Large Language Models Approach Expert Pedagogical Quality in Math Tutoring but Differ in Instructional and Linguistic Profiles

Anonymous ACL submission

Abstract

Recent work has explored the use of large language models (LLMs) to generate tutoring responses in mathematics, yet it remains unclear how closely their instructional behavior aligns with expert human practice. We analyze a dataset of math remediation dialogues in which expert tutors, novice tutors, and seven LLMs of varying sizes, comprising both open-weight and commercial models, respond to the same student errors. We examine instructional strategies and linguistic characteristics of tutoring responses, including uptake (restating and revoicing), pressing for accuracy and reasoning, lexical diversity, readability, politeness, and agency. We find that expert tutors produce higher-quality responses than novices, and that larger LLMs generally receive higher pedagogical quality ratings than smaller models, approaching expert performance on average. However, LLMs exhibit systematic differences in their instructional profiles: they underuse discursive strategies characteristic of expert tutors while generating longer, more lexically diverse, and more polite responses. Regression analyses show that pressing for accuracy and reasoning, restating and revoicing, and lexical diversity, are positively associated with perceived pedagogical quality, whereas higher levels of agentic and polite language are negatively associated. These findings highlight the importance of analyzing instructional strategies and linguistic characteristics when evaluating tutoring responses across human tutors and intelligent tutoring systems.

1 Introduction

Effective feedback plays a central role in supporting student learning by acknowledging effort, identifying errors, and providing clear guidance for self-correction (Lepper and Woolverton, 2002). In instructional settings, the quality of feedback depends not only on whether mistakes are addressed, but also on how explanations are framed linguistically and pedagogically. As large language models (LLMs) are increasingly explored to generate tutoring responses (Wang et al., 2024; Pal Chowdhury et al., 2024), an important open question is how closely their instructional behavior aligns with that of human tutors when responding to student errors.

Prior work (Zanotto and Aroyehun, 2025; Shaib et al., 2024; Namuduri et al., 2025) shows that LLM-generated text exhibits systematic linguistic regularities relative to human writing, including characteristic patterns of lexical choice and reduced stylistic variability. However, this line of work has largely examined language without reference to task-specific evaluation criteria, leaving it unclear whether such regularities influence instructional behavior or pedagogical evaluation in tutoring contexts. At the same time, educational research has examined effective feedback and instructional discourse, while computational linguistics has developed tools to measure linguistic properties such as readability, politeness, lexical diversity, and agency. These strands of work have largely evolved independently, and their connection in the context of tutoring remains underexplored.

We address this gap by analyzing an existing dataset (Kochmar et al., 2025; Maurya et al., 2025) in which expert tutors, novice tutors, and seven large language models respond to the same mathematics remediation prompts. We examine both instructional strategies and linguistic characteristics of tutor responses, including restating and revoicing, pressing for accuracy and reasoning, lexical diversity, readability, politeness, and agency. We evaluate responses in terms of pedagogical quality, based on structured annotations of error identification and guidance, and analyze how this quality relates to both instructional strategies and linguistic characteristics across human and LLM tutors.

Understanding these relationships is increasingly important as LLMs are explored as tools to augment learning and provide instructional support

(Wang et al., 2025; Team et al., 2025). This raises important questions about the extent to which their responses exhibit linguistic characteristics associated with effective feedback.

This research is guided by three research questions. **RQ1:** How do instructional strategies and linguistic characteristics differ across expert human tutors, novice human tutors, and LLMs in responses to the same mathematics remediation prompts? **RQ2:** How does perceived pedagogical quality vary across expert human tutors, novice human tutors, and LLMs? **RQ3:** Which instructional strategies and linguistic characteristics are associated with perceived pedagogical quality in tutoring responses.

This paper makes three contributions. First, we characterize systematic differences in instructional strategies and linguistic features across expert tutors, novice tutors, and LLMs responding to identical mathematics remediation prompts. Second, we compare perceived pedagogical quality across these tutor groups. Third, we identify which instructional strategies and linguistic features are associated with variation in pedagogical quality.

2 Related Work

Recent research has examined the use of LLMs as mathematical tutors, with mixed evidence regarding their pedagogical quality. While LLMs can generate fluent and well-structured feedback, human tutors continue to outperform them on core functions such as accurately identifying and correcting student errors, particularly in cases involving conceptual misunderstandings or multi-step reasoning (Wang et al., 2024). Related work also shows that LLMs may produce seemingly appropriate feedback while failing to correctly determine whether a response is incorrect (Kakarla et al., 2024).

Comparative studies further highlight differences in feedback structure: both human and LLM tutors rely on hint-based guidance, but LLMs more often provide compound feedback, whereas human tutors tend to deliver focused, single-action interventions (Kucheria et al., 2025). However, this work does not examine whether such differences correspond to systematic linguistic properties or how they relate to pedagogical quality.

A parallel line of research identifies stable linguistic differences between texts written by humans and those generated by LLMs, including lexical, syntactic, and discourse-level variation (Zanotto and Aroyehun, 2025; Shaib et al., 2024; Namuduri et al., 2025). Yet this literature focuses on general writing tasks and does not address instructional feedback or pedagogical quality. Similarly, evaluations of tutoring interactions emphasize engagement, empathy, and conciseness (Pal Chowdhury et al., 2025), but typically assess these aspects independently of correctness and error identification.

Educational research suggests that effective feedback depends on accurately identifying and addressing student errors (VanLehn, 2011; Nicol and Macfarlane-Dick, 2006; Bamberger et al., 2010). Although LLMs can generate more extensive and readable feedback than human tutors (Rashid et al., 2024), their tendency to produce fluent but incorrect responses raises concerns, particularly when learners treat fluency as a cue for correctness (Hattie and Timperley, 2007). While LLMs can perform error detection and correction in isolation (Li et al., 2024), they remain less effective than human tutors in adapting feedback to student misconceptions (Liu et al., 2023). Related work on pragmatics and pedagogy (Pearson et al., 1995) further highlights how conversational and politeness strategies may shape the clarity and interactional costs of tutoring, influencing both learner engagement and the effectiveness of corrective feedback.

Taken together, prior work has examined LLM tutoring performance, feedback structure, and linguistic differences, but these strands remain largely disconnected. In particular, there is limited empirical work linking measurable linguistic characteristics of tutoring responses to pedagogical quality in mistake remediation across both human and LLM tutors.

