# From Solver to Tutor: Evaluating the Pedagogical Intelligence of LLMs with KMP-Bench

**Weikang Shi**[*1], **Houxing Ren**[*1], **Junting Pan**[*1,3], **Aojun Zhou**[1], **Ke Wang**[1], **Zimu Lu**[1], **Yunqiao Yang**[1], **Yuxuan Hu**[1], **Linda Wei**[1], **Mingjie Zhan**[1], **Hongsheng Li**[†1,2,3]

[1]Multimedia Laboratory (MMLab), The Chinese University of Hong Kong
[2]Shanghai Artificial Intelligence Laboratory
[3]CPII under InnoHK
{wkshi@link, hsli@ee}.cuhk.edu.hk

## Abstract

Large Language Models (LLMs) show significant potential in AI mathematical tutoring, yet current evaluations often rely on simplistic metrics or narrow pedagogical scenarios, failing to assess comprehensive, multi-turn teaching effectiveness. In this paper, we introduce KMP-Bench, a comprehensive K-8 Mathematical Pedagogical Benchmark designed to assess LLMs from two complementary perspectives. The first module, KMP-Dialogue, evaluates holistic pedagogical capabilities against six core principles (e.g., Challenge, Explanation, Feedback), leveraging a novel multi-turn dialogue dataset constructed by weaving together diverse pedagogical components. The second module, KMP-Skills, provides a granular assessment of foundational tutoring abilities, including multi-turn problem-solving, error detection and correction, and problem generation. Our evaluations on KMP-Bench reveal a key disparity: while leading LLMs excel at tasks with verifiable solutions, they struggle with the nuanced application of pedagogical principles. Additionally, we present KMP-Pile, a large-scale (150K) dialogue dataset. Models fine-tuned on KMP-Pile show substantial improvement on KMP-Bench, underscoring the value of pedagogically-rich training data for developing more effective AI math tutors.

## Introduction

The rapid development of Large Language Models (LLMs) has transformed numerous fields, with education emerging as a particularly promising application area. LLMs have demonstrated considerable potential in AI tutoring, assisting with tasks like essay composition (Shao et al. 2024) and providing emotional support (Shin et al. 2022; Kang et al. 2025). However, mathematical tutoring presents unique challenges due to its demand for specialized knowledge, distinct pedagogical scenarios, and the difficulty in assessing true teaching effectiveness. While LLMs are increasingly adept at solving mathematical problems (Wang et al. 2023; Yang et al. 2024; DeepSeek-AI et al. 2025), proficiency in problem-solving does not equate to skilled teaching. Expert instruction requires a broader suite of capabilities, including adaptive challenging, effective modeling, guided practice, constructive feedback, etc. (Allison and Tharby 2015). Consequently, comprehensively evaluating the pedagogical abilities of LLMs for mathematical tutoring presents a significant and pressing challenge.

Current evaluation efforts for LLMs in mathematical tutoring present some limitations. One stream relies on objective metrics like problem-solving accuracy (Liang et al. 2024; Mishra et al. 2024) or textual similarity scores (e.g., BLEU, BERTScore) (Ding et al. 2024; Christ, Kropko, and Hartvigsen 2024), which, while easily quantifiable, inadequately reflect a model's teaching efficacy and interactional quality within dynamic, real-world educational settings. Another stream incorporates pedagogical dimensions but mainly uses existing datasets (e.g., MathDial (Macina et al. 2023), Bridge (Wang et al. 2024)) where evaluations are confined to the narrow scenario of error correction. This overlooks a spectrum of critical tutoring functions, such as proactive follow-up questioning, student confusion clarification, and the guided generation of practice problems—all hallmarks of effective, adaptive teaching. This focus also restricts problem scope (often elementary arithmetic problems) and dataset size. This gap means current assessments may not fully reveal how well AI tutors align with effective teaching practices.

To bridge these gaps, we introduce a systematic framework for building and evaluating AI math tutors. The foundation of this framework is a multi-stage data curation pipeline that generates pedagogically-rich, multi-functional tutoring dialogues. This pipeline first crafts four distinct pedagogical components—follow-up questions, error analysis scenarios, similar practice problems, and confusion clarifications—and then meticulously weaves them into coherent conversational flows that simulate authentic K-8 tutoring sessions. Applying this pipeline, we create two principal assets: an evaluation set of 4.6K dialogues, and a large-scale (150K) training dataset.

Built upon our evaluation set, we present **KMP-Bench**, a benchmark suite designed to assess AI math tutors from two complementary perspectives. The first module, **KMP-Dialogue**, evaluates *holistic pedagogical capabilities*. It as-

[*]These authors contributed equally.
[†]Corresponding author.

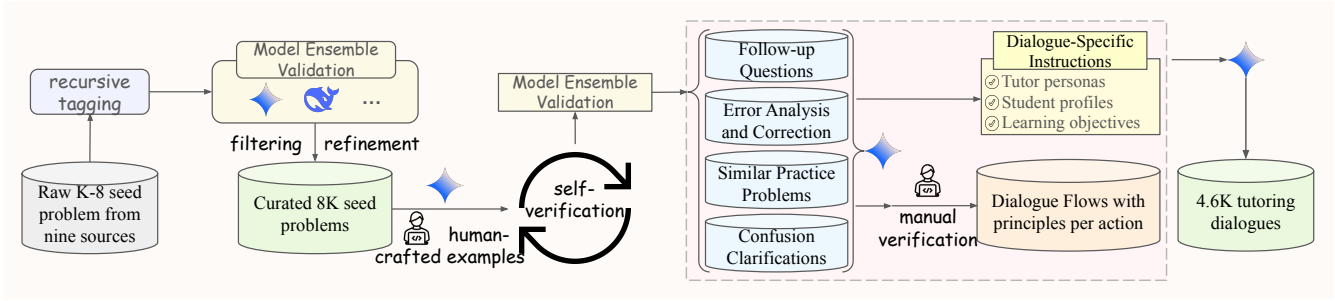



Figure 1: The curation pipeline for the KMP-Bench dataset. The process generates four distinct pedagogical components from curated K-8 problems using an LLM guided by human-crafted few-shot examples. After a rigorous quality control process (including LLM-based self-verification and model ensemble validation), these components are organized into Dialogue Flows, which are then manually verified for pedagogical soundness before being woven into final tutoring dialogues.

sesses a model's ability to generate contextually appropriate responses within truncated dialogues, measuring performance against six core pedagogical principles adapted from established educational guidelines (Challenge, Explanation, Modelling, Practice, Questioning, and Feedback) (Allison and Tharby 2015).

