IALab UC at BEA 2025 Shared Task: LLM-Powered Expert Pedagogical Feature Extraction

Sofía Correa Busquets[1, 2, 3], Valentina Córdova Véliz[1,3], Jorge Baier[1, 2, 3],

[1]Pontificia Universidad Católica de Chile, [2]Millenium Institute Foundational Research on Data,

[3]National Center for Artificial Intelligence

{sbcorrea,avcordova,jbaier}@uc.cl

Abstract

As AI's presence in educational environments grows, it becomes critical to evaluate how its feedback may impact students' learning processes. Pedagogical theory, with decades of effort into understanding how human instructors give good-quality feedback to students, may provide a rich source of insight into feedback automation. In this paper, we propose a novel architecture based on pedagogical-theory feature extraction from the conversation history and tutor response to predict pedagogical guidance on MRBench. Such features are based on Brookhart's canonical work in pedagogical theory, and extracted by prompting the language model LearnLM. The features are then used to train a random-forest classifier to predict the Track 3: Pedagogical Guidance of the BEA 2025 shared task. Our approach ranked 8$^{\text{th}}$ in the dimension's leaderboard with a test Macro F1-score of ~ 0.54. Our work provides some evidence in support of using pedagogical theory qualitative factors treated separately to provide clearer guidelines on how to improve low-scoring intelligent tutoring systems. Finally, we observed several inconsistencies between pedagogical theory and MRBench's inherent relaxation of the tutoring problem implied in evaluating on a single-conversation basis, calling for the development of more elaborate measures which consider student profiles to serve as true heuristics of AI tutors' usefulness.

1 Introduction

As part of the AI revolution, AI tutors will gain a growing role in education. Their use, however, should be preceded by rigorous evaluation, as omitting this step would be as unthinkable as hiring untrained teachers. To contribute to the development of evaluation standards for AI tutors, this paper describes an approach to automatically classify certain aspects of pedagogical ability on the Mistake Remediation Benchmark (MRBench) dataset of grade-school math tutoring chats (Maurya et al., 2025a). The dataset contains annotations for the dimensions of identifying that the student has made a mistake, correctly individualizing what that mistake was, providing the student with relevant and helpful guidance, and cueing the student on how to follow the conversation. Of these, our approach attempts to classify whether feedback did, did not, or did to some extent, provide pedagogical guidance (PG) on the, Track 3: Pedagogical Guidance of the BEA 2025 shared task (Kochmar et al., 2025).

PG as an object of study is richly explored in the theory of pedagogy. For instance, the area of math didactics has studied phenomena such as students' capacity to grasp concepts progressing from the concrete, to the pictorial, to the abstract (Bruner, 1966); how to develop an academic math discourse to support understanding (Chapin et al., 2009); and best practices for orchestrating productive student discussions (Smith and Stein, 2011). Also, assessment theory compiles frameworks on how to construct feedback as a powerful tool to improve student understanding and performance (Brookhart, 2008; Tunstall and Gipps, 1996). Our approach attempts to transfer knowledge from pedagogical theory by proposing a set of engineered features for PG classification strongly based on Brookhart's work. With these features in hand, we propose a two-phase classification process. In the first phase, we use an LLM to query the text, which includes the conversation history between the student and the tutor, for the presence, or lack thereof, of our features in the tutor's feedback. In the second phase, we use a random-forest classifier which is given a binary vector representing the output of the previous phase and attempts to predict the PG dimension.

2 Related Work

MRBench's dimensions on which to assess the pedagogical ability of AI tutors result from the distil-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1187–1193

lation of a body of previous work in NLP addressing ITS evaluation (Tack and Piech, 2022; Macina et al., 2023; Daheim et al., 2024; Wang et al., 2024). Tack and Piech (2022), in their "AI Teacher Test" evaluated the dialogic pedagogical ability of certain LLMs in a mathematics-domain educational dialogue from the dimensions of whether they speak like a teacher, understand a student, and help a student. Specifically in math mistake remediation in the tutoring context, Macina et al. (2023) dimensions included coherence, correctness, and equitable tutoring. In the same context, Daheim et al. (2024) create the dimensions of targetedness, correctness, and actionability. Finally, and also within said context, Wang et al. (2024) put forth usefulness, care, and humanness. Maurya et al. (2025b) compile MRBench to address this need for a unified evaluation framework, and the present Shared Task is proposed as a challenge because all the aforementioned work is not, as of yet, fully independent from manual evaluation.