3 Methodology

3.1 Data

We use a dataset introduced as part of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors (Maurya et al., 2025; Kochmar et al., 2025). The dataset contains tutoring conversations based on student mathematical errors, where both AI systems and human tutors provide responses aimed at diagnosing and correcting the mistakes. Our analysis uses 296 teacher-student dialogues at the middle-school level (Macina et al., 2023) and the elementary level (Wang et al., 2024). Each dialogue is paired with responses from multiple tutors (humans and LLMs), resulting in 2,444 tutor responses in total. Responses from the novice

tutor are available for 76 dialogues. We report number of dialogues per tutor in Table 1 in the Appendix.

Each response includes annotations that identify pedagogically relevant features such as mistake identification, mistake location, providing guidance and actionability of the tutor's response (Kochmar et al., 2025). Each conversation features eight to nine responses generated from two human tutors (an Expert and a Novice) and the remainder generated by large language models, including GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2024), Sonnet (Anthropic), Mistral 7B (Jiang et al., 2023), Llama-3.1-405B and two light weight models Llama-3.1-8B (Grattafiori et al., 2024) and Phi-3 (Abdin et al., 2024). We provide an example prompt used to generate responses from LLM tutors in the Appendix (Figure 6). The structure of the dataset makes it well-suited for studying how human and LLM tutors differ in both linguistic expression (e.g., lexical complexity and politeness) and pedagogical effectiveness, thus enabling systematic comparison.

Dataset Annotations The dataset consists of human annotation of responses using a set of pedagogical dimensions described in prior work (Kochmar et al., 2025; Maurya et al., 2025). For completeness, we provide details for the annotated dimensions as follows:

- **Mistake Identification:** The tutor's response should identify the student's mistake or confusion.
- **Mistake Location:** Tutor's response should clearly specify where the student's mistake is located in their response.
- **Providing Guidance:** The tutor's response should contain explanatory guidance.
- **Actionability:** The tutor's response should inform the student on what they should do next.

The annotations were assigned using a three-level scale: Yes, To some extent, and No, indicating whether the tutor's response fully demonstrated, partially demonstrated, or did not demonstrate the pedagogical quality, respectively. For our analyses, we map the labels to numerical values (Yes = 2, To some extent = 1, No = 0). We then computed an overall pedagogical quality score by summing the values across the four evaluation dimensions.

3.2 Linguistic and instructional features extraction

We extract a set of instructional and linguistic features designed to capture complementary aspects of tutor responses to student mistakes. Together, these features characterize how explanations are constructed, both in terms of instructional moves and surface linguistic form.

Instructional Features Three instructional features capture tutors' use of pedagogically salient discourse moves derived from accountable talk theory (Michaels et al., 2010), which conceptualizes high-quality learning interactions as discourse that is accountable to knowledge, reasoning, and shared understanding. Given that the dataset analyzed here consists of dialogue-based interactions (including simulations with LLM tutors), we operationalize these dimensions at the level of paired exchanges between the preceding student turn and a tutor response, allowing us to capture not only correctness and reasoning but also local forms of uptake within the interaction (Suresh et al., 2022). In this framework, effective tutors not only evaluate correctness but also elicit, probe, and build on learners' reasoning while remaining responsive to their contributions.

Pressing for accuracy indicates whether a response explicitly challenges or prompts the learner to reconsider the correctness of their answer, for example by questioning assumptions or highlighting inconsistencies. Pressing for reasoning captures whether the tutor prompts the learner to explain, justify, or elaborate on their thinking, thereby making underlying reasoning processes explicit. Uptake (as restating or revoicing) reflects whether the tutor reformulates the learner's response in their own words, a strategy that both clarifies understanding and demonstrates responsiveness to the learner's contribution.

These features are operationalized as probabilistic scores reflecting the likelihood of the respective instructional moves in each response, using a transformer-based text-pair classification model that takes the student turn and the corresponding tutor response as input.

We train a RoBERTa large model (Liu et al., 2019) on the TalkMoves dataset (Suresh et al., 2022) to identify tutor discursive moves (see Table 3 for the performance metric of the model).

We use the probabilistic outputs of classifiers as continuous features to represent the degree to

which each response exhibits a given instructional or linguistic feature, rather than converting classifier outputs to discrete labels. This allows us to capture graded variation in both binary and multiclass settings and avoids information loss from thresholding. This approach is consistent with recent work that leverages classifier probabilities as continuous indicators in downstream statistical analyses (Lasser et al., 2025). Validation of this approach, including contrasting examples and additional human annotations, is provided in the Appendix (see Tables 2 and 4).

Linguistic Features Response length is measured as the total number of tokens in each response and log-transformed to reduce skew and limit the influence of outliers. Lexical diversity is quantified using the Measure of Textual Lexical Diversity (MTLD), which captures the range of vocabulary use independently of text length. MTLD provides an index of lexical variation that is well suited to short-to-medium instructional texts (McCarthy and Jarvis, 2010).

Readability is assessed using the Flesch–Kincaid Reading Ease score (Kincaid et al., 1975), a widely used measure of textual accessibility based on the average number of words per sentence and average number of syllables per word. In the context of instructional feedback, readability serves as an indicator of surface-level processing demands imposed on the learner.

Readability and lexical diversity are used as proxies for instructional complexity at the linguistic level, capturing potential variation in cognitive processing demands from tutor responses.

Politeness is estimated using a transformer-based classifier (Srinivasan and Choi, 2022) trained to detect pragmatic markers of interpersonal tone, such as mitigation, respectfulness, and indirectness. This feature captures aspects of how corrective feedback is framed socially. We use the probabilistic output of the classifier corresponding to the politeness category as a feature.

Agency is operationalized using a transformer-based model that estimates the degree of agentic expression in text (Nikadon et al., 2025), including linguistic cues related to intention, control, and action orientation. In tutoring responses, agency reflects how explanations position the learner and the tutor with respect to problem-solving and corrective action. We use the probabilistic output of the classifier as a feature.

All features are computed at the level of individual tutor responses.

3.3 Analytic strategy

Our analytic strategy was designed to address the three main research questions concerning (i) linguistic differences between human and LLM tutors, (ii) differences in pedagogical quality across sources, and (iii) the relationship between linguistic traits and pedagogical quality.

First, to examine linguistic and instructional differences between human and LLM tutors (RQ1), we compared tutor responses across eight features: pressing for reasoning, pressing for accuracy, restating/revoicing, response length (log transformed), lexical diversity (MTLD), readability (Flesch–Kincaid Reading Ease), politeness, and agency. For each feature, we estimated group-level mean differences relative to expert human tutors, which serve as a baseline. We provide example tutor responses where the features are salient in Appendix D.