The second module, **KMP-Skills**, provides a more granular assessment of *foundational tutoring abilities*. It leverages the individual pedagogical components to create targeted evaluations for three specific skills: (1) multi-turn follow-up problem-solving, (2) error detection and correction, and (3) the generation of valid and pedagogically-sound mathematical problems. Together, KMP-Dialogue and KMP-Skills offer a robust evaluation of an AI tutor's effectiveness across a wide spectrum of K-8 mathematical content.

Our extensive evaluations using KMP-Bench reveal a significant disparity: while current LLMs show promise in structured tasks like problem-solving, they struggle with the nuanced application of pedagogical principles and the reliable generation of pedagogically-sound responses. Furthermore, models fine-tuned on KMP-Pile dataset demonstrate substantial improvements across the KMP-Bench evaluations, underscoring the critical value of pedagogically-rich training data. Our main contributions are:

1. We introduce a comprehensive benchmark, KMP-Bench, composed of two distinct modules for multi-faceted evaluation: KMP-Dialogue for assessing holistic, principle-based pedagogical capabilities, and KMP-Skills for performing a granular assessment of foundational tutoring abilities.
2. We provide a detailed empirical analysis that reveals a key weakness in modern LLMs: a trade-off between task-solving proficiency and pedagogical quality. Our results highlight the urgent need to focus research on improving pedagogical reasoning and reliable content generation.
3. We release KMP-Pile, a large-scale (150K) training dataset of multi-turn tutoring dialogues. We demonstrate that training on this dataset significantly enhances AI tutor performance, serving as a valuable resource for the community.

## Tutoring Dialogue Curation

This section details our data curation pipeline, illustrated in Figure 1, which systematically transforms curated K-8 seed problems into a corpus of pedagogically diverse and instructionally sound tutoring dialogues. The process unfolds in three main phases: establishing a seed problem foundation, crafting distinct pedagogical components, and weaving these components into coherent conversational flows.

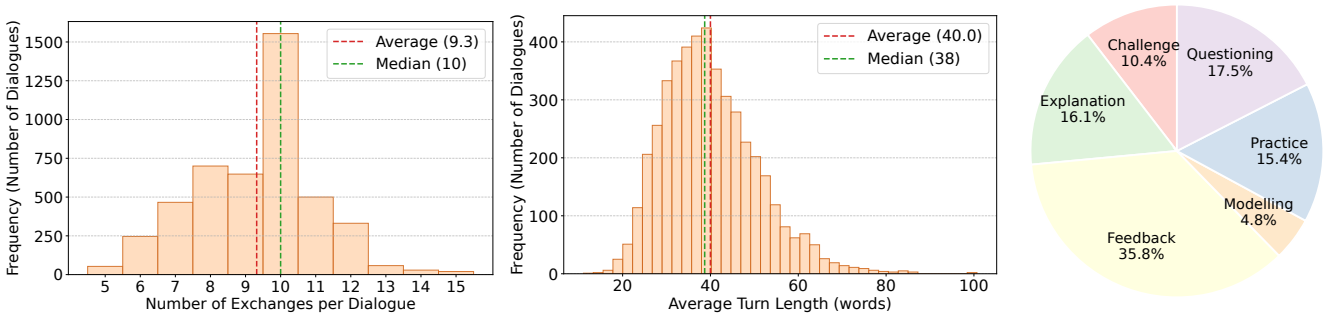
### Seed Problem Foundation

To overcome the limited K-8 curriculum coverage and small scale of existing mathematical educational evaluation datasets like MathDial and Bridge, we established a comprehensive foundation of seed problems. This involved an initial collection from nine diverse elementary to middle school mathematics datasets, followed by a systematic categorization using the recursive tagging algorithm from Lucy et al. (2024) based on Common Core Standards, resulting in problems spanning 11 distinct mathematical domains. To meet the rigorous quality standards required for a benchmark, each problem underwent model ensemble validation for correctness, while its corresponding solution was subsequently refined and standardized. This meticulous process, detailed further in Appendix, yielded 8K validated K-8 mathematical problems that serve as the foundation for our tutoring dialogues.

### Crafting Pedagogical Interaction Components

Building upon the curated seed problems, we systematically generated four types of core conversational components designed to embody distinct pedagogical functions. To ensure the quality and relevance of the outputs, each generation task was guided by a set of human-crafted few-shot examples. This foundational step provided the LLM with clear, exemplary precedents for each component type. This generation process primarily utilized the Gemini-2.0-Flash model.

**Knowledge-Extending Follow-up Questions** are designed to deepen conceptual understanding and foster cognitive extension. For each seed problem, we generated a sequence of follow-up questions where each subsequent question was carefully scaffolded to incrementally increase dif-




(a) Distribution of Exchanges (b) Distribution of Avg. Turn Length (c) Distribution of Principles

Figure 2: Statistical distributions of the dataset.

ficulty and explore related mathematical concepts. This process included guided reasoning to ensure pedagogical relevance and robust validation of the questions' accuracy.

**Error Analysis and Correction** scenarios were constructed to develop students' critical thinking and error analysis capabilities. This process began by systematically generating plausible incorrect solutions for the seed problems, by employing varied modeling techniques and parameters designed to elicit common mathematical errors. To ensure a consistent basis for analysis, all solutions, whether correct or incorrect, were then standardized into a uniform step-by-step format, preserving the original reasoning path. Subsequently, an LLM was utilized to generate detailed correction feedback. This feedback involved analyzing the presented solution, pinpointing the first erroneous step if an error existed, and providing a rectified solution from that step forward. A multi-stage verification protocol was also implemented to ensure the quality, consistency, and pedagogical value of these error analysis components.

