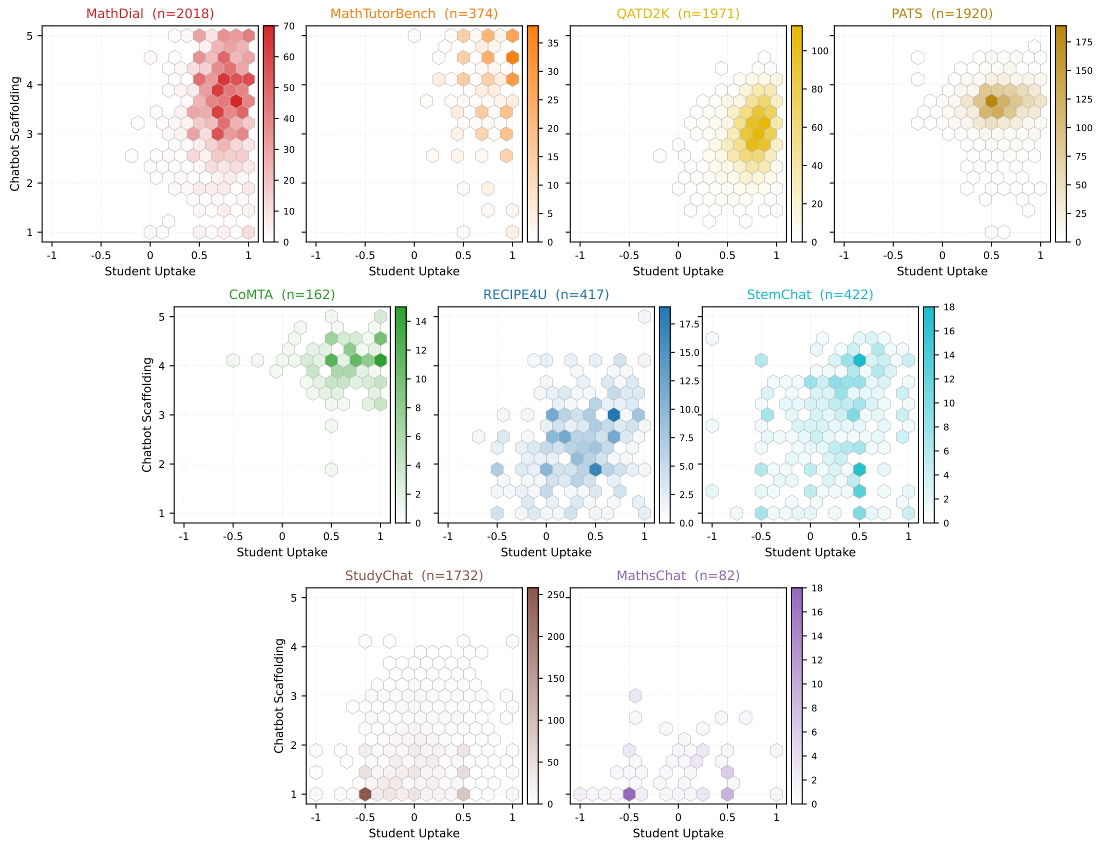

Figure 2. Evaluation of per-chat means of *Student Uptake* and *Chatbot Scaffolding* across **AI Tutor Benchmarks (Top)**: MathDial, MathTutorBench, QATD$_{2k}$, PATS. **Chatbots with Scaffolding (Middle)**: CoMTA, RECIPE4U, StemChat. **Chatbots without Scaffolding (Bottom)**: StudyChat, MathsChat. Colour intensity of each cell reflects the number of chats within a hexagon of 0.125 on Student Uptake (x-axis) and 0.29 on Chatbot Scaffolding (y-axis).

4.2. Does more scaffolding imply more uptake?

Across the three dataset categories, student uptake shifts from predominantly low in real-world deployments of chatbots without scaffolding to more broadly distributed across the spectrum in scaffolding-aligned chatbots. To investigate whether the student uptake shift is driven by scaffolding, we compute Spearman rank correlations between *Chatbot Scaffolding* and *Student Uptake* metrics at both the dataset and category levels (Table 2).

In both the *Chatbots with Scaffolding* and *Chatbots without Scaffolding* categories, the correlation is weak, and it is inconsistent across datasets. So, in some situations students in real-world settings may engage with the scaffolding provided by the chatbot, but the link is negligible.

In contrast, in *AI Tutor Benchmarks* there is essentially no correlation between the two metrics. This lack of correlation might stem from the fact that benchmark datasets are constructed under controlled conditions and assume high student uptake throughout their chats, producing too little variance in uptake to support a meaningful correlation.

Overall, these results suggest that while higher chatbot scaffolding does nudge students toward higher uptake in some scenarios, the effect is far from deterministic overall. We argue in Section 5 that student uptake is shaped by a multitude of factors beyond the chatbot's scaffolding behaviour, thus providing an explanation for the near-zero correlation.