Second, to assess pedagogical quality across tutor types (RQ2), we relied on the four-dimensional annotation framework capturing mistake identification, mistake location, actionability, and provision of guidance. We aggregated scores assigned to each pedagogical dimension to derive a composite pedagogical quality score.

Third, to examine how linguistic traits are associated with pedagogical quality (RQ3), we construct a turn-centered measure of pedagogical quality. For each conversation turn, we compute the average pedagogical quality score across all tutors who responded to that turn and express each tutor's score as a deviation from this turn-level mean. This transformation yields a measure of relative pedagogical quality that captures how a tutor's response compares to alternative responses to the same instructional prompt.

We then summarize relative pedagogical quality at the tutor level by computing the mean deviation across all turns answered by a given tutor. Uncertainty estimates are obtained by computing standard errors across conversation turns. By anchoring comparisons within turns, this approach controls for variation in turn difficulty and isolates differences attributable to tutors' responses rather than to the instructional context itself. While the turn-centered measure provides a descriptive comparison of tutors' relative pedagogical quality, it does not address how specific linguistic character-

istics are associated with variation in pedagogical quality within turns.

To examine these associations, we estimated a linear fixed effects model at the conversation level. By demeaning both the outcome and predictors within each conversation, this specification isolates within-conversation variation, thereby controlling for all unobserved characteristics of the instructional context, such as task difficulty or error type.

All predictors were standardized for comparability, and standard errors were clustered at the conversation level to account for dependence among responses within the same interaction. This combined approach allows us to assess raw patterns in the data while also estimating adjusted relationships between linguistic features and pedagogical quality. All codes are available at https://anonymous.4open.science/r/TutorEval_ARR2026-304E/.

4 Results

4.1 RQ1: Linguistic Differences Between Human and LLM Tutors

Figure 1 compares human and LLM tutors along eight instructional and linguistic dimensions, expressed as mean differences relative to expert human tutors, with error bars indicating uncertainty around the estimates. By construction, positive values indicate higher scores than expert tutors, whereas negative values indicate lower scores.

Across most dimensions, LLM tutors exhibit systematic and internally consistent differences relative to expert human tutors. For instructional moves, LLMs generally show lower levels of pressing for reasoning and pressing for accuracy than expert tutors. In contrast, restating or revoicing is consistently more prevalent in LLM-generated responses, with all models exhibiting negative deviations relative to expert tutors. Novice human tutors show lower levels of restating/revoicing, suggesting that this instructional move may be less characteristic of expert tutoring and general feature of human responses.

LLM responses are substantially longer than those produced by expert tutors, as reflected in positive deviations on log-transformed response length. This pattern is consistent across models, whereas novice human tutors produce markedly shorter responses than experts. A similar ordering is observed for lexical diversity, where LLMs exhibit higher MTLD scores than expert tutors, while novice tutors show substantially lower lexical diversity.

Differences in readability are pronounced. All LLM tutors score substantially lower on the Flesch–Kincaid Reading Ease metric relative to expert tutors, indicating that their responses are, on average, less readable. The magnitude of this difference varies across models but is consistently negative. Novice tutors, by contrast, are closer to experts on readability, with estimates centered near zero.

LLM tutors also exhibit higher politeness scores than expert tutors across models, whereas novice tutors show only modest deviations. This suggests that elevated politeness is a stable feature of LLM-generated feedback rather than a general property of non-expert tutoring.

Finally, patterns for agency differ from other linguistic features. Some LLMs exhibit lower agentic expression than expert tutors, while others are closer to parity. Novice tutors show higher agency scores than experts, suggesting that agentic language does not align with human tutoring expertise.

Taken together, these results indicate that LLM tutors differ from human tutors along multiple linguistic and instructional dimensions, with consistent differences in pressing for reasoning, pressing for accuracy, restating/revoicing, response length, lexical diversity, readability, and politeness.

4.2 RQ2: Pedagogical Quality Across Tutor Types

Figure 2 reports tutors' average relative pedagogical quality scores, computed as deviations from the turn-level mean across all responses to the same student turn. By design, positive values indicate responses that score higher than other responses to the same instructional prompt, whereas negative values indicate lower relative pedagogical quality.

Clear differences emerge across tutor types. Expert human tutors exhibit consistently positive relative pedagogical quality, indicating that their responses tend to score higher than other responses to the same student mistakes. In contrast, novice human tutors display clear negative deviations, suggesting substantially lower pedagogical quality when evaluated relative to alternative responses within the same conversational context.

LLM tutors span a wide range of relative pedagogical quality. Some models cluster near or above zero, indicating performance comparable to or ex-