**Similar Practice Problems** were generated to reinforce skill acquisition by providing targeted practice across a spectrum of difficulty levels. Guided by Common Core Standards annotations, we prompted an LLM to create new problems at three distinct difficulty tiers—easy, medium, and hard—using the original seed problem as the medium-difficulty reference. The generation process incorporated guided reasoning stages for the LLM and subsequent validation, including measures to ensure problem diversity and avoid trivial variations of the seed problem.

**Confusion Clarifications** dialogues were designed to resolve common points of confusion encountered by K-8 students and provide targeted, explanatory interactions. Using curated examples of genuine student queries, we prompted an LLM to generate multi-turn interactions between a simulated K-8 student and a tutor. These dialogues specifically aim to clarify conceptual understanding (What), elucidate solution steps and methods (How), and explore the underlying rationale for chosen approaches (Why), thereby fostering a deeper comprehension of the mathematical content.

Upon completing the generation process for all components, we filtered our collection to retain the 6K seed problems for which all four pedagogical component types were successfully created. This curated set forms the foundation for our KMP-Skills evaluation, as detailed in Section .

## Dialogue Weaving and Pedagogical Enrichment

To weave the individual pedagogical components into coherent, multi-turn dialogues, we developed a pipeline designed to simulate K-8 tutoring sessions. The process systematically structures the conversation and enriches it with pedagogical diversity and personalized context.

**Dialogue Flow Generation and Verification.** The pipeline begins with a Flow Generator LLM that creates a "Dialogue Flow"—a structured blueprint for the entire tutoring interaction. This LLM designs a pedagogical "storyline" by selecting and sequencing the relevant material components. This stage is where the core pedagogical enrichment occurs: the LLM defines tutor personas, student profiles with potential knowledge gaps, and specific learning objectives. Furthermore, it assigns one or two of six core pedagogical principles to each planned tutor action, creating a foundation for varied and adaptive interactions.

Crucially, these LLM-generated Dialogue Flows then undergo manual verification. Human reviewers assessed each flow against shared quality criteria. They were tasked with identifying and addressing critical flaws, such as the hallucination of materials or illogical pedagogical sequencing, and truncating flows to remove conversational turns that lacked pedagogical value. This process resulted in the removal of 451 (7.6%) problematic flows.

**Dialogue Expansion.** Following human approval, the verified Dialogue Flow is passed to a Dialogue Expander LLM, which translates the blueprint into a complete, natural-language dialogue. It expands the planned actions, referenced materials, and persona details from the flow into fluent and K-8 appropriate conversational turns. This two-tiered process, with its human-in-the-loop oversight, enables the creation of dialogues that are both instructionally structured and context-aware.

Using this pipeline, we generated the 4.6K dialogues for our benchmark. To scale up data creation, we then adapted the process by replacing the manual verification step with an LLM-based verifier. Applying this modified pipeline to the Orca-Math (Mitra et al. 2024) seed problems yielded KMP-

Pile, a training dataset of 150K dialogues.

### Dataset Statistics

The seed problems for our dataset are sourced from a diverse collection of 9 educational platforms, encompassing 11 distinct mathematical domains and spanning 9 grade levels from Kindergarten to Grade 8 (K-8).

Figure 2 highlights the substantial nature of our generated dialogues. The dialogues average a robust 9.3 exchanges (Figure 2 (a)), a length sufficient for facilitating meaningful, multi-step pedagogical interactions. Furthermore, the considerable average turn length of 40 words (Figure 2 (b)) indicates that the tutor's utterances are detailed and explanatory, which is a hallmark of effective instructional content.

## Evaluation Framework

### KMP-Dialogue: Evaluating Holistic Pedagogical Capabilities

To rigorously assess the pedagogical capabilities of tutor LLMs, we developed a multifaceted evaluation framework, as show in Appendix. The framework comprises three main stages: setting up evaluation instances through dialogue truncation, generating dialogue-specific instructions, and curating a nuanced set of evaluation criteria.

**Evaluation Instance Preparation via Dialogue Truncation.** First, to prepare evaluation instances, we adapt our coherent tutoring interaction dataset. Dialogues are strategically truncated at a tutor's turn, preserving the conversational history up to the preceding student utterance. This creates a test case where a tutor LLM is prompted to generate the subsequent pedagogical response. Each selected tutor turn is guided by one or two of our six core pedagogical principles (Challenge, Explanation, Modelling, Practice, Questioning, and Feedback). Figure 2 (c) shows the precise distribution of these principles across the truncated turns used in our evaluation set. Furthermore, an LLM assists in filtering these truncated dialogues, retaining only those deemed to possess significant value for evaluation.

**Dialogue-Specific Instructions Generation.** Second, recognizing that effective tutoring responses are highly dependent on context, we generate dialogue-specific instructions for the tutor LLM under evaluation. Since each truncated dialogue originates from a unique scenario with specific tutor personas, student profiles, and learning objectives, this contextual information is vital. We employ an LLM to craft a concise set of instructions for each test case, serving as a system prompt for the tutor LLM. Crucially, these instructions not only convey the necessary scenario details but also explicitly specify the target pedagogical principle(s) that the tutor LLM's response should embody for that particular turn. This approach facilitates targeted, principle-based evaluation and assesses the model's capacity to adhere to specific pedagogical directives.

**Context-Aware Evaluation Criteria and Method.** Finally, to evaluate the generated tutor responses, we curated a comprehensive set of 22 evaluation criteria, moving beyond generic metrics. This involved both manual expert specification and LLM assistance. Our evaluation framework is distinguished by its two-tiered structure—comprising 4 general criteria applicable to all responses and 3 principle-specific criteria for each of the six pedagogical principles—and its dialogue-specific nature. The criteria are adapted to consider the unique context of each interaction, such as the established tutor persona or student needs, enhancing evaluation validity. For assessment, the tutor LLM's generated response is compared against the original tutor turn (which was truncated from the dialogue and serves as a reference response). This comparison yields a Win, Tie, or Lose outcome for each applicable criterion and an overall judgment for the tutor's entire response. This multi-layered feedback enables a nuanced analysis of the tutor LLM's pedagogical strengths and weaknesses in multi-turn conversational settings. The grading prompt is available in Appendix.