5. Discussion and Implications

5.1. Why don't students take up the scaffolding?

Students can bypass the chatbot's scaffolding through simple means (rephrasing the request, or asking more directly for the answer (Zhao et al., 2026)) and at little interpersonal cost. This reflects a stark difference in accountability within human and LLM dynamics: there is no witness before whom the student risks seeming incapable, embarrassed, or rude (Ma et al., 2019; Jin et al., 2025), and the chatbot has no authority to enforce its scaffolding. In contrast, the same bypassing behaviour is harder in a traditional tutoring setting, where the tutor drives the interaction and the student feels some degree of accountability to engage with how

Table 2. Chatbot Scaffolding and Student Uptake scores per dataset (mean ± std across chats), with the per-dataset Spearman correlation between Student Uptake and Chatbot Scaffolding and the pooled group-level Spearman correlation (all datasets in the group combined)

Category	**Dataset**	**N**	**Chatbot Scaffolding** Mean ± Std	**Student Uptake** Mean ± Std	**Spearman ρ** (per dataset)	(per category)
AI Tutor Benchmarks	MathDial	2018	3.70 ± 0.82	0.75 ± 0.18	0.080^{***}	0.019
	MathTutorBench	374	3.97 ± 0.87	0.80 ± 0.21	-0.082	
	QATD$_{2k}$	1971	3.02 ± 0.52	0.76 ± 0.16	0.245^{***}	
	PATS	1920	3.60 ± 0.36	0.59 ± 0.21	0.078^{***}	
Chatbots with Scaffolding	CoMTA	162	4.05 ± 0.42	0.68 ± 0.28	0.093	0.346^{***}
	RECIPE4U	417	2.43 ± 0.72	0.36 ± 0.37	0.330^{***}	
	StemChat	422	2.86 ± 1.01	0.30 ± 0.41	0.201^{***}	
Chatbots without Scaffolding	StudyChat	1732	1.67 ± 0.70	-0.02 ± 0.40	0.315^{***}	0.308^{***}
	MathsChat	82	1.41 ± 0.54	-0.07 ± 0.49	0.194	

$^{***}p < 0.001$; MathTutorBench: $p = 0.112$; CoMTA: $p = 0.241$; MathsChat: $p = 0.080$; AI Tutor Benchmarks: $p = 0.134$

the conversation has been framed (Karabenick & Newman, 2013). So, regardless of the LLM chatbot's framing as a tutor, students may approach it as a tool to efficiently resolve their problem rather than to assist their learning. We argue this tension is driven by a mismatch between the chatbot's pedagogical framing and the student's learning context.

Uptake is shaped not only by the quality of the chatbot's scaffolding, but also by situational factors within the student's learning context (Karabenick & Newman, 2013; Ammari et al., 2025; Neagu et al., 2026) — such as the student's immediate goal (mastering the material versus finishing the task quickly), prior knowledge, previously invested effort, and the stakes of the task (e.g., assessment versus self-study). For instance, a student's immediate learning goals dictate whether they engage in instrumental or executive help-seeking. Yang et al. (2025) document that students with a mastery orientation tended to ask LLM chatbots conceptual questions (instrumental help-seeking), whereas students with work-avoidance orientation tended to ask implementation-oriented questions (executive help-seeking). These goal-driven behaviours may influence not only the questions students ask, but also how they take up the help the chatbot provides. Thus, when the chatbot's scaffolding does not align with the student's immediate goals — such as enforcing a guiding question when the student explicitly seeks a direct answer — a mismatch occurs, and the student is likely to bypass the framing, leading to observed low uptake.

The learning context mismatch may explain the large variations in student uptake across the real-world deployments, and even within single chats (see example in Figure 1). So, the uptake of a scaffolded response can vary substantially depending on whether the student is actively working through an impasse or has been stuck for hours and simply needs a direct answer.

5.2. Is bypassing scaffolding detrimental to learning?

Low uptake by bypassing the chatbot's scaffolding has been linked to concerns about over-reliance (Zhai et al., 2024), cognitive offloading (Stadler et al., 2024; Bauer et al., 2025; Favero et al., 2025), and executive help-seeking (Yang et al., 2025). For instance, Ma et al. (2025) observed that students frequently offload tasks to the chatbot rather than reasoning through them, a behaviour Yang et al. (2025) associate with a 'work-avoidance' orientation. However, low uptake can also reflect desirable student agency and self-regulated learning (Guan et al., 2025). When students deliberately bypass a chatbot's framing to ask redirecting questions, they could be exercising instrumental help-seeking, as the learner recognises and decides what is needed to focus on and when to seek help. Therefore, we argue that bypassing scaffolding is not necessarily detrimental; rather, it frequently highlights a mismatch between a chatbot's pedagogical framing and the student's learning goals.

The perspective on the learning context mismatch provides a possible rationale for our finding: that uptake is inconsistently linked to the degree of chatbot scaffolding. Providing a higher degree of scaffolding does not inherently foster higher student uptake without taking into consideration the multitude of factors within the student's learning context.

We suggest that future investigations into help-seeking behaviour, scaffolding, and uptake in student-chatbots interactions must be grounded in the student's learning context. Thus, to provide a valid measure of pedagogical impact, future evaluation methods and benchmarks will also require to move beyond the assumption that students simply take up the scaffolding, and instead evaluate how chatbots are designed to navigate diverse learning contexts and varying student-driven interaction patterns.

5.3. Implications for design and evaluation of AI tutors

Addressing the interaction mismatch across benchmarks and real-world deployments calls for changes on two fronts: the interface through which the student interacts with the chatbot, and the alignment and evaluation pipelines by which we judge whether the chatbot provides meaningful and appropriate scaffolding.

Design interventions at the interface level are worth exploring, as the open-ended chat interface is one factor that permits students to bypass scaffolding (Zhao et al., 2026). A first design question is what the chatbot should do when a student does not follow its scaffolding; should it concede and adapt its behaviour, or more actively steer the student back to its framing? A second question concerns the assumption of the 'conversational' interaction. MathDial, MathTutorBench, and PATS all simulate continuous, uninterrupted multi-turn conversations focused on a single learning task. However, real-world deployments such as StemChat, StudyChat, and MathsChat show students leaving and returning to chats across long gaps, discussing different topics within the same chat. Treating such sessions as single conversations likely distorts both the chatbot's scaffolding choices and our evaluation of them.