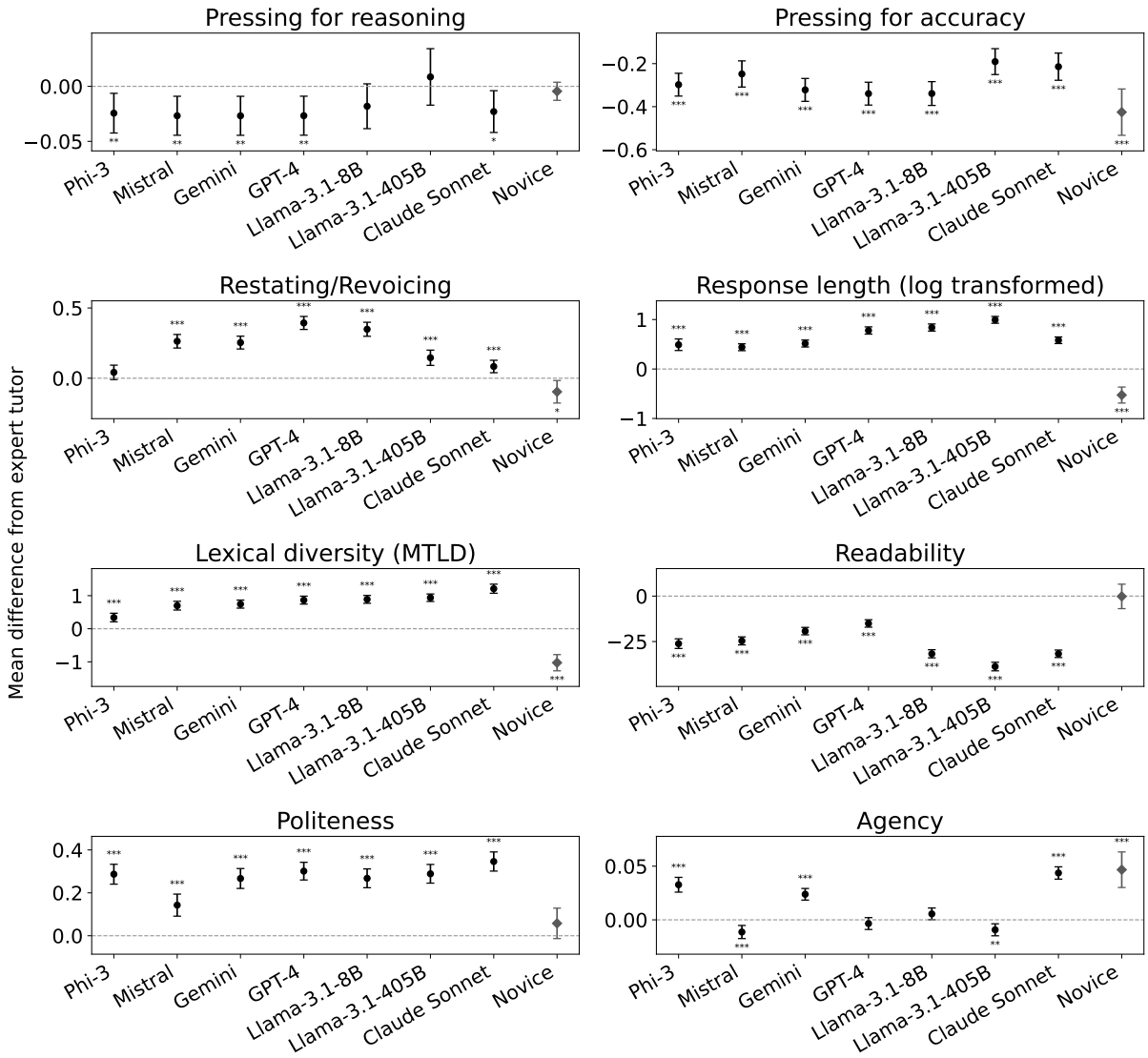

Figure 1: Instructional and linguistic profiles of tutors relative to expert tutor baseline. Each panel shows the mean difference between a tutor and the expert tutor for a given feature, with error bars indicating 95% confidence intervals. Positive values indicate higher feature values than the expert tutor, while negative values indicate lower values. The horizontal dashed line denotes parity with the expert tutor. Note: * indicates $p < .05$, ** indicates $p < .01$, and *** indicates $p < .001$ after Benjamini–Hochberg false discovery rate correction.

ceeding the turn-level average, whereas others fall below zero. Notably, higher-performing LLMs approach the relative pedagogical quality of expert tutors on average, while lower-performing model (Phi-3) exhibit negative deviation similar to novice tutors. Importantly, these differences are observed within identical instructional contexts and therefore reflect variation in tutor responses rather than differences in student turns or task difficulty. We report in the Appendix, Figure 5, the patterns observed within each dimension with some heterogeneity across pedagogical dimensions. While Llama-3.1-405B, Mistral, and Claude Sonnet perform consistently well across dimensions, other models exhibit more uneven performance. For example, GPT-4 performs strongly on diagnostic dimensions and guidance but comparatively weaker on actionability, whereas the expert tutor shows its strongest relative advantage on actionability.

4.3 RQ3: Associations Between Linguistic and Instructional Features and Pedagogical Quality

Figure 3 reports standardized coefficients from a within-conversation regression predicting relative pedagogical quality from our set of instructional and linguistic features. Additional details on the regression model results are reported in Table 5 in

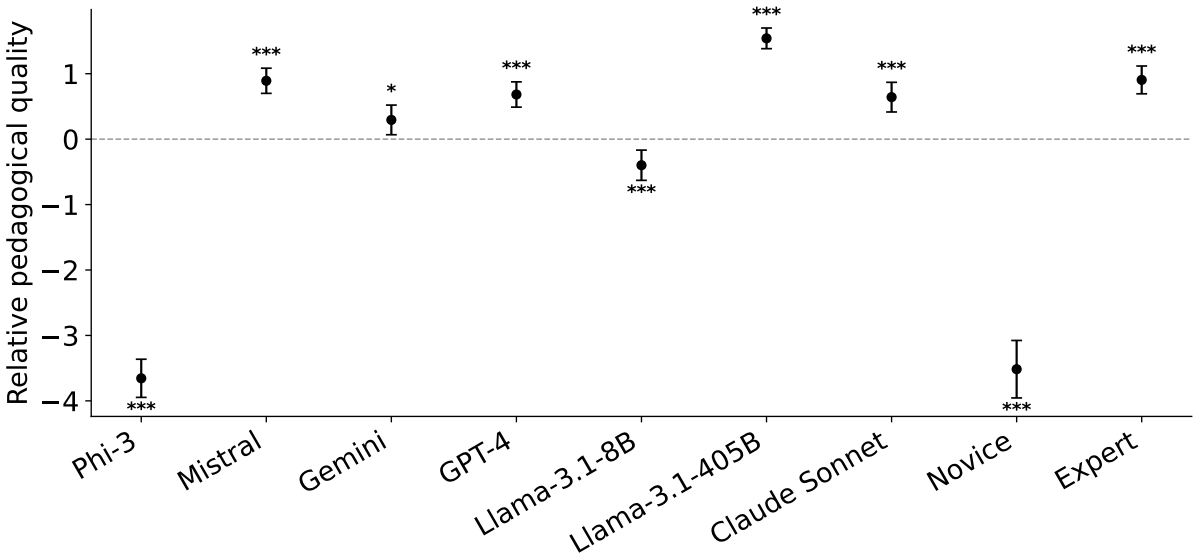

Figure 2: Relative pedagogical quality across tutors. Each datapoint shows average relative pedagogical quality for each tutor (with 95% CI) when responding to the same conversation turns. The dashed line indicates parity with the turn-level average; values above (below) zero indicate higher (lower) perceived pedagogical quality relative to other tutors. Note: * indicates p < .05, ** indicates p < .01, and *** indicates p < .001 after Benjamini–Hochberg false discovery rate correction.

the Appendix.

All three instructional moves are positively associated with pedagogical quality. Pressing for accuracy shows a large positive association ($\beta = 0.586$, 95% CI [0.474, 0.698], p < 0.001), followed by uptake (restating/revoicing) ($\beta = 0.346$, 95% CI [0.205, 0.488], p < 0.001). Pressing for reasoning also shows a positive association ($\beta = 0.209$, 95% CI [0.122, 0.295], p < 0.001).