### KMP-Skills: Evaluating Foundational Tutoring Abilities

While the evaluation of our math tutoring dialogues provides insight into a model's overall pedagogical and instructional capabilities, the proficiency in core, granular tutoring skills is also critical. Abilities such as multi-turn problem solving, accurate error detection and correction, and the generation of valid practice problems form the bedrock of effective teaching. We use the 6K Pedagogical Interaction Components from Section  to construct targeted evaluations for these three skills. Collectively, we refer to this suite of evaluations as KMP-Skills.

**Multi-turn Follow-up Problem-Solving.** This evaluation assesses a model's continuous reasoning and its ability to handle increasing problem complexity within a dialogue. Adopting a multi-turn conversational structure similar to MathChat (Liang et al. 2024), we use sequences comprising a seed problem and two connected, progressively harder follow-up questions. The evaluation proceeds by making the model's response to each question the context for the subsequent one. We calculate solution accuracy independently for each of the three turns. This per-turn metric reveals how a model's performance changes as conversational depth and problem difficulty increase.

**Error Detection and Correction.** An effective tutor must identify and correct student errors. This evaluation uses our set of generated student solutions, which includes both correct examples and incorrect ones containing annotated error steps and reasons. Inspired by ProcessBench (Zheng et al. 2024), we evaluate the model's core ability to find and fix mistakes. In this setup, the model performs a two-turn task: (1) it identifies whether a solution is correct and, if not, pinpoints the first erroneous step; (2) it provides a fully corrected solution. Performance is measured by two fundamental metrics: the F1-score for the error identification in Turn 1 and the correction accuracy of the solution in Turn 2.

Further, to gain a more nuanced understanding of a model's ability to diagnose not just where an error is but why it occurred, we adopt the MR-Score from the MR-GSM8K framework (Zeng et al. 2024). This composite score offers deeper diagnostic insight by jointly evaluating three model outputs: the final corrected solution, the identified first error step, and the validity of the stated error reason.

| Model | Overall Judgement | General-Level Acc | Principle-Level Acc | | | | | | Overall Acc |
|---|---|---|---|---|---|---|---|---|---|
| | | | Challenge | Explanation | Modelling | Practice | Questioning | Feedback | |
| Closed-Source Models | | | | | | | | | |
| GPT-4o | 60.7 | 44.5 | 43.4 | 62.2 | 57.5 | 47.5 | 59.9 | 41.3 | 48.2 |
| Gemini-2.0-Flash | 58.5 | 38.0 | 51.1 | 49.9 | 34.2 | 42.2 | 50.8 | 37.6 | 41.1 |
| Claude-3.7-Sonnet | 87.1 | 69.8 | 80.0 | 79.3 | 71.2 | 69.1 | 79.3 | 72.2 | 72.5 |
| LearnLM-2.0-Flash | 74.8 | 58.9 | 77.3 | 42.7 | 24.9 | 67.6 | 73.5 | 63.9 | 58.6 |
| Open-Source Models | | | | | | | | | |
| DeepSeek-V3 | 84.6 | 71.6 | 70.3 | 81.5 | 71.4 | 69.6 | 83.9 | 70.8 | 73.1 |
| Qwen2.5-Math-72B-Inst | 31.0 | 27.1 | 25.6 | 51.3 | 42.9 | 33.9 | 29.5 | 23.7 | 30.8 |
| Qwen2.5-72B-Inst | 68.5 | 55.5 | 50.5 | 77.3 | 71.5 | 60.7 | 54.0 | 52.4 | 58.3 |
| Qwen2.5-7B-Inst | 62.5 | 52.9 | 41.2 | 74.2 | 66.5 | 56.6 | 49.7 | 50.6 | 54.7 |
| LLaMa-3.1-8B | 51.3 | 40.3 | 33.0 | 50.1 | 29.0 | 38.3 | 55.0 | 46.2 | 41.1 |
| SocraticLM | 21.2 | 15.4 | 17.0 | 25.0 | 21.0 | 21.4 | 27.3 | 17.8 | 18.5 |
| $\text{MathChat}_{\text{sync}}$-LLM | 16.1 | 12.0 | 16.0 | 23.0 | 15.4 | 17.6 | 23.1 | 16.3 | 15.3 |
| TutorChat-LLM | 32.1 | 30.0 | 21.1 | 41.3 | 33.2 | 37.1 | 40.6 | 29.9 | 31.9 |
| Qwen2.5-Math-7B-Inst | 21.0 | 19.7 | 17.5 | 41.7 | 34.5 | 23.2 | 25.5 | 22.6 | 23.6 |
| **KMP-LM-7B** | 40.4 | 35.1 | 29.3 | 47.0 | 38.5 | 39.8 | 43.0 | 36.1 | 37.0 |
| $\Delta$ | +19.4 | +15.4 | +11.8 | +5.3 | +4.0 | +16.6 | +17.5 | +13.5 | +13.4 |

Table 1: Overall pedagogical performance of various LLMs in KMP-Dialogue. Tutor LLM responses are compared against reference responses from our curated dialogues. The best and second-best results among all the models are highlighted in red and blue, respectively. The $\Delta$ row represents the improvement of KMP-LM-7B over Qwen2.5-Math-7B-Instruct.

**Mathematical Problem Generation.** The ability to generate valid and contextually relevant mathematical problems is crucial for adaptive tutoring systems. However, ensuring that auto-generated problems and their solutions are logically sound and error-free remains a significant challenge. This evaluation, therefore, assesses a model's ability to generate two key types of problems: *follow-up problems*, which extend concepts from a seed problem with increasing complexity, and *similar problems*, which target the same skills using different contexts or values for reinforcement.