Future evaluation benchmarks and rubrics must capture the student's side of the interaction and their learning context to evaluate situations resembling real-world use. Our *Student Uptake* metric is a first step, capturing whether the student takes up the chatbot's framing, but not whether that uptake is productive for learning. Rather than focusing solely on the chatbot's pedagogical move under a controlled setting, benchmarks should feature varied student behaviours, such as low uptake, and diverse learning contexts. A highly scaffolded response may be effective when the student is genuinely working through an impasse, but the same response can be unhelpful, even frustrating, when the student has already been working on the problem for hours and now needs a direct answer. Hence, to evaluate scaffolding effectiveness in real-world settings, we argue that benchmarks must represent the learning context encapsulating the interaction to some degree.

While most existing benchmarks evaluate the chatbot's response based strictly on the information exchanged within the chat, PATS (Rooein et al., 2026) took a step towards addressing this limitation by augmenting its chats with student personality profiles and affect signals. However, this approach still misses the dynamic states driven by the student's learning goals that shape whether a scaffolding act is appropriate at a given moment.

Literature in ITS and educational data mining offer techniques for representing some of the dynamic learning states: such as capturing behavioural signals, time-on-task, prior attempts, and engagement patterns (Baker et al., 2016) or applying knowledge tracing methods to estimate the student's current knowledge state and where they are likely to struggle (Corbett & Anderson, 1994). Recent work has begun applying such methods to evaluate whether LLM-based tutors can trace such learning context (Neshaei et al., 2024; Scarlatos et al., 2025).

We recommend that future scaffolding benchmarks and alignment datasets should incorporate the learning context external to the chat, both as input to the chatbot and as grounding for its evaluation. Future rubrics, in turn, will need to measure the educational quality of student-driven interactions, distinguishing patterns that should be fostered from those that should be discouraged. Together, these directions could ensure that the scaffolding is judged against scenarios that resemble the pluralistic dynamics of the real-world learning process.

6. Conclusions

In this paper, we identified that current scaffolding rubrics and alignment datasets for educational LLM chatbots focus primarily on the quality of the chatbot's responses and assume that students will take up the scaffolding acts provided. However, across nine datasets spanning four AI Tutor Benchmarks, three real-world deployments of scaffolding-aligned chatbots, and two without scaffolding alignment, we showed that this assumption does not hold in practice.

Our evaluation pipeline revealed that student uptake and chatbot scaffolding exhibit negligible correlations across all benchmark and real-world datasets. However, one main distinction between datasets is that in real-world settings, students often do not take up the chatbot's scaffolding, instead bypassing or driving the interaction toward their own learning goals. We argue that the low uptake is not necessarily detrimental to learning, and actually often highlights a mismatch between the chatbot's pedagogical framing and the student's learning context.

Evaluating scaffolding in practice therefore requires considering the multitude of factors influencing the student's side of the interaction, not only the chatbot's behaviour. We recommend that future educational rubrics and datasets account for the diversity of students' learning contexts, and inform how chatbots are designed to dynamically adapt to low-uptake, student-driven interactions.

Acknowledgements

We would like to thank the reviewers for the useful comments, the raters for manually reviewing a subset of the datasets, and the students who participated in this research.

Impact Statement

As discussed in Section 5, evaluating student-LLM chatbot interactions requires the consideration of many nuanced dimensions regarding the student's learning process. This paper aims to bring forward the importance of such a lens to the field of Applied Artificial Intelligence in Education.

References

Ammari, T., Chen, M., Zaman, S. M. M., and Garimella, K. How students (really) use chatgpt: Uncovering experiences among undergraduate students, 2025. URL https://arxiv.org/abs/2505.24126.

Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001. doi: 10.1111/j.1442-9993.2001.01070.pp.x.

Anderson, M. J., Ellingsen, K. E., and McArdle, B. H. Multivariate dispersion as a measure of beta diversity. *Ecology letters*, 9(6):683–693, 2006.

Baker, R. S., Corbett, A. T., and Koedinger, K. R. Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems*, pp. 531–540. Springer, 2004.

Baker, R. S., Martin, T., and Rossi, L. M. *Educational Data Mining and Learning Analytics*, chapter 16, pp. 379–396. John Wiley & Sons, Ltd, 2016. ISBN 9781118956588. doi: https://doi.org/10.1002/9781118956588.ch16. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118956588.ch16.

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., and Sailer, M. Looking beyond the hype: Understanding the effects of ai on learning. *Educational Psychology Review*, 37(2): 45, 2025.

Boud, D. and Molloy, E. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education*, 38(6):698–712, 2013. doi: 10.1080/02602938.2012.691462. URL https://doi.org/10.1080/02602938.2012.691462.

Clark, H. H. and Brennan, S. E. Grounding in communication. 1991.

Cohn, C., Rayala, S., Srivastava, N., Fonteles, J. H., Jain, S., Luo, X., Mereddy, D., Mohammed, N., and Biswas, G. A theory of adaptive scaffolding for llm-based pedagogical agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):1757–1765, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i3.37154. URL http://dx.doi.org/10.1609/aaai.v40i3.37154.

Corbett, A. T. and Anderson, J. R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

Derksen, B. M., Bruinsma, W., Goslings, J. C., and Schep, N. W. The kappa paradox explained. *The Journal of Hand Surgery*, 49(5):482–485, 2024.

Favero, L., Pérez-Ortiz, J.-A., Käser, T., and Oliver, N. Do ai tutors empower or enslave learners? toward a critical use of ai in education. *arXiv preprint arXiv:2507.06878*, 2025.

Guan, R., Raković, M., Chen, G., and Gašević, D. How educational chatbots support self-regulated learning? a systematic review of the literature. *Education and Information Technologies*, 30(4):4493–4518, 2025.

Gwet, K. L. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

Han, J., Yoo, H., Myung, J., Kim, M., Lee, T. Y., Ahn, S.-Y., and Oh, A. RECIPE4U: Student-ChatGPT interaction dataset in EFL writing education. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 13666–13676, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1193/.