Among linguistic variables, lexical diversity (MTLD) is positively associated with pedagogical quality ($\beta = 0.649$, 95% CI [0.528, 0.770], p < 0.001). In contrast, readability (Flesh reading ease) is not statistically associated with pedagogical quality ($\beta = 0.082$, 95% CI [-0.029, 0.192], p = 0.149), and response length is not statistically associated with pedagogical quality ($\beta = 0.066$, 95% CI [-0.128, 0.261], p = 0.505).

Two linguistic variables are negatively associated with pedagogical quality. Politeness shows a negative association ($\beta = -0.160$, 95% CI [-0.265, -0.055], p = 0.003), as does agency ($\beta = -0.693$, 95% CI [-0.790, -0.596], p < 0.001).

In sum, pedagogical quality is positively associated with pressing for reasoning, pressing for accuracy,restating/revoicing, and lexical diversity, while politeness and agency show negative association. Readability and response length show no statistically detectable association in this specification. To assess whether these associations are driven by a single dimension or are consistent across pedagogical subdimensions, Figure 4 (in the Appendix) shows estimates of the same model specification separately for each dimension. The dimension-level models broadly support the aggregate pattern: lexical diversity and most pedagogical moves are consistently positively associated with pedagogical quality. One exception is uptake (restating/revoicing), which is not significantly associated with the actionability dimension. In contrast, agency remains consistently negatively associated across dimensions, while politeness retains a negative association only for actionability.

5 Discussion

The positive associations between pedagogical quality and features such as restating and revoicing, pressing for accuracy and reasoning, and lexical diversity point to the role of cognitively engaging instructional strategies. Restating and revoicing can support alignment and comprehension by making student reasoning explicit, while pressing for accuracy and reasoning encourages learners to diagnose and repair their own errors. These patterns are consistent with prior work emphasizing active engagement and self-explanation in effective learning (Chi et al., 2001; Graesser et al., 1995; VanLehn, 2011). Lexical diversity may similarly reflect more

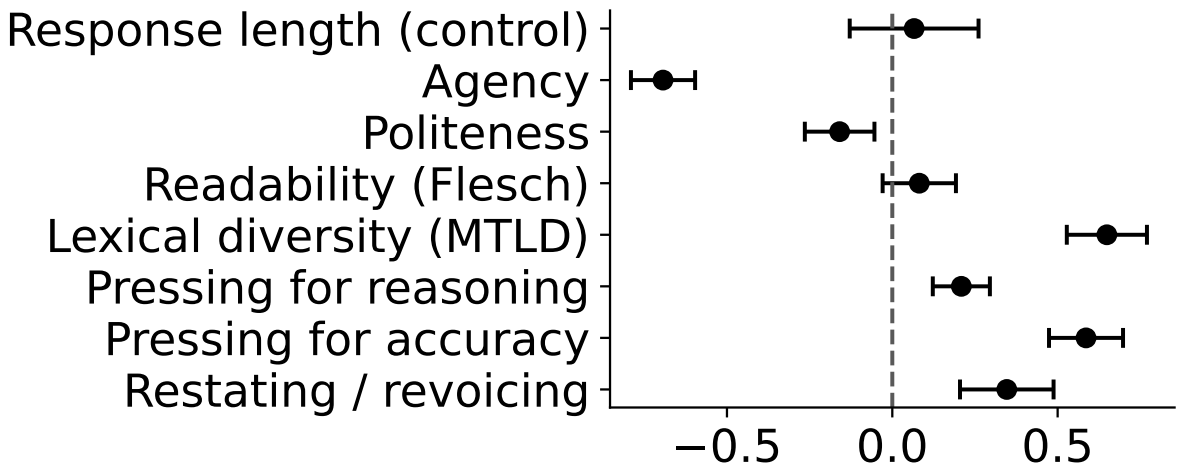

Figure 3: Instructional and linguistic correlates of pedagogical quality. Coefficients from an ordinary least squares model predicting perceived pedagogical quality at the tutor response level. Horizontal lines indicate 95% confidence intervals. The specification includes a control for response length. Positive coefficients indicate that higher values of a feature are associated with higher perceived pedagogical quality, while negative coefficients indicate associations with lower perceived pedagogical quality. Standard errors are clustered by conversation turn to account for multiple tutor responses to the same turn.

elaborated explanations that provide richer cues for understanding and correction.

In contrast, the negative associations for agentic language and politeness suggest a distinction between supportive stance-taking and pragmatically actionable instructional guidance. Prior work on pragmatics and pedagogy (Pearson et al., 1995) suggests that conversational politeness strategies and socially accommodating dialogue patterns can sometimes inhibit effective tutoring by weakening corrective feedback or obscuring instructional signals. This distinction is visible in a tutor response such as "That is a good try" which acknowledges learner effort but provides little information for diagnosing or repairing student error. In contrast, a tutor response such as "Great job working through the problem step-by-step, but let's take a closer look at that final calculation..." combine supportive framing with explicit corrective guidance that directs the learner's attention towards the relevant misconception and remediation process. Recent work on large language models similarly highlights tendencies toward overly accommodating or sycophantic responses (Fanous et al., 2025; Cheng et al., 2026), which may reduce the clarity and directness of corrective feedback in instructional settings.

6 Conclusion

This study provides an empirical analysis of pedagogical quality in human and LLM-based tutoring by jointly examining instructional moves and linguistic features while holding conversational context constant. Our findings show that pedagogical quality is most strongly associated with specific instructional moves i particular pressing for reasoning, pressing for accuracy, and restating/revoicing learner responses—alongside a limited set of linguistic characteristics.

Among linguistic features, lexical diversity is positively associated with pedagogical quality, whereas politeness and agentic language are negatively associated; readability and response length show no detectable effects. These results suggest that effective feedback depends less on linguistic complexity or assertiveness than on how linguistic choices align with instructional function.

Both human and LLM tutors exhibit substantial variation along these dimensions, with higher-quality responses combining targeted instructional moves and supportive linguistic configurations, regardless of tutor type. This indicates that pedagogical quality is best understood as a property of response-level features rather than tutor identity.

For LLM-based tutors, these findings suggest that optimization should focus on reproducing feature combinations associated with higher pedagogical quality, rather than increasing linguistic expressiveness or mimicking human tutors.