This generation task employs a three-turn dialogue. Given a seed problem and solution, models sequentially generate two new problems (either follow-up or similar types) along with their solutions, maintaining conversational context. These generated problem-solution pairs are then evaluated by a strong LLM using a strict "fail-by-default" binary classification across key dimensions: Problem Construction, Solution Correctness, Solution Quality, and (for similar problems) Similarity Appropriateness. Performance is measured as the average proportion of "Pass" judgments across all evaluated dimensions.

# Experiments

## Experimental Setup

**Metrics.** We use Gemini-2.0-Flash as the LLM evaluator. For KMP-Dialogue, all accuracies are win rates. Key metrics include: Overall Judgement Acc. (from the evaluator's holistic Win/Tie/Lose decision); General-Level Acc. (average win rate across 4 general criteria); six Principle-Level Acc. scores (average win rate of 3 specific criteria for each of the 6 pedagogical principles); and a composite Overall Acc. (an average of General-Level Acc. and the mean of the six Principle-Level Accs.). For KMP-Skills, we report only the final-round performance of follow-up problem-solving and problem generation for brevity.

**Models.** Our evaluation spans a diverse set of LLMs. This includes: (a) closed-source models: GPT-4o (OpenAI 2024b), Gemini-2.0-Flash (Google 2025), Claude-3.7-Sonnet (Anthropic 2025), and LearnLM-2.0-Flash (Team et al. 2024); (b) a broad range of open-source general-purpose and math-specialized models such as the LLaMA series, Qwen2.5 series, and DeepSeek-V3 (DeepSeek-AI 2025); and (c) models with a tutoring focus. Among these are established tutor-specific models like SocraticLM (Liu et al. 2024). Furthermore, we fine-tuned Qwen2.5-Math-7B on two tutoring datasets, $\text{MathChat}_{\text{sync}}$ (Liang et al. 2024) and TutorChat (Chevalier et al. 2024), resulting in $\text{MathChat}_{\text{sync}}$-LLM and TutorChat-LLM.

**Training Details.** We developed our primary model, KMP-LM-7B, by fine-tuning Qwen2.5-Math-7B base model. The training data consisted of 150K KMP-Pile directly mixed with 208K general instruction samples, which were created by combining 100K randomly sampled conversations from UltraChat (Ding et al. 2023) with 108K multi-turned WildChat (Zhao et al. 2024) data. More training details, including hyperparameter selection, are presented in Appendix. The $\text{MathChat}_{\text{sync}}$-LLM and TutorChat-LLM were fine-tuned with the same number of steps and comparable training configurations.

## Main Results

**Overall Pedagogical Performance.** The overall pedagogical performance of various LLMs on KMP-Dialogue is presented in Table 1. Among prominent closed-source models,

| Model | Size | KMP-FQA | KMP-EC | | | KMP-Gen | |
|---|---|---|---|---|---|---|---|
| | | | MR-Score | F1-Score | Correction | Follow-up | Similar |
| Closed-Source Models | | | | | | | |
| GPT-4o | - | 84.1 | 64.4 | 78.3 | 90.9 | 74.8 | 72.2 |
| GPT-4o mini | - | 79.4 | 56.9 | 66.1 | 84.3 | 76.5 | 64.8 |
| Gemini 2.0 Flash | - | 91.2 | 86.3 | 87.5 | 96.7 | 75.0 | 73.2 |
| LearnLM-1.5-Pro | - | 87.7 | 69.8 | 74.6 | 94.4 | 77.0 | 75.0 |
| Open-Source Models | | | | | | | |
| Qwen2.5-72B-Inst | 72B | 82.2 | 65.8 | 78.6 | 89.5 | 81.3 | 78.1 |
| Qwen2.5-Math-72B-Inst | 72B | 83.2 | 53.8 | 73.5 | 81.0 | 86.3 | 82.6 |
| Qwen2.5-7B-Inst | 7B | 71.0 | 34.9 | 44.3 | 73.2 | 66.2 | 66.2 |
| Llama-3.1-8B-Inst | 8B | 50.1 | 28.6 | 9.0 | 55.9 | 41.7 | 38.5 |
| $\text{MathChat}_{\text{sync}}$-LLM | 7B | 53.5 | 23.5 | 9.8 | 18.9 | 70.9 | 71.3 |
| TutorChat-LLM | 7B | 57.7 | 24.0 | 38.0 | 48.5 | 65.8 | 69.5 |
| Qwen2.5-Math-7B-Inst | 7B | 76.1 | 3.4 | 44.1 | 52.3 | 48.6 | 52.8 |
| **KMP-LM-7B** | 7B | 77.5 | 49.0 | 57.5 | 63.4 | 81.1 | 83.2 |
| $\Delta$ | | +1.4 | +45.6 | +13.4 | +11.1 | +32.5 | +30.4 |

Table 2: Foundational tutoring performance of various LLMs on KMP-Skills. Abbreviations: KMP-FQA denotes Multi-Turn Problem Solving, KMP-EC denotes Error Detection and Correction, and KMP-Gen denotes Problem Generation. The best and second-best results among all the models are highlighted in red and blue, respectively. The $\Delta$ row represents the improvement of KMP-LM-7B over Qwen2.5-Math-7B-Instruct.

| Model | Grading Method | General-Level Acc | Principle-Level Acc | Overall Acc | Alignment Rate (%) |
|---|---|---|---|---|---|
| LearnLM-2.0-Flash | Gemini-2.0-Flash | 61.8 | 60.4 | 61.1 | 89.7 |
| | Manual | 62.5 | 57.9 | 60.2 | |
| Claude-3.7-Sonnet | Gemini-2.0-Flash | 72.5 | 75.7 | 74.1 | 87.1 |
| | Manual | 72.9 | 80.5 | 76.7 | |
| DeepSeek-V3 | Gemini-2.0-Flash | 75.0 | 75.5 | 75.2 | 92.5 |
| | Manual | 75.8 | 76.5 | 76.1 | |

Table 3: Comparison of evaluation results between our LLM evaluator (Gemini-2.0-Flash) and manual human annotation on a subset of KMP-Dialogue. The Alignment Rate measures the percentage of criteria where the LLM evaluator's judgment aligns with the human ground truth.