3 Preliminaries

To determine qualities that make feedback effective, the pedagogical perspective generally follows Brookhart's (2008) four-dimension framework: content, specificity, timing and audience. Rather than assigning intrinsic value to hints, explanations or other information the tutor might provide, these dimensions promote that feedback's potential depends on every point that it communicates complying with certain characteristics. For example, when amending any student misconceptions (content-focus), to unambiguously identify the misconception (specificity-clarity), feedback should explicitly distinguish it from what the student has understood correctly (content-valence). The same would be true for the offering of procedural guidance (content-focus): a hint about the right direction may confuse the student into undoing correct steps taken. Furthermore, the clarity of all the aforementioned depends on the student's level of prior knowledge (audience-individual), which in this case we may approximate as the school year. This framework thus offers a theoretically grounded approach to tackle the interdependence of feedback dimensions in function of the ultimate goal: helping the student.

4 Methodology

To distill a set of features from pedagogical theory, we first asked the virtual assistant Claude (Anthropic, 2024) to create a feedback checklist from the key takeaways of seminal books on assessment and math didactics (Chapin et al., 2009; Smith and Stein, 2011; Brookhart, 2008; Tunstall and Gipps, 1996). Second, we merged redundant points together and discarded factors that were outside the scope of MRBench: anything that required knowing the student personally, communicating nonverbally and/or interacting in a classroom environment. The few remaining factors came chiefly from Brookhart (2008). Third, we stress-tested these for what we anticipated as possible AI tutor failures and edited accordingly by hand. For example, we added "accurately and specifically" at the beginning of Claude's sentence "identify what was done correctly before addressing the error". Then, each quality was separated into its own feature (identifies / identifies accurately / identifies specifically), so that binary tags on these features would be as informative as possible. Fourth, we phrased each feature as a yes-or-no question to prompt LearnLM (Team et al., 2024). Finally, we performed prompt engineering on the questions using a subset of 20 random tutor responses. The full resulting list of questions is available in Appendix A.

5 Architecture

Our proposed architecture, shown in Figure 1, is composed of two models working sequentially: a feature extractor, and a classifier.

First, features are extracted by prompting LearnLM (Team et al., 2024), a domain-specific Gemini fine-tuning, currently in experimental phase. We chose this model because of its expert training on tutoring data and pedagogical theory sources. For each feature, the conversation history is concatenated to the tutor response and a yes-or-no question representing the feature (see Appendix B), to which the model is prompted to respond with a binary 0/1 tag. Since preliminary tests yielded no relevant difference resulting from temperature variation, the model's hyperparameters were left at their default values. The full feature extraction prompt is in Appendix B.

To accommodate the low dimensionality of the data, decision tree (DT) and random forest (RF) models were included in the trials for the final classifier. These were chosen for their structural

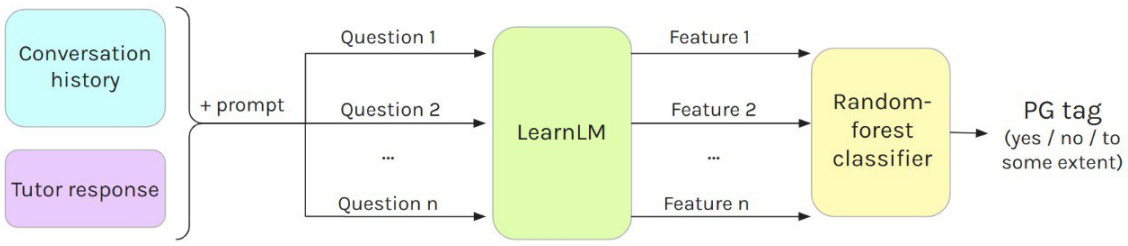

Figure 1: Proposed architecture to classify using expert pedagogical features.

mimicking of the decision process that pedagogy professionals described while annotating sample data.

6 Results

The variety of classifiers trained resulted from a different selection of extracted features as input, hyperparameter combinations, and the choice of DT versus RF model. A total 17320 DTs and 1400 RFs were trialed, each with 5-fold cross-validation, and the best candidates were then iterated using SMOTE oversampling. The highest performing model was an RF excluding some features from the input data, the hyperparameters of which we include in Appendix C, with training metrics detailed in Table 1.

Phase	**Exact macro F1**	Exact accuracy	Lenient macro F1	Lenient accuracy
Train	**0.5662**	0.6373	0.7529	0.8214
Test	**0.5369**	0.6244	0.7379	0.7822

Table 1: Performance of selected classifier model.

The final architecture using this model ranked 8th in the leaderboard for the pedagogical guidance dimension, with test metrics detailed in Table 1.

7 Conclusions

We have presented an approach to PG classification that combines LLMs and traditional AI techniques with a theoretical framework on PG. The features we propose offer a perspective that considers the interdependence of the original MRBench dimensions, but puts them all in service of how well the tutor guides the student.