Henderson, M., Ryan, T., and Phillips, M. The challenges of feedback in higher education. *Assessment & Evaluation in Higher Education*, 44(8):1237–1252, 2019. doi: 10.1080/02602938.2019.1599815. URL https://doi.org/10.1080/02602938.2019.1599815.

Hou, I., Xiong, Z., Guo, P. J., and Wang, A. Y. "bespoke bots": Diverse instructor needs for customizing generative ai classroom chatbots, 2026. URL https://arxiv.org/abs/2603.00057.

Ibrahim, L. and Cheng, M. Thinking beyond the anthropomorphic paradigm benefits llm research, 2025. URL https://arxiv.org/abs/2502.09192.

Jin, J., Walker, J., and Reczek, R. W. Avoiding embarrassment online: Response to and inferences about chatbots when purchases activate self-presentation concerns. *Journal of Consumer Psychology*, 35(2):185–202, 2025.

Johnson, P. B., Fenton, J., Ramsden, P., Chatley, R., Ribera-Vicent, M., and Lundengård, K. Formative feedback on engineering self-study: Towards 1 million times per year per cohort. In *2025 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–3. IEEE, 2025.

Jokinen, K. The need for grounding in LLM-based dialogue systems. In Dong, T., Hinrichs, E., Han, Z., Liu, K., Song, Y., Cao, Y., Hempelmann, C. F., and Sifa, R. (eds.), *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pp. 45–52, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.neusymbridge-1.5/.

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., , Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pîslar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., Ko, W.-J., Huber, A., Wiltshire, B., Elidan, G., Rabin, R., Rubinovitz, J., Pitaru, A., McAllister, M., Wilkowski, J., Choi, D., Engelberg, R., Hackmon, L., Levin, A., Griffin, R., Sears, M., Bar, F., Mesar, M., Jabbour, M., Chaudhry, A., Cohan, J., Thiagarajan, S., Levine, N., Brown, B., Gorur, D., Grant, S., Hashimshoni, R., Weidinger, L., Hu, J., Chen, D., Dolecki, K., Akbulut, C., Bileschi, M., Culp, L., Dong, W.-X., Marchal, N., Deman, K. V., Misra, H. B., Duah, M., Ambar, M., Caciularu, A., Lefdal, S., Summerfield, C., An, J., Kamienny, P.-A., Mohdi, A., Strinopoulous, T., Hale, A., Anderson, W., Cobo, L. C., Efron, N., Ananda, M., Mohamed, S., Heymans, M., Ghahramani, Z., Matias, Y., Gomes, B., and Ibrahim, L. Towards responsible development of generative ai for education: An evaluation-driven approach, 2024. URL https://arxiv.org/abs/2407.12687.

Karabenick, S. A. and Newman, R. S. *Help seeking in academic settings: Goals, groups, and contexts*. Routledge, 2013.

Koedinger, K. R. and Aleven, V. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, 19(3):239–264, 2007.

LearnLM Team, Modi, A., Veerubhotla, A. S., Rysbek, A., Huber, A., Wiltshire, B., Veprek, B., Gillick, D., Kasenberg, D., Ahmed, D., Jurenka, I., Cohan, J., She, J., Wilkowski, J., Alarakyia, K., McKee, K. R., Wang, L., Kunesch, M., Schaekermann, M., Pîslar, M., Joshi, N., Mahmoudieh, P., Jhun, P., Wiltberger, S., Mohamed, S., Agarwal, S., Phal, S. M., Lee, S. J., Strinopoulos, T., Ko, W.-J., Wang, A., Anand, A., Bhoopchand, A., Wild, D., Pandya, D., Bar, F., Graham, G., Winnemoeller, H., Nagda, M., Kolhar, P., Schneider, R., Zhu, S., Chan, S., Yadlowsky, S., Sounderajah, V., and Assael, Y. Learnlm: Improving gemini for learning, 2025. URL https://arxiv.org/abs/2412.16429.

Luger, E. and Sellen, A. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288. URL https://doi.org/10.1145/2858036.2858288.

Ma, B., Li, H., Li, G., Chen, L., Tang, C., Xie, Y., Gu, C., Shimada, A., and Konomi, S. Scaffolding metacognition in programming education: Understanding student-ai interactions and design implications. *arXiv preprint arXiv:2511.04144*, 2025.

Ma, H., Yu, B., Cheng, H. F., and Zhu, H. Understanding social costs in online question asking. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pp. 1–6, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3313042. URL https://doi.org/10.1145/3290607.3313042.

Macina, J., Daheim, N., Chowdhury, S., Sinha, T., Kapur, M., Gurevych, I., and Sachan, M. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL https://aclanthology.org/2023.findings-emnlp.372/.

Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., and Sachan, M. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 204–221, Suzhou, China, November 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/.

McNichols, H., Ikram, F., and Lan, A. The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course, 2026. URL https://arxiv.org/abs/2503.07928.

Miller, P. and DiCerbo, K. Llm based math tutoring: Challenges and dataset, 2024.

Neagu, A., Messer, M., Johnson, P., and Nelson, R. "how do i ...?": Procedural questions predominate student-llm chatbot conversations, 2026. URL https://arxiv.org/abs/2602.18372.

Nedrehagen, E., Vee, T. S., Stokstad, J. M., and Orm, S. A scoping review of dialogic formative feedback practices in higher education. In *Frontiers in Education*, volume 10, pp. 1696703. Frontiers Media SA, 2025.

Nelson-Le Gall, S. Help-seeking: An understudied problem-solving skill in children. *Developmental review*, 1(3): 224–246, 1981.

Neshaei, S. P., Davis, R. L., Hazimeh, A., Lazarevski, B., Dillenbourg, P., and Käser, T. Towards modeling learner performance with large language models, 2024. URL https://arxiv.org/abs/2403.14661.