Claude-3.7-Sonnet achieves the leading Overall Accuracy at 72.5. In the open-source domain, DeepSeek-V3 stands out with the highest Overall Accuracy (73.1), demonstrating that top-tier open-source models are highly competitive. Beyond the leaders, the Qwen2.5 model family offers a particularly insightful finding: the general-purpose Qwen2.5-72B-Inst comprehensively outperforms its math-tuned counterpart, Qwen2.5-Math-72B-Inst, suggesting that while math-specific models are optimized for solving problems, this specialization does not readily translate to the nuanced skill of pedagogical dialogue. The principle-level analysis further reveals diverse strengths across models: DeepSeek-V3 excels in Explanation (81.5) and Questioning (83.9), while Claude-3.7-Sonnet is dominant in Challenge (80.0) and Feedback (72.2).

Our fine-tuned model, KMP-LM-7B, demonstrates the effectiveness of our approach, achieving an Overall Accuracy of 37.0. This marks a substantial improvement over Qwen2.5-Math-7B-Inst model, with absolute gains of +13.4 in Overall Accuracy and +19.4 in the holistic Overall Judgement score. This performance also positions KMP-LM-7B favorably against other open-source tutoring-focused models, surpassing SocraticLM (18.5), $\text{MathChat}_{\text{sync}}$-LLM (15.3), and TutorChat-LLM (31.9). These results underscore that fine-tuning on the pedagogically structured dialogues within our KMP-Pile dataset is a highly effective strategy for boosting the pedagogical capabilities of LLMs, validating our data curation methodology.

**Foundational Tutoring Skills Evaluation.** The performance of various models on KMP-Skills is detailed in Table 2. In general, leading closed-source models like Gemini-2.5-Flash show high proficiency in the more structured tasks, achieving a score of 91.2 in Multi-Turn Follow-up Problem-Solving and a 96.7 correction accuracy in Error Detection and Correction. Our fine-tuned model, KMP-LM-7B, achieves a competitive 77.5 in problem-solving and demonstrates substantial improvements over Qwen2.5-Math-7B-Inst model across the board. The gains are most pronounced

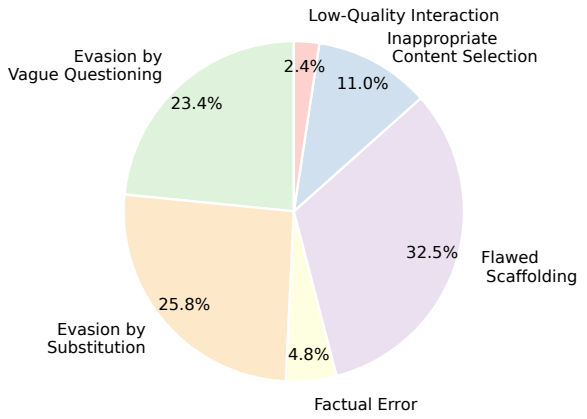



Figure 3: The distribution of the task errors.

in Error Detection and Correction (a remarkable +45.6 increase in MR-Score) and Problem Generation (Follow-up: +32.5, Similar: +30.4).

Taken together, our results reveal a critical insight into current AI tutors. While LLMs have become exceptionally proficient at tasks with verifiable solutions, such as problem-solving and error correction, a significant performance gap remains in tasks requiring deep pedagogical awareness. This gap is evident in both the inconsistent performance on teaching principles in KMP-Dialogue and the broader challenge of creating pedagogically sound questions in the Problem Generation task. This distinction underscores that the frontier for AI tutors is shifting from being accurate problem-solvers to becoming effective, student-centric educators.

## Ablation Studies

**Analysis of the Accuracy of KMP-Dialogue Evaluator Results.** To validate our automated evaluation pipeline, we conducted a human evaluation to assess the reliability of our LLM evaluator, Gemini-2.0-Flash. We randomly selected 300 instances from KMP-Dialogue, covering responses from three top-performing models, and had them annotated by human experts to establish a ground truth. As detailed in Table 3, the results from Gemini-2.0-Flash closely mirror the manual evaluations, with an average Alignment Rate of 89.8% between the LLM's judgments and the human ground truth. This high level of agreement confirms that our automated evaluation is effective and that the LLM evaluator serves as a reliable proxy for human judgment. Further details on the human evaluation protocol are provided in Appendix.

**Analysis of the Errors in KMP-Dialogue Tasks.** We analyze the errors and pedagogical flaws that occur in KMP-Dialogue evaluation using mainstream LLMs such as DeepSeek-V3, as illustrated in Figure 3. A detailed explanation of the error types and representative examples can be found in Appendix. Currently, the most prevalent issues are flawed scaffolding (32.5%) and evasion by substitution (25.8%), indicating that LLMs struggle with providing appropriate pedagogical guidance and tend to avoid directly addressing student needs by offering tangential responses.

## Related Works

**Evaluation of AI-based Math Tutors.** Evaluating AI math tutors requires frameworks that assess the interactive pedagogical process, beyond single-turn problem-solving accuracy. However, many current evaluations still rely on isolated accuracy metrics for discrete skills like multi-turn reasoning (Zheng et al. 2023; Shi et al. 2025; Mishra et al. 2024), error correction (Zheng et al. 2024; Zeng et al. 2024), or problem generation (Liang et al. 2024; Christ, Kropko, and Hartvigsen 2024), failing to capture the quality of the holistic interaction. While datasets like MathDial (Macina et al. 2023) construct tutoring dialogues for error correction scenarios, their reliance on textual similarity scores (BLEU, BERTScore) correlates poorly with pedagogical quality. Furthermore, the problem scope of existing benchmarks is often confined to arithmetic word problems. Even specialized benchmarks like MathTutorBench (Macina et al. 2025) and MRBench (Maurya et al. 2025), while evaluating pedagogical dimensions, tend to assess specific tutoring aspects in isolation. KMP-Bench addresses these gaps by providing a holistic evaluation framework. It is designed to assess a model's ability to seamlessly integrate key tutoring functions across a diverse K-8 curriculum, using pedagogically-grounded criteria rather than simplistic metrics.