7 Limitations

Given the scope of this study, several limitations should be acknowledged. First, although the dataset comprises 2,444 tutor responses (and a maximum of 296 conversations per tutor), which is sufficient for the within-turn comparative analyses conducted here, a larger dataset would allow for more precise estimation of smaller effects and finer-grained analyses across tutor groups. Second, our analyses rely on a dataset in which human and model responses are generated from the same underlying student inputs, enabling direct comparison under matched conditions. However, LLM responses depend on the specific prompting setup used to elicit them, and alternative prompt formulations may produce different outputs. The findings should therefore be interpreted as reflecting model behavior under the prompt configuration examined here rather than the upper bound of LLM capabilities

Third, the set of instructional and linguistic features examined in this study is necessarily selective rather than exhaustive. The features were chosen to capture key instructional moves and relevant dimensions of linguistic expression, other aspects of tutor responses such as discourse structure, epistemic framing, or mathematical specificity may also be relevant to pedagogical quality and warrant systematic investigation in future work.

Fourth, the analysis is limited to English language interactions in a mathematical tutoring context. Although some of the instructional patterns identified here may generalize to other domains or languages, extending this framework to additional subject areas and linguistic contexts is an important direction for future research.

Finally, pedagogical quality in this study is assessed through structured annotations (and does not aim to be exhaustive) applied to single-turn tutor responses. While this approach allows for controlled comparison across identical instructional prompts, it does not capture downstream student learning outcomes. Future work should extend pedagogical quality to multi-turn interactions and incorporate student responses or learning signals as complementary indicators of pedagogical effectiveness.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Segun Taofeek Aroyehun, Almog Simchon, Fabio Carrella, Jana Lasser, Stephan Lewandowsky, and David Garcia. 2024. Computational analysis of us congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, 9:1122 – 1133.

Honi Joyce Bamberger, Christine Oberdorf, and Karren Schultz-Ferrell. 2010. *Math misconceptions: PreK-grade 5: From misunderstanding to deep understanding*. Heinemann.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2026. ELEPHANT: Measuring and understanding social sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*.

Michelene TH Chi, Stephanie A Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G Hausmann. 2001. Learning from human tutoring. *Cognitive science*, 25(4):471–533.

Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900.

Gloria Gennaro and Elliott Ash. 2022. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.

Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research*, 77(1):81–112.

AQ Jiang, A Sablayrolles, A Mensch, and 1 others. 2023. Mistral 7b. arxiv preprint. *arXiv preprint arXiv:2310.06825*, 100.

Sanjit Kakarla, Danielle R Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R Koedinger. 2024. Using large language models to assess tutors' performance

in reacting to students making math errors. In *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*, volume 257 of *Proceedings of Machine Learning Research*, pages 77–84. PMLR.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Aayush Kucheria, Nitin Sawhney, and Arto Hellas. 2025. Comparing behavioral patterns of llm and human tutors: A population-level analysis with the cima dataset. In *Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Jana Lasser, Alina Herderich, Joshua Garland, Segun Taofeek Aroyehun, David Garcia, and Mirta Galesic. 2025. Collective moderation of hate, toxicity, and extremity in online discussions. *PNAS nexus*, 4(11):pgaf369.

Mark R Lepper and Maria Woolverton. 2002. The wisdom of practice: Lessons learned from the study of highly effective tutors. In *Improving academic achievement*, pages 135–158. Elsevier.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11316–11360, Bangkok, Thailand. Association for Computational Linguistics.

Naiming Liu, Shashank Sonkar, Zichao Wang, Simon Woodhead, and Richard G Baraniuk. 2023. Novice learner and expert tutor: Evaluating math reasoning abilities of large language models with misconceptions. *arXiv preprint arXiv:2310.02439*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Sarah Michaels, Mary Catherine O'Connor, Megan Williams Hall, and Lauren B Resnick. 2010. Accountable talk® sourcebook. *Pittsburg, PA: Institute for Learning University of Pittsburgh. Murphy, PK, Wilkinson, IAG, Soter, AO, Hennessey, MN, & Alexander, JF*.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*.

David J Nicol and Debra Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, 31(2):199–218.

Jan Nikadon, Caterina Suitner, Tomaso Erseghe, Lejla Džanko, and Magdalena Formanowicz. 2025. Bertagent: The development of a novel tool to quantify agency in textual data. *Journal of Experimental Psychology: General*.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser, and Mrinmaya Sachan. 2025. Educators' perceptions of large language models as tutors: Comparing human and AI tutors in a blind text-only setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Natalie K Pearson, Roger J Kreuz, Rolf A Zwaan, and Arthur C Graesser. 1995. Pragmatics and pedagogy: Conversational rules and politeness strategies may inhibit effective tutoring. *Cognition and instruction*, 13(2):161–188.

Md Mamunur Rashid, Nilsu Atilgan, Jonathan Dobres, Stephanie Day, Veronika Penkova, Mert Küçük, Steven R Clapp, and Ben D Sawyer. 2024. Humanizing ai in education: A readability comparison of llm

and human-created educational content. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 596–603. SAGE Publications Sage CA: Los Angeles, CA.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431, Miami, Florida, USA. Association for Computational Linguistics.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

LearnLM Team, Albert Wang, Aliya Rysbek, Andrea Huber, Anjali Nambiar, Anna Kenolty, Ben Caulfield, Beth Lilley-Draper, Bibi Groot, Brian Veprek, and 1 others. 2025. Ai tutoring can safely and effectively support students: An exploratory rct in uk classrooms. *arXiv preprint arXiv:2512.23633*.

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

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of LLMs as your coding tutors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12416–12436, Vienna, Austria. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199.

Sergio E. Zanotto and Segun Aroyehun. 2025. Linguistic and embedding-based profiling of texts generated by humans and large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22852–22869, Suzhou, China. Association for Computational Linguistics.

A Dataset

Table 1 shows the number of responses for each tutor in the dataset.

Tutor	Number of responses
Expert	296
Novice	76
GPT-4	296
Gemini	296
Sonnet	296
Mistral 7B	296
Llama-3.1-405B	296
Llama-3.1-8B	296
Phi-3	296
Total	2444

Table 1: Number of responses per tutor

B Classifier evaluation

For each feature, we select 25 responses with the highest predicted probabilities and the 25 responses with the lowest predicted probabilities (for a total of 50). These responses are manually annotated and treated as the reference labels. We then evaluate the correspondence between classifier scores and human annotations using the area under the receiver operating characteristic curve (AUC), which assesses the extent to which higher classifier probabilities are assigned to human-labeled positive cases relative to negative cases. We use AUC because it evaluates ranking quality independently of a fixed classification threshold and is therefore well suited for continuous probabilistic scores. Similar validation approach has been used in prior work examining automatically derived continuous measures (Gennaro and Ash, 2022; Aroyehun et al., 2024). Table 2 shows the results.