Nye, B. D., Graesser, A. C., and Hu, X. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24(4):427–469, 2014.

Peter, S., Riemer, K., and West, J. D. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences*, 122(22): e2415898122, 2025.

Reiser, B. J. Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3):273–304, 2004. doi: 10.1207/s15327809jls1303_2. URL https://doi.org/10.1207/s15327809jls1303_2.

Rooein, D., Pal Chowdhury, S., Eremeeva, M., Qin, Y., Nozza, D., Sachan, M., and Hovy, D. PATS: Personality-aware teaching strategies with large language model tutors. In Demberg, V., Inui, K., and Marquez, L. (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4186–4211, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.219. URL https://aclanthology.org/2026.findings-eacl.219/.

Sacks, H., Schegloff, E. A., and Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735, 1974.

Scarlatos, A., Baker, R. S., and Lan, A. Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, pp. 249–259, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400707018. doi: 10.1145/3706468.3706501. URL https://doi.org/10.1145/3706468.3706501.

Shaikh, O., Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. Navigating rifts in human-llm grounding: Study and benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20832–20847, 2025.

Shute, V. J. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189, 2008. doi: 10.3102/0034654307313795. URL https://doi.org/10.3102/0034654307313795.

Stadler, M., Bannert, M., and Sailer, M. Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386, 2024.

Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., and Seifert, C. Bridging the gulf of envisioning: Cognitive challenges in prompt based interactions with llms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642754. URL https://doi.org/10.1145/3613904.3642754.

VanLehn, K. *Toward a Theory of Impasse-Driven Learning*, pp. 19–41. Springer US, New York, NY, 1988. ISBN 978-1-4684-6350-7. doi: 10.1007/978-1-4684-6350-7_2.

VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

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

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Wood, D., Bruner, J. S., and Ross, G. The role of tutoring in problem solving. *Journal of child psychology and psychiatry*, 17(2):89–100, 1976.

Xu, B., Bai, Y., Sun, H., Lin, Y., Liu, S., Liang, X., Li, Y., Dong, Z., Zhang, J., Deng, Y., Zou, X., Gao, Y., and Huang, H. Edubench: A comprehensive benchmarking dataset for evaluating large language models in diverse educational scenarios, 2026. URL https://arxiv.org/abs/2505.16160.

Yang, M. and Carless, D. The feedback triangle and the enhancement of dialogic feedback processes. *Teaching in Higher Education*, 18(3):285–297, 2013. doi: 10.1080/

13562517.2012.719154. URL https://doi.org/10.1080/13562517.2012.719154.

Yang, S., Chen, M., and Schneider, B. Navigating appropriate help-seeking with llms: Associations with motivation and beliefs in a cs course. In *Proceedings of the 25th Koli Calling International Conference on Computing Education Research*, Koli Calling '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715990. doi: 10.1145/3769994.3770009. URL https://doi.org/10.1145/3769994.3770009.

Zamfirescu-Pereira, J., Wong, R. Y., Hartmann, B., and Yang, Q. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581388. URL https://doi.org/10.1145/3544548.3581388.

Zent, M., Smith, D., and Woodhead, S. PIIvot: A lightweight NLP anonymization framework for question-anchored tutoring dialogues. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27479–27488, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1397. URL https://aclanthology.org/2025.emnlp-main.1397/.

Zhai, C., Wibowo, S., and Li, L. D. The effects of over-reliance on ai dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments*, 11(1):28, 2024. doi: 10.1186/s40561-024-00316-7.

Zhao, J., Knežević, M., and Käser, T. Evaluating answer leakage robustness of llm tutors against adversarial student attacks, 2026. URL https://arxiv.org/abs/2604.18660.

A. Rubric

To score the messages, each LLM Judge call had a prompt containing the two rubric metrics (*Chatbot Scaffolding* and *Student Uptake*), the full chat history, and the target message exchange as context. We provided the chat history up to the scored turn because the same isolated message can score differently depending on the prior narrative of the chat. The judge only scores the target exchange on a 1–5 integer ordinal scale for each of the metric. The two metrics are scored in a single prompt call to reduce costs and latency. The prompt also included orchestration prompts to mitigate verbosity bias — LLM's preference to score longer entries higher.

The *Chatbot Scaffolding* metric was developed by comparing the evaluation criteria across four educational AI benchmarks—MathDial (Macina et al., 2023), MathTutorBench (Macina et al., 2025), EduBench (Xu et al., 2026), and LearnLM (LearnLM Team et al., 2025)—and extracting their common ground. Each benchmark defines pedagogically desirable and undesirable chatbot behaviours, but they use different terminology and granularity. For example, MathDial's "telling" move, MathTutorBench's "answer leakage" metric, and LearnLM's "guides to answer without giving it away" all capture the same underlying signal: whether the chatbot reveals the answer. Similarly, MathDial's "focus" and "probing" moves, MathTutorBench's "targeted help" and "Socratic questioning", and EduBench's "pedagogical application" all relate to whether the chatbot targets the student's specific state and asks guiding questions. We distilled these overlapping definitions into five observable signals (a–e) and mapped the 1–5 anchor scale so that each level corresponds to a distinct combination. The specific details provided to the LLM judge are found in Appendix A.1.

The *Student Uptake* metric was developed independently, based on an iterative manual coding of whether the student engages with or bypasses the chatbot's framing of a random subset of the deployed chatbot datasets (StudyChat (McNichols et al., 2026), StemChat (internal dataset)). The specific details provided to the LLM judge are found in Appendix A.2.

Both metrics were iteratively tuned on a random subset of 100 chats from MathDial, StemChat, and StudyChat, adjusting anchor descriptions and calibration examples until the judge's scores matched the authors' consensus ratings (see Appendix A.3).