**LLMs for Math Education.** Recent advances in LLMs have spurred their application in building intelligent tutoring systems (ITS) with high pedagogical value (Wu et al. 2022; Létourneau et al. 2025; Son 2024). Two primary approaches have emerged. The first involves creating agentic systems (Bidarian 2023; Wang et al. 2025), which leverage powerful foundation models like GPT-4 (OpenAI 2024a). While these systems offer immediate, personalized support, they can lack deep pedagogical grounding and struggle with real-world alignment. The second approach involves fine-tuning LLMs on tutoring datasets to instill specific capabilities. These datasets are either human-annotated, like MathDial (Macina et al. 2023), which are pedagogically authentic but costly to scale, or synthetically generated (Liu et al. 2024; Chevalier et al. 2024). Synthetic approaches offer scalability but often suffer from factual hallucinations and a lack of pedagogical diversity. Our work addresses these limitations by introducing a pipeline that generates dialogues integrating multiple core tutoring functions, ensuring both pedagogical richness and factual accuracy through a structured generation and verification process.

## Conclusion

This paper introduced KMP-Bench, a comprehensive K-8 benchmark with two modules, KMP-Dialogue and KMP-Skills, to assess both holistic, principle-based pedagogical capabilities and foundational tutoring skills. Our evaluations highlight a critical weakness: LLMs are proficient at tasks with verifiable solutions like problem-solving but require significant improvement in applying pedagogical principles and generating valid mathematical problems. Training on our pedagogically-rich dataset, KMP-Pile, improved performance on KMP-Bench, underscoring that such data is crucial for advancing AI from mere solvers to effective tutors.

## Acknowledgements

This study was supported in part by National Key R&D Program of China Project 2022ZD0161100, in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, in part by NSFC-RGC Project N_CUHK498/24, and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW).

## References

Allison, S.; and Tharby, A. 2015. *Making Every Lesson Count: Six Principles to Support Great Teaching and Learning*. Carmarthen, Wales: Crown House Publishing Limited. ISBN 978-1845909734.

Anthropic. 2025. Claude 3.7 Sonnet System Card. https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf. Accessed: 2025-08-01.

Bidarian, N. 2023. Meet Khan Academy's chatbot tutor. *CNN Business*.

Chevalier, A.; Geng, J.; Wettig, A.; Chen, H.; Mizera, S.; Annala, T.; Aragon, M. J.; Fanlo, A. R.; Frieder, S.; Machado, S.; Prabhakar, A.; Thieu, E.; Wang, J. T.; Wang, Z.; Wu, X.; Xia, M.; Jia, W.; Yu, J.; Zhu, J.-J.; Ren, Z. J.; Arora, S.; and Chen, D. 2024. Language Models as Science Tutors. arXiv:2402.11111.

Christ, B. R.; Kropko, J.; and Hartvigsen, T. 2024. MATHWELL: Generating Educational Math Word Problems Using Teacher Annotations. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11914–11938. Miami, Florida, USA: Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437.

DeepSeek-AI; Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; Zhang, X.; Yu, X.; Wu, Y.; Wu, Z. F.; Gou, Z.; Shao, Z.; Li, Z.; Gao, Z.; Liu, A.; Xue, B.; Wang, B.; Wu, B.; Feng, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; Dai, D.; Chen, D.; Ji, D.; Li, E.; Lin, F.; Dai, F.; Luo, F.; Hao, G.; Chen, G.; Li, G.; Zhang, H.; Bao, H.; Xu, H.; Wang, H.; Ding, H.; Xin, H.; Gao, H.; Qu, H.; Li, H.; Guo, J.; Li, J.; Wang, J.; Chen, J.; Yuan, J.; Qiu, J.; Li, J.; Cai, J. L.; Ni, J.; Liang, J.; Chen, J.; Dong, K.; Hu, K.; Gao, K.; Guan, K.; Huang, K.; Yu, K.; Wang, L.; Zhang, L.; Zhao, L.; Wang, L.; Zhang, L.; Xu, L.; Xia, L.; Zhang, M.; Zhang, M.; Tang, M.; Li, M.; Wang, M.; Li, M.; Tian, N.; Huang, P.; Zhang, P.; Wang, Q.; Chen, Q.; Du, Q.; Ge, R.; Zhang, R.; Pan, R.; Wang, R.; Chen, R. J.; Jin, R. L.; Chen, R.; Lu, S.; Zhou, S.; Chen, S.; Ye, S.; Wang, S.; Yu, S.; Zhou, S.; Pan, S.; Li, S. S.; Zhou, S.; Wu, S.; Ye, S.; Yun, T.; Pei, T.; Sun, T.; Wang, T.; Zeng, W.; Zhao, W.; Liu, W.; Liang, W.; Gao, W.; Yu, W.; Zhang, W.; Xiao, W. L.; An, W.; Liu, X.; Wang, X.; Chen, X.; Nie, X.; Cheng, X.; Liu, X.; Xie, X.; Liu, X.; Yang, X.; Li, X.; Su, X.; Lin, X.; Li, X. Q.; Jin, X.; Shen, X.; Chen, X.; Sun, X.; Wang, X.; Song, X.; Zhou, X.; Wang, X.; Shan, X.; Li, Y. K.; Wang, Y. Q.; Wei, Y. X.; Zhang, Y.; Xu, Y.; Li, Y.; Zhao, Y.; Sun, Y.; Wang, Y.; Yu, Y.; Zhang, Y.; Shi, Y.; Xiong, Y.; He, Y.; Piao, Y.; Wang, Y.; Tan, Y.; Ma, Y.; Liu, Y.; Guo, Y.; Ou, Y.; Wang, Y.; Gong, Y.; Zou, Y.; He, Y.; Xiong, Y.; Luo, Y.; You, Y.; Liu, Y.; Zhou, Y.; Zhu, Y. X.; Xu, Y.; Huang, Y.; Li, Y.; Zheng, Y.; Zhu, Y.; Ma, Y.; Tang, Y.; Zha, Y.; Yan, Y.; Ren, Z. Z.; Ren, Z.; Sha, Z.; Fu, Z.; Xu, Z.; Xie, Z.; Zhang, Z.; Hao, Z.; Ma, Z.; Yan, Z.; Wu, Z.; Gu, Z.; Zhu, Z.; Liu, Z.; Li, Z.; Xie, Z.; Song, Z.; Pan, Z.; Huang, Z.; Xu, Z.; Zhang, Z.; and Zhang, Z. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. arXiv:2305.14233.