Our work shows the potential of using PG-theory-based features, which is a fine-grained way of assessing elements of good-quality feedback while exploiting an LLM. Future work should explore other ways in which identification of these features may be exploited to iterate the construction of good-quality feedback via LLMs. In addition, we think that PG theory invites developers of AI tutors to take two other complementary routes for future work. The first is to design tutors aware of learning objectives, since this is fundamental to understand how to guide the student. The second is that AI tutors should build and exploit a student profile over time, considering the student's previous knowledge, degree of metacognition, and learning strategies that have previously worked or failed. Tackling these two action points would expand the frontier of AI tutor evaluation beyond the biggest limitations of this work from the standpoint of PG theory.

Limitations

Our architecture first assumes the limitations of our theoretical alignment: following Brookhart (2008) may better describe certain Western learning contexts than other sociocultural realities. Then, the architecture's reliance on LearnLM means it inherits any of the model's possible inaccuracies and biases, and that implementation depends on proprietary API use. Finally, the classifier model's performance should be improved with further trials using cleaned and augmented data.

Regarding the last point, the MRBench dataset carries limitations that transfer to our architecture. In terms of quality, we found conversation histories that we considered to be noisy: some lacking the original word problem being solved, with alternative tutor responses embedded within, or exchanging tutor/student speaker tags. We also did not find tagging criteria to be self-evident: the question of what constituted relevant "explanation, elaboration, hint, examples, and so on" seemed both open and necessitating at least some degree of expert pedagogical knowledge. Finally, the dataset is limited to the English language, mathematics school subject, arithmetic content and grade-school instruc-

tion level. Asymmetric advances in low-resource languages and higher influence of culture in other subjects of instruction limit the applicability of the benchmark for the range of intelligent tutoring systems currently on the market.

Finally, the strongest limitation surrounding this shared task was scarcity of context. In the pedagogical theory that we reviewed and that we believe is key to incorporate to these systems, the majority of factors contributing to PG are considered to be based on the student as a subject of learning. As such, factors that are regarded as key to PG are the student's individual previous knowledge, metacognitive ability, optimal learning strategies, personal relationship to the contents being taught, role in the classroom social dynamics, and sociocultural context (Brookhart, 2008; Smith and Stein, 2011; Chapin et al., 2009). Though we expanded as much as possible on the factors inferable from a single conversation via text, existing PG literature would suggest that an AI tutor's quality of PG can only be realistically estimated against a constructed learner profile of the student. Moreover, these considerations all defined their value only in relation to learning objectives and how they advanced the student towards them, meanwhile the context of what learning objectives were being reinforced in the tutoring sessions was not present in the MRBench dataset.

Acknowledgments

This work was supported by the National Center for Artificial Intelligence (CENIA) under Grant FB210017 Basal ANID. We would like to thank all of the following people. For their input to this work: Francisco Gazitúa and Juan Pablo Fuentes. For their support in timely experiment execution: to our colleagues at IALab, friends and family. For their contribution of expert knowledge in pedagogy: Francisca Ubilla, Edgar Valencia, Carolina Véliz, Teresita Fuentes, Emilia Deichler, Chiara Hiraizumi, Javier Riquelme, Marcelo Mena, Constanza Del Solar, Camila Sánchez, Andrea Vilca and Martín Pino. For their contribution of non-expert pedagogy knowledge: Luzhania Céspedes, Joaquín Handal, Guillermo Staudt, Raimundo Labbé, Daniel Villaseñor, Alexandre Icaza, Cristobal Soto, Vicente Muñoz, Victor Marques, José Chong, Matías Valenzuela, Maximiliano Berríos and Maximiliano Navia.

References

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic.

Susan M. Brookhart. 2008. *How to Give Effective Feedback to Your Students*. ASCD, Alexandria, VA.

Jerome Seymour Bruner. 1966. *Toward a Theory of Instruction*. Belknap Press of Harvard University.

Suzanne H. Chapin, Catherine O'Connor, and Nancy Canavan Anderson. 2009. *Classroom Discussions: Using Math Talk to Help Students Learn*. Math Solutions, California.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise Verification and Remediation of Student Reasoning Errors with Large Language Model Tutors. *Preprint*, arXiv:2407.09136.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, K. V. 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*.

Jakub Macina, Nico Daheim, Sankalan Pal 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. *Preprint*, arXiv:2305.14536.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025a. 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.

Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025b. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. *Preprint*, arXiv:2412.09416.

Margaret Schwan Smith and Mary Kay Stein. 2011. *5 Practices for Orchestrating Productive Mathematics Discussions*. National Council of Teachers of Mathematics, Reston, VA.