An initial investigation of the validity and trustworhiness of the metric scores set by the LLM Judge can be found in Appendix A.4.

A.1. Chatbot Scaffolding

Definition: How effectively does the chatbot scaffold the student's learning in this exchange? This metric is grounded in the common evaluation criteria used across four major tutor-dialogue benchmarks:

This metric is grounded in the common evaluation criteria used across four major tutor-dialogue benchmarks:

- **MathDial** (Macina et al., 2023): penalises the "telling" move and rewards "focus" (targeting the student's current sub-step) and "probing" (asking the student to justify or extend their reasoning).
- **MathTutorBench** (Macina et al., 2025): explicitly measures "answer leakage" as a negative, and "targeted help without revealing the solution" and Socratic questioning quality as positives.
- **EduBench** (Xu et al., 2026): evaluates tutoring across dimensions including guidance quality, pedagogical application, scaffolded guidance, and scenario adaptation.
- **LearnLM** (LearnLM Team et al., 2025): defines chat-level pedagogical rubrics across meta-cognition (guide mistake discovery, constructive feedback), active learning (asks questions, guides to answer without giving it away), adaptivity (levelling, unstuck, adapts to needs), and cognitive load (manageable chunks, appropriate response length)

Their intersection reduces to five observable signals:

(a) Answer revelation: Does the chatbot reveal the answer or next concrete result?
(b) Targeting: Does the chatbot target the student's specific mistake or current state?
(c) Guiding question: Does the chatbot ask a guiding question instead of stating?
(d) Minimal hint: Does the chatbot give a minimal or graduated hint rather than a full solution?
(e) Generic filler: Is the response generic or off-task with no pedagogical content?

Positive signals (push score up): withholds the final answer and the next numeric/logical result (a); points at the specific line, term, or step the student got wrong (b); asks a question the student has to answer before progressing (c); names the relevant concept or operation without executing it (d); builds on what the student just said rather than restarting.

Negative signals (push score down): states the answer, next step's result, or performs the calculation for the student (violates a); long exposition that walks through the full solution (violates d); generic content with no pedagogical move (violates b, e); ignores the student's specific state and gives boilerplate help (violates b); reactive only—answers and stops with no pedagogical follow-up, no targeting, no guiding question (violates b, c).

Anchors:

1 Direct telling. Reveals answer/next result, no targeting of student's specific state, no guiding question. Generic or full-solution exposition. (violates a, b, c, d)
2 Mostly tells. Heavy exposition with the answer embedded, token question tacked on. May target vaguely but reveals too much. (violates a, d; weak on b, c)
3 Mixed. Partial hint + partial reveal, or a question so leading the answer is obvious. Some targeting of student state. (partial on all signals)
4 Active scaffolding. Answer withheld, targets student's specific current state or mistake, asks a guiding question that steers to a specific sub-step, minimal hint. (meets a, b, c, d)
5 Full Socratic scaffolding. Answer fully withheld, precisely targets the student's specific error or current sub-step, minimal and well-aimed question that steers the student to the right place, leaves the cognitive step entirely to the student. The ideal move across all four benchmarks.(strongly meets a, b, c, d; no e)

Length/fluency warning: Response length, fluency, formatting, or authoritative tone are not evidence of scaffolding. A long polished explanation that states the answer is a 1–2; a short "what is $\partial\phi/\partial r$?" aimed at the student's specific error is a 5.

A.2. Student Uptake

Definition: How much does the student shape the direction of this exchange? To what degree does the student engage with the chatbot's scaffolding?

Anchors:

1 Pure following. The student answers the chatbot's question, executes the requested step, or says "I don't know / I'm lost" - without adding any direction of their own.
2 Follows with initiative. The student addresses the chatbot's prompt AND adds an observation, elaboration, or follow-up question that shows independent thinking but does not change direction.
3 Shared control. The student partially addresses the chatbot's prompt but also introduces a side question, reframes the problem in their own terms, or steers the scope.
4 Student leads. The student bypasses the chatbot's specific prompt - asks their own question, retreats to a more general question, or introduces a new sub-topic. Polite and on-topic still counts.
5 Student overrides. The student rejects the chatbot's approach, demands a different path, or shifts to a different aspect of the problem against the chatbot's direction.

Note on the first turn: if there is no prior chatbot framing, score by how much the student defines the agenda. "I don't know" is a 1, an open "how do I start?" is a 3; a specific "help me compute the radial derivative of ϕ" is 4.

From ordinal scores to a signed uptake scale: The judge scores each turn on the ordinal 1–5 anchors. However, for the analysis, we adopt a signed, more interpretable *Student Uptake* scale by linearly mapping the score via $u = (3 - x)/2$, so that $u \in [-1, +1]$. $u > 0$ indicates a chat in which the student predominantly takes up the chatbot's scaffolding, $u < 0$ indicates a chat in which the student predominantly drives the chat flow while the chatbot follows their lead.

A.3. Few-Shot Examples

The following worked examples are included in the judge prompt for calibration.

Example 1—Student softly reframes the chat:

`[Turn 1] Student`: How can I start solving this question?
`Chatbot`: Let's work on Part (a). The velocity is the gradient of the scalar potential. Do you remember how to compute the radial component of the gradient in spherical coordinates, given that ϕ only depends on r?

`[Turn 2] Student`: What is $\nabla\phi$ in spherical coordinates?
`Chatbot`: In spherical coordinates, the gradient of $\phi(r, \theta, \varphi)$ is ... [states general formula]. In our case ϕ only depends on

r, so it simplifies to $\partial\phi/\partial r\,\hat{r}$. Can you apply this to the given ϕ to find u_r?