Ding, Y.; Hu, H.; Zhou, J.; Chen, Q.; Jiang, B.; and He, L. 2024. Boosting Large Language Models with Socratic Method for Conversational Mathematics Teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 3730–3735.

Google. 2025. Gemini 2.0 Flash Model Card. https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf. Accessed: 2025-08-01.

Kang, D.; Kim, S.; Kwon, T.; Moon, S.; Cho, H.; Yu, Y.; Lee, D.; and Yeo, J. 2025. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. arXiv:2402.13211.

Liang, Z.; Yu, D.; Yu, W.; Yao, W.; Zhang, Z.; Zhang, X.; and Yu, D. 2024. MathChat: Benchmarking Mathematical Reasoning and Instruction Following in Multi-Turn Interactions. *ArXiv*, abs/2405.19444.

Liu, J.; Huang, Z.; Xiao, T.; Sha, J.; Wu, J.; Liu, Q.; Wang, S.; and Chen, E. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 85693–85721. Curran Associates, Inc.

Lucy, L.; August, T.; Wang, R. E.; Soldaini, L.; Allison, C.; and Lo, K. 2024. Mathfish: Evaluating Language Model Math Reasoning via Grounding in Educational Curricula. arXiv:2408.04226.

Létourneau, A.; Martineau, M. D.; Charland, P.; Karran, J. A.; Boasen, J.; and Léger, P. M. 2025. A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1): 29.

Macina, J.; Daheim, N.; Chowdhury, S.; Sinha, T.; Kapur, M.; Gurevych, I.; and Sachan, M. 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*, 5602–5621. Singapore: Association for Computational Linguistics.

Macina, J.; Daheim, N.; Hakimi, I.; Kapur, M.; Gurevych, I.; and Sachan, M. 2025. MathTutorBench: A Benchmark for

Measuring Open-ended Pedagogical Capabilities of LLM Tutors. arXiv:2502.18940.

Maurya, K. K.; Srivatsa, K. A.; Petukhova, K.; and Kochmar, E. 2025. Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors. arXiv:2412.09416.

Mishra, S.; Poesia, G.; Mo, B.; and Goodman, N. D. 2024. MathCAMPS: Fine-grained Synthesis of Mathematical Problems From Human Curricula. *ArXiv*, abs/2407.00900.

Mitra, A.; Khanpour, H.; Rosset, C.; and Awadallah, A. 2024. Orca-Math: Unlocking the potential of SLMs in Grade School Math. arXiv:2402.14830.

OpenAI. 2024a. GPT-4 Technical Report. arXiv:2303.08774.

OpenAI. 2024b. GPT-4o System Card. arXiv:2410.21276.

Shao, Y.; Jiang, Y.; Kanell, T. A.; Xu, P.; Khattab, O.; and Lam, M. S. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. arXiv:2402.14207.

Shi, W.; Yu, A.; Fang, R.; Ren, H.; Wang, K.; Zhou, A.; Tian, C.; Fu, X.; Hu, Y.; Lu, Z.; Huang, L.; Liu, S.; Liu, R.; and Li, H. 2025. MathCanvas: Intrinsic Visual Chain-of-Thought for Multimodal Mathematical Reasoning. arXiv:2510.14958.

Shin, D.; Park, S.; Kim, E. H.; Kim, S.; Seo, J.; and Hong, H. 2022. Exploring the Effects of AI-assisted Emotional Support Processes in Online Mental Health Community. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, CHI '22. ACM.

Son, T. 2024. Intelligent Tutoring Systems in Mathematics Education: A Systematic Literature Review Using the Substitution, Augmentation, Modification, Redefinition Model. *Computers*, 13(10).

Team, L.; 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. 2024. LearnLM: Improving Gemini for Learning. arXiv:2412.16429.

Wang, K.; Ren, H.; Zhou, A.; Lu, Z.; Luo, S.; Shi, W.; Zhang, R.; Song, L.; Zhan, M.; and Li, H. 2023. MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning. arXiv:2310.03731.

Wang, R. E.; Ribeiro, A. T.; Robinson, C. D.; Loeb, S.; and Demszky, D. 2025. Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise. arXiv:2410.03017.

Wang, R. E.; Zhang, Q.; Robinson, C.; Loeb, S.; and Demszky, D. 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*. Association for Computational Linguistics.

Wu, S.; Xu, X.; Liu, R.; Liang, G.; Meng, H.; and He, B. 2022. An Intelligent Tutoring System for Math Word Problem Solving with Tutorial Solution Generation. In *2022 International Conference on Intelligent Education and Intelligent Research (IEIR)*, 183–188.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; Lu, K.; Xue, M.; Lin, R.; Liu, T.; Ren, X.; and Zhang, Z. 2024. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. arXiv:2409.12122.

Zeng, Z.; Chen, P.; Liu, S.; Jiang, H.; and Jia, J. 2024. MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation. arXiv:2312.17080.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. arXiv:2405.01470.

Zheng, C.; Zhang, Z.; Zhang, B.; Lin, R.; Lu, K.; Yu, B.; Liu, D.; Zhou, J.; and Lin, J. 2024. ProcessBench: Identifying Process Errors in Mathematical Reasoning. arXiv:2412.06559.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.