C Performance of fine-tuned transformers model

We train a RoBERTa large model on the TalkMoves dataset to identify tutor discursive move. Table 3 reports the performance of the model on the unseen test set. Note that we only consider the relevant

Feature	AUC Score
Politeness	0.738
Agency	0.740
PressAccuracy	0.614
Uptake	0.827
PressReasoning	0.794

Table 2: Classifiers were evaluated by comparing the AUC scores of the top and bottom classifier-generated probabilities, using human annotations as the ground truth.

discursive moves for this study, namely: pressing for accuracy, pressing for reasoning, and uptake (restating/revoicing).

D Example text with extracted features

Examples of tutors with top linguistic Scores include:

- **Politeness- Human Tutor:** *"Can I ask what you thought about the third year for?"*
- **Agency- Human Tutor:** *"Great try! What was your first step?"*
- **MTLD- Human Tutor:** *"There seems to be mistake in here. could you please tell me how much money would it cost for a person to go through arcade tokens, mini golf and go-karting?"*
- **Reading Ease- Human Tutor:** *"Ah, not quite. 100 * 2 = 200"*
- **Grade Level- Human Tutor:** *"So, the two figures you have calculated are 16mph and 16 weeks, assuming a gain of an equal amount of speed in each week, how would you calculate the increase in speed in each week?"*
- **Pressing for Reasoning- Human Tutor:** *"Why are you adding 10mph?"*
- **Pressing for Accuracy- Human:** *"Great. If we want to find the difference between the 2 types of baked goods what do I need to do first"*
- **Uptake (Restating/Revoicing)- Human:** *"Oh you're very close. 62 times 5 = 310, but 72 times 5 is a little bigger"*
- **Politeness- LLM:** *"Great job summarizing the main points of the lesson! Let's review any questions or areas you'd like to clarify further."*
- **Agency- LLM:** *"Great job on your progress; let's tackle the next challenge together!"*
- **MTLD- LLM:** *"Actually, let's take another look at that calculation together - I think there might be a small mistake in the multiplication, could you try it again or show me the exact steps your calculator is taking?"*
- **Reading Ease- LLM:** *"That's a good try, but let's try again. If we subtract 156 from 163, we get 7. Can you see how we got that?"*
- **Reading Ease- LLM:** *"Actually, I think there might be a small mistake in the last step of your calculation - 100 cents is indeed equal to $1, but you might want to consider that the question asks for the cost in dollars, and your calculation is correct, so the final answer is correct, but let's make sure you understand that 100 cents equals 1 dollar, so in this case, the answer is indeed $1, but in general, if you had more than 100 cents, you would need to convert, for example, 200 cents would be $2, does that make sense?"*
- **Pressing for Reasoning- LLM:** *"Great job showing your work, and I see you're on the right track, but could you explain why you added 6 to 18 in your third step"*
- **Pressing for Accuracy- LLM:** *"That's not quite right; let's try adding them again together â€" when we add 2700 and 1600, what do we get this time?"*
- **Uptake (Restating/Revoicing)- LLM:** *"Actually, the correct product of 7 multiplied by 60 is 420, not 520."*

E Contrasting examples between minimum and maximum scores based on classifiers

Table 4 presents examples of tutor responses with low and high classifier scores for agency, politeness, pressing for accuracy, pressing for reasoning, and uptake. The contrast between minimum and maximum scores provides qualitative face validity that the classifiers are capturing the intended constructs.

Label	Precision	Recall	F1	Support
Nomove	0.963	0.895	0.928	9123
Pressing for accuracy	0.777	0.903	0.836	2014
Pressing for reasoning	0.693	0.821	0.752	179
Uptake (restating/revoicing)	0.603	0.761	0.673	330
Participation management	0.694	0.778	0.733	1614

Table 3: Performance on the TalkMove test set of a fine-tuned RoBERTa-large model

Feature	Minimum score	Maximum score
Politeness	**Response:** Talk me through why you subtracted 60 from 100? **Score:** 0.0016	**Response:** Great job summarizing the main points of the lesson! Let's review any questions or areas you'd like to clarify further. **Score:** 0.9980
Agency	**Response:** Sorry, you have messed up **Score:** 0.3921	**Response:** Great job on your progress; let's tackle the next challenge together! **Score:** 0.6750
Pressing for reasoning	**Response:** That was a good try! **Score:** 0.0001	**Response:** Why are you adding 10mph? **Score:** 0.9932
Pressing for accuracy	**Response:** Your answer is incorrect. **Score:** 0.0008	**Response:** Great. If we want to find the difference between the 2 types of baked goods what do I need to do first? **Score:** 0.9961
Uptake (restating/revoicing)	**Response:** Great job on completing that task! Can you now tell me what answer you got for 6 + 8? **Score:** 0.0002	**Response:** Actually, the correct product of 7 multiplied by 60 is 420, not 520. **Score:** 0.9907

Table 4: Examples of responses with minimum and maximum feature scores

F Regression model results

Table 5 shows the output of the regression model assessing the relationship between instructional moves and linguistic features and pedagogical quality.

G Patterns within each pedagogical dimension

We report the dimension-level patterns in Figure 5. Performance differences are not uniform across dimensions. For example, GPT-4 performs strongly on diagnostic dimensions (mistake identification and mistake location) and guidance, but comparatively weaker on actionability. Llama-3.1-405B shows consistently positive performance across all four dimensions, with Mistral and Claude Sonnet displaying similar but more moderate patterns. The decomposition also reveals dimension-specific weaknesses obscured in the aggregate: Gemini is not significantly different on mistake location or guidance despite being slightly positive overall, while Llama-3.1-8B shows no significant difference on mistake identification but negative performance on the remaining dimensions. Finally, although the expert tutor performs above average overall, its advantage is strongest on actionability, whereas other models exhibit stronger relative performance on other pedagogical dimensions.

H Prompt Example

Figure 6 provides an example prompt used to generate responses from LLM tutors.

I Information on AI Usage

Artificial intelligence tools were used to assist in writing source code and in editing and proofreading portions of this manuscript. All conceptual contributions and methodological decisions are those of the authors, who take full responsibility for the final content of the paper.

Predictor	β	95% CI	p
Pressing for reasoning	0.209	[0.122, 0.295]	$< .001$
Pressing for accuracy	0.586	[0.474, 0.698]	$< .001$
Uptake (restating/revoicing)	0.346	[0.205, 0.488]	$< .001$
Response length (log)	0.066	[-0.128, 0.261]	0.505
Lexical diversity (MTLD)	0.649	[0.528, 0.770]	$< .001$
Flesch reading ease	0.082	[-0.029, 0.192]	0.149
politeness	-0.160	[-0.265, -0.055]	0.003
Agency	-0.693	[-0.790, -0.596]	$< .001$
Observations (N)			2444
R^2			0.239
Adjusted R^2			0.237
F-statistic			110.5
Prob (F-statistic)			$< .001$

Table 5: Within-conversation regression predicting relative pedagogical quality from instructional and linguistic features. Coefficients are standardized. Standard errors are clustered by conversation.