Anaïs Tack and Chris Piech. 2022. The AI Teacher Test: Measuring the Pedagogical Ability of Blender and GPT-3 in Educational Dialogues. *Preprint*, arXiv:2205.07540.

LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire,

Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, Irina Jurenka, James Cohan, Jennifer She, Julia Wilkowski, Kaiz Alarakyia, Kevin R. McKee, Lisa Wang, Markus Kunesch, Mike Schaekermann, and 27 others. 2024. Learnlm: Improving Gemini for Learning. *Preprint*, arXiv:2412.16429.

Patricia Tunstall and Caroline Gipps. 1996. Teacher Feedback to Young Children in Formative Assessment: a typology. *British Educational Research Journal*, 22(4):389–404.

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, Mexico City, Mexico. Association for Computational Linguistics.

A Full list Pedagogical Features

Table 2

Scope	Criterion	Question
Transversal	Psicosocial	Does the tutor's response focus on the specific task/process rather than the student personally?
Transversal	Psicosocial	Does the tutor's response frame mistakes as learning opportunities?
Transversal	Psicosocial	Does the tutor's response begin by affirming any partial success, even if minor?
Transversal	Metacognition	Throughout the conversation history and final response, does the tutor show preference for asking, rather than stating, to the student what their error could have been and/or how to fix it?
Local	Achieved	Does the tutor's response express that the student has taken some steps correctly?
Local	Achieved	Is the tutor's final response specific about which portion of the student's messages are going in the right direction to solve the proposed problem?
Local	Achieved	Is the tutor's final response correct about which portion of the student's messages are going in the right direction to solve the proposed problem?
Local	Mistaken	Does the tutor's final response imply that the student has made a mistake of some sort?
Local	Mistaken	Is the tutor's final response fully accurate in pointing out the student's mistake(s)?
Local	Mistaken	When communicating that the student has made a mistake, is the tutor's final repsonse specific with regards to what the alleged error was?
Local	Mistaken	Does the tutor's final response provide an explanation for why the student's approach was incorrect?
Local	Mistaken	Regarding the tutor's explanation for why the student's approach was incorrect, is it clear and understandable at a 6th grade level?
Local	Mistaken	Regarding the tutor's explanation for why the student's approach was incorrect, is it fully accurate?
Local	Remediate	Does the tutor offer the student a strategy or hint to solve the word problem?
Local	Remediate	Does the tutor offer the student a correct strategy or hint that would allow them to successfully solve the word problem?
Local	Remediate	Does the tutor offer the student a strategy to solve the word problem that is clear and understandable at the 6th-grade level?
Local	Remediate	Does the tutor offer the student an example problem or fact to correct a misinterpretation of the original problem?

B Feature Extraction Prompt

```
"""You will be presented with the conversation history from a grade-school math
tutoring session happening over computer chat, where the student makes a mistake or
evidences confusion.
Your task is to evaluate the tutor's final response in terms of the question:
{question}
---------------------------------
{conversation_history}
---------------------------------
Tutor Response: {tutor_response}
---------------------------------
Question: {question} (0 for No, 1 for Yes)
Answer: """
```

C Best-Performing Classifier Configuration

```
model_config = {
    'input_features': [
      ' Throughout the conversation history and final response, does the tutor show
       preference for asking, rather than stating, to the student what their error
         could have been and/or how to fix it?',
        'Does the tutor\'s response express that the student has taken some steps
         correctly?',
      'Is the tutor\'s final response specific about which portion of the student's
       messages are going in the right direction to solve the proposed problem?',
      'Is the tutor\'s final response correct about which portion of the student's
       messages are going in the right direction to solve the proposed problem?',
      'Does the tutor\'s final response imply that the student has made a mistake of
         some sort?',
      'Is the tutor\'s final response fully accurate in pointing out the student\'s
         mistake(s)?',
        'When communicating that the student has made a mistake, is the tutor\'s
        final response specific with regards to what the alleged error was?',
        'Does the tutor\'s final response provide an explanation for why the
        student\'s approach was incorrect?',
        'Regarding the tutor\'s explanation for why the student\'s approach was
        incorrect, is it clear and understandable at a 6th grade level?',
        'Regarding the tutor\'s explanation for why the student\'s approach was
        incorrect, is it fully accurate?',
        'Does the tutor offer the student a strategy or hint to solve the word
        problem?',
       'Does the tutor offer the student an example problem or fact to correct a
        misinterpretation of the original problem?',
    ]
    'preprocessing': {
        'oversampling': 'SMOTE'
    },
    'rf_hyperparameters': {
        'max_depth': None,
        'max_features': 'sqrt',
        'min_samples_leaf': 4,
        'n_estimators': 500
    }
}
```