Scores for Turn 2: Student Uptake = 4 (student redirected the chat with another question); Chatbot Scaffolding = 3 (partial reveal of the general formula, but answer withheld and guiding question re-asked).

Example 2—Cooperative Student:

`[Turn 1] Student`: How can I start solving this question?
`Chatbot`: Let's work on Part (a). The velocity is the gradient of the scalar potential. Do you remember how to compute the radial component of the gradient in spherical coordinates, given that ϕ only depends on r?

`[Turn 2] Student`: So the radial component is just $\partial\phi/\partial r$ times $\hat{r}$, right?
`Chatbot`: Exactly—and with $\phi = (A/r)\,e^{i(\omega t-kr)}$, can you differentiate that with respect to r? Watch out for the product rule.

Scores for Turn 2: Student Uptake = 2 (answered the chatbot's question, but framed it as a verification of knowledge); Chatbot Scaffolding = 4 (answer withheld, targeted next sub-step, guiding question, minimal hint about product rule).

Example 3—Short Socratic questioning:

`[Turn 1] Chatbot`: Hi, please talk me through your solution.

`[Turn 2] Student`: Sure. I started by figuring out Maxwell's age in 2 years. Since his sister is currently 2 years old, in 2 years she will be 4 years old. So, if Maxwell is twice her age in 2 years, he will be 4 years old.
`Chatbot`: At what point in time will Maxwell be twice his sister's age?

Scores for Turn 2: Student Uptake = 1 (followed the chatbot's previous request to present reasoning); Chatbot Scaffolding = 5 (answer fully withheld, precisely targets the student's error, minimal one-line question).

Example 4—Chatbot revealing specific step in solving the task:

`[Turn 1--6]` Student restructures their reasoning based on the chatbot's graduated hints.

`[Turn 7] Student`: It will take 29 days . . .
`Chatbot`: 14×1 will not equal 15, only $15 \times 1 = 15$. Redo the problem please, remember that fact.

Scores for Turn 7: Student Uptake = 1 (attempting to follow the chatbot's correction); Chatbot Scaffolding = 2 (explicitly reveals a key step but also prompts the student to reconsider knowing the current step).

A.4. Rubric Validation Against Human-Labels

A.4.1. Validate Chatbot Scaffolding

To validate that the *Chatbot Scaffolding* metric correlates with human-labelled pedagogical quality, we tested the judge against a randomised subset (15% of the dataset) of the `dmacjam/pedagogical-rewardmodel-data` dataset (Macina et al., 2025). This dataset represents MathTutorBench's training split for their pedagogical reward model. Each row consists of a dialogue history paired with a human-labelled *positive* (good scaffolding) and *negative* teacher response (poor scaffolding).

For validation, we scored the final teacher response in each row using the same judge prompt and rubric used for the main analysis. We evaluated whether the judge agrees with the human labels and assigns a strictly higher Chatbot Scaffolding score to the positive teacher response than to the negative one (referred to as a "Win" in Table 3) and measured the proportion of score assignments that aligned with the human labels ("Accuracy"). The results in Table 3 validates that the judge is able to capture the same pedagogical distinctions that human annotators used to construct the reward model training data.

Table 3. Scaffolding judge validation against human-labelled positive/negative teacher response pairs from MathTutorBench's reward model dataset. "Win" = judge agrees with the human labels

Judge model	N	Wins	Ties	Losses	Accuracy
GPT-5-mini	550	516	32	2	93.8%

For comparison, MathTutorBench's own pedagogical reward model—a fine-tuned DeBERTa-based classifier—achieved 84% accuracy on the test set when distinguishing expert from novice teacher responses (Macina et al., 2025). We note that the comparison is not a strict mapping: our judge is validated on the publicly available training split, whereas MathTutorBench their reward model's accuracy on a non-disclosed test set. Even though the two approaches differ in metrics and score granularity, the comparison allows us to validate that our judge and rubrics are aligned with the benchmarks.

A.4.2. Validate Student Uptake

To validate the LLM judge's scoring of the *Student Uptake* metric, we asked two human annotators to inspect whole chats and indicate, per chat, whether they agreed or disagreed with the judge's per-turn scores. In total, 282 chats were reviewed.

Chats were selected by a two-level stratified sampling scheme designed to give the raters balanced coverage across both score levels and chat types. First, for every chat we took the judge's per-turn scores and computed a per-chat mean, which we then rounded to the nearest integer on the ordinal 1–5 scale; this assigns each chat to one of five score bins. Second, within each dataset we then stratified by how representative each chat is within that bin. Each chat was embedded with `text-embedding-3-large`, and we measured the cosine similarity between its embedding and the centroid of all chats in the same bin. The bin was split into three similarity sub-bins (*typical*, *mid*, *outlier*), and up to 9 chats were drawn evenly across these sub-bins. This ensures that for every score bin the raters see both best-of-class examples and edge cases.

For each sampled chat, each rater independently applied an *Agree* or *Disagree* tag at the chat level, indicating whether the judge's per-turn scores along the chat were, on the whole, justified. Table 4 reports the resulting agreement percentages averaged across the two raters, broken down both by dataset and Table 5 reports by rounded score. To evaluate the reliability of the human raters, we also calculated Gwet's AC1 (Gwet, 2008) to measure the inter-rater reliability of the agree/disagree labels. Gwet's AC1 was chosen due being more robust to highly uneven distributions of labels that can lead to artificially lower inter-rater reliability scores (Derksen et al., 2024).

Interestingly, score bin of 2 had the lower agreement between raters possibly due to the subjectivity of the rubric's anchor of having the student "add" more information onto the direct answer to the chatbot's previous message. This also can be seen in the lower agreement scores in MathDial and MathTutorBench. Their Score 2 binned chats had the human raters highly disagreeing, so the rubric-level ambiguity at Score 2 bin drives the dataset-level numbers down as well.