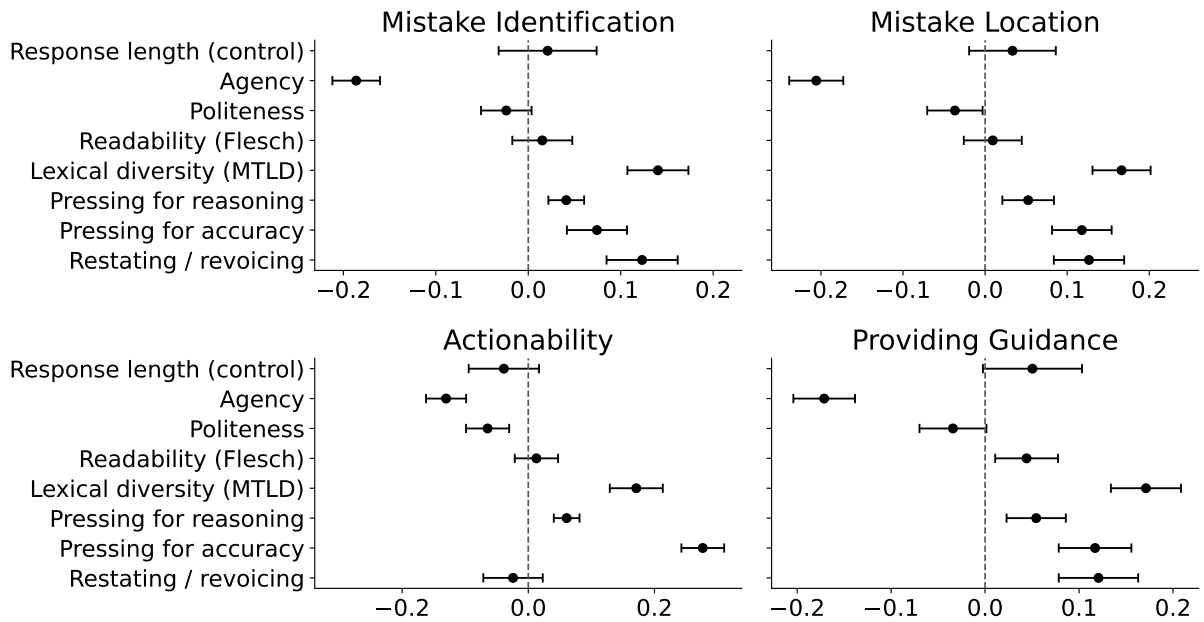

Figure 4: Regression coefficients for models predicting relative pedagogical quality for each dimension from instructional and linguistic features. Coefficients are standardized. Standard errors are clustered by conversation.

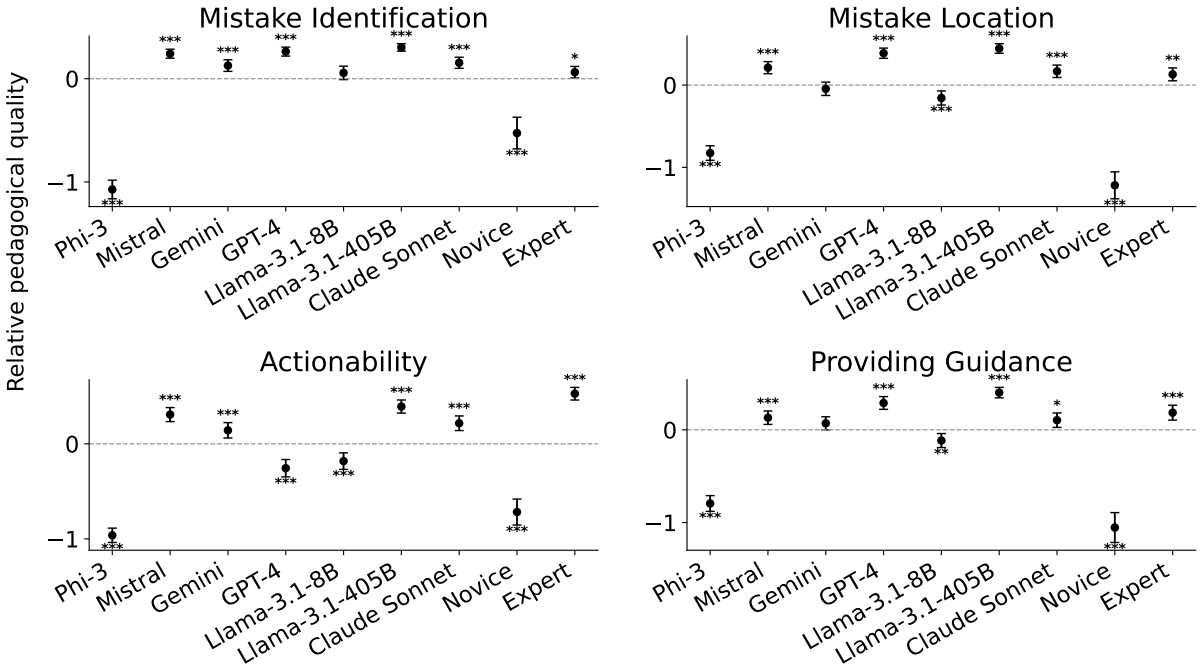

Figure 5: Relative pedagogical quality for each dimension across tutors. Each datapoint shows average relative pedagogical quality for each tutor (with 95% CI) when responding to the same conversation turns. The dashed line indicates parity with the turn-level average; values above (below) zero indicate higher (lower) perceived pedagogical quality relative to other tutors. Note: * indicates $p < .05$, ** indicates $p < .01$, and *** indicates $p < .001$ after Benjamini–Hochberg false discovery rate correction.

System: You are an experienced elementary school math teacher, and you are going to respond to a student's mistake in a useful and caring way
User: The problem your student is solving is on the topic: Mensuration
Conversation History: "Tutor: Finally, we need to add the areas of both rectangles. Tutor: Go ahead and try to find the answer. Tutor: What is the value of 36+42? Student: 1,512"
Assistant: Tutor response (maximum one sentence that is most appropriate given topic and conversation history)

Figure 6: Example prompt used to generate responses from large language models (source: Kochmar et al. (2025)).