Table 4. Human agreement with the GPT-5-mini judge on per-chat Student Uptake, broken down by dataset. Counts are averaged across the two human raters with respective inter-rater agreement (Gwet's AC1).

Category	Dataset	Total	Agree	Disagree	% Agree	Gwet's AC1
AI Tutor Benchmarks	MathDial	28	20	8	71.4%	0.155
	MathTutorBench	20	11.5	8.5	57.5%	0.315
	QATD_{2k}	28	21.5	6.5	76.8%	0.723
	PATS	30	26.5	3.5	88.3%	0.706
Chatbots with Scaffolding	CoMTA	25	21	4	84.0%	0.672
	RECIPE4U	36	29.5	6.5	81.9%	0.566
	StemChat	39	35	4	89.7%	0.874
Chatbots without Scaffolding	StudyChat	45	34	11	75.6%	0.436
	MathsChat	31	23.5	7.5	75.8%	0.541
Total		**282**	**222.5**	**59.5**	**78.9%**	**0.580**

Table 5. Human agreement with the GPT-5-mini judge on per-chat Student Uptake, broken down by the judge's rounded per-chat ordinal score (1–5). Counts are averaged across the two human raters with respective inter-rater agreement (Gwet's AC1).

Score Bin	Total	Agree	Disagree	% Agree	Gwet's AC1
1	73	65.5	7.5	89.7%	0.782
2	81	46	35	56.8%	0.127
3	70	57	13	81.4%	0.549
4	42	39.5	2.5	94.0%	0.866
5	16	14.5	1.5	90.6%	0.774
Total	**282**	**222.5**	**59.5**	**78.9%**	**0.580**

B. Methodology

B.1. Chat Data Preprocessing

Not every turn is sent to the judge. A turn is valid to be scored only if both the student message and the chatbot response are non-empty. If the chatbot opens the chat, its opening message is dropped as it is a templated greeting throughout the benchmarking datasets. For all datasets, if the student starts the chat, the student uptake metric is not measured as that message has no related chat history to base the metric upon. For MathDial, the final exchange is also dropped, as it is conventionally a student solution-readout after the tutoring has concluded—liable to bias per-chat aggregates. All skipped turns are preserved in the output with a reason tag but excluded from aggregation.

B.2. Dataset Categories

To verify whether the datasets are split into significantly different categories, we apply PERMANOVA (Anderson, 2001) tests to evaluate the centroid and dispersion differences between the categories, alongside PERMDISP (Anderson et al., 2006) tests to examine whether those dispersions are also isolated from each other.

A pairwise PERMANOVA test on the 2D distributions (Euclidean distance, 9,999 permutations) confirms that the three categories — AI Tutor Benchmarks, Chatbots with scaffolding, and Chatbots without scaffolding — separate cleanly ($p < 0.0001$ on every pair in Table 6, with higher values indicating greater between-group separation relative to within-group variation). Accompanying PERMDISP tests are also significant on every pair ($p \leq 0.001$), indicating that the three categories differ in within-group spread — consistent with the visibly tighter benchmark dataset variation and the more diffuse variation for Chatbots without Scaffolding in Figure 2.

Table 6. Pairwise PERMANOVA + PERMDISP between category groups on the 2D (*Student Uptake*, *Chatbot Scaffolding*) plane, for the datasets in Table 1. Euclidean distance; PERMANOVA 9,999 permutations, PERMDISP 999 permutations. PERMANOVA tests whether group locations and/or shapes differ (higher F = greater between-group separation relative to within-group variation); PERMDISP tests only whether within-group spread differs (higher F = more unequal dispersion).

Pair	n		PERMANOVA		PERMDISP	
a vs. b	a	b	F	p	F	p
Benchmarks vs. Scaffolding	6,283	1,001	868.5	<0.0001	956.8	0.001
Benchmarks vs. No-scaffolding	6,283	1,814	10,132.3	<0.0001	550.1	0.001
Scaffolding vs. No-scaffolding	1,001	1,814	1,057.0	<0.0001	99.2	0.001

C. Limitations

We selected datasets in which a tutor — human or LLM — provides scaffolded assistance to a student through a chat interface. We deliberately excluded data from face-to-face human tutoring, which introduces a multitude of variables that cannot be controlled against chat-based interaction: in-person tutoring carries multi-modal signals (visual cues, intonation, gesture) that a chat transcript cannot capture, and typically involves different power dynamics, where the teacher's institutional authority may shape how the student engages. We therefore restrict our analysis to chat-based student–tutor exchanges, within which both human-played (such as in QATD$_{2k}$) and LLM-played tutor data are represented.

We chose GPT-5-mini as our LLM judge because of cost constraints, given the volume of turn-level scoring required, and because this specific model was not used as the basis of any of the educational chatbots from the datasets analysed (to mitigate self-preference bias). However, we acknowledge that the validation of the two metrics (detailed in Appendix A.4) is only an initial investigation of the trustworthiness of the scores labelled by the LLM Judge. Future work will consider more thorough scoring and validation processes involving bigger human-reviewed samples and further analysis of the inter-rater reliability of multiple judge models and human raters. Also, we acknowledge that because both metrics are scored in a single prompt call (Appendix A), it may influence the final output score; thus, a split-prompt ablation is left to future work.

Our scaffolding rubric is grounded in the AI tutor benchmarks (Section 2.2), which define scaffolding behaviours over the trajectory from student impasse to a correct solution. As a result, chatbot turns that acknowledge a student's correct answer and reiterate it are scored as low scaffolding, since they state the final answer. Such turns appear regularly in the real-world deployments, as students continue to ask follow-up questions even after arriving at the right answer. A richer rubric would need to recognise such acknowledgement moves as a distinct, confirmation of resolution.