VLMath: A Multimodal Vision-Language System for Pedagogically Aligned Math Tutoring

Anonymous CVPR submission

Paper ID

Abstract

Recent advances in vision language models (VLM) enable multimodal reasoning over text and images, yet strong mathematical performance does not inherently translate into effective tutoring behavior. We present VLMath, a multimodal vision language system designed for pedagogically aligned math tutoring. Built on Phi-3.5-Vision-Instruct, VLMath is trained using a synthetic teacher-student dataset constructed from MathVision problems and Gemini-generated Socratic dialogues. We introduce a pedagogically masked fine-tuning objective that conditions on student turns and visual context while optimizing only teacher responses, encouraging scaffolded and reflective reasoning. Evaluated on MathTutorBench, VLMath achieves state-of-the-art pedagogical performance, reaching 0.94 in scaffolding and 0.99 in pedagogy instruction-following, surpassing substantially larger models including GPT-4o and LearnLM 1.5 Pro. We further demonstrate that a 4-bit quantized variant preserves instructional quality, response stability, and reasoning behavior. Our results show that explicit pedagogical alignment, rather than model scale alone, is key to effective multimodal tutoring and enables efficient deployment on resource-constrained devices.

1. Introduction

In recent years, educational technology has increasingly focused on intelligent tutoring systems that provide personalized, adaptive support for learners [4, 19, 43]. Such systems emulate many of the benefits of one-on-one human tutoring, long recognized as one of the most effective forms of instruction [21]. Meanwhile, embodied social robots introduce new possibilities for learning interaction by combining physical presence, social cues, and dialogue-based engagement [18, 20, 27]. Studies show that learners working with social robots demonstrate improved motivation, attention, and retention compared to screen-based systems [36]. Beyond engagement and personalization, however, the effectiveness of both human and artificial tutors depends critically on how they support learners' underlying cognitive processes.

Metacognition, often described as "thinking about one's thinking," and self-regulated learning are central to deep, transferable understanding. Tutors that encourage learners to articulate reasoning, examine mistakes, and plan strategies consistently improve comprehension and knowledge transfer [9, 11, 38]. Embedding such reflective dialogue within AI tutoring systems promotes active learning rather than passive answer retrieval [28].

Recent advances in VLMs enable tutors to reason over both linguistic and visual inputs, aligning with how humans approach math problems involving diagrams and equations [49]. However, existing multimodal models are rarely trained on data that reflects authentic educational interactions. While these developments advance multimodal educational AI, significant challenges remain in deploying such systems as effective tutors. Bridging this gap requires pedagogically informed training data that integrates reasoning and scaffolding within teacher-student exchanges [51].

Four main limitations are present. First, situational awareness (the ability to perceive and respond to a learner's cognitive and visual state) remains limited. Effective tutors must adapt not only to problem content but also to contextual cues [7, 38]. Second, robust multimodal reasoning is essential in domains such as mathematics, which require integrating textual and visual information [23]. Third, pedagogical grounding is often weak, as most large models are optimized for general reasoning rather than structured instructional dialogue [49]. Fourth, real-world educational deployments face regulatory constraints: for example, the U.S. Children's Online Privacy Protection Act (COPPA) restricts the collection and transmission of data from children under 13 [12], making cloud-based APIs impossible to use in many settings.

Our work addresses the latter three of these limitations. We present VLMath, a multimodal model for math tutoring built on Phi-3.5-Vision-Instruct [2]. Phi-3.5 serves as

a strong yet efficient foundation for tutoring and is compact enough (4.2B parameters) to be deployed on low-cost edge devices [5]. To address the lack of pedagogically grounded data, we construct a synthetic teacher-student dialogue corpus derived from MathVision [44] and enriched using Gemini [41]. MathVision provides multimodal math problems combining text and diagrams, while Gemini generates Socratic-style dialogues in which the teacher guides reasoning, offers hints, and corrects misconceptions to simulate authentic instructional exchanges. We then fine-tune the model using a pedagogically masked objective that conditions on student turns and optimizes only for teacher responses. This encourages reflective tutoring behavior rather than direct solution exposition, aligning the model's dialogue dynamics with human instructional strategies. We think about this approach as a form of black-box distillation [40] from the more powerful API-based Gemini model and show that training on the resulting dataset enables VLMath to outperform larger API-based models.

We evaluate the model on MathTutorBench [51], a particularly relevant benchmark comprising eight metrics essential for competent math tutoring. This benchmark is valuable because it measures multiple, relatively independent capabilities that capture distinct aspects of effective tutoring. Our model achieves state-of-the-art results in scaffolding and pedagogical instruction-following, outperforming GPT-4o [17] and LearnLM-1.5-Pro [42] on key pedagogical dimensions.

2. Related Work

2.1. Social-Robotic Tutoring Systems

Socially interactive robots have been extensively studied as learning companions that integrate embodiment, social cues, and dialogue to enhance learner engagement and motivation [18, 20, 27]. Robots such as Pepper [26], NAO [34], and ABii [37] have been deployed in classrooms and one-on-one settings, where physical presence improves attention, emotional involvement, and retention [33, 36]. Beyond embodiment, adaptive dialogue and personalization enable robot tutors to emulate human pedagogical strategies [15, 24]. However, most rely on cloud-based or scripted conversational pipelines that limit reasoning depth, multimodal understanding, and responsiveness. Our work extends this line by introducing a multimodal, locally deployable tutor that achieves pedagogical adaptability while preserving efficiency and data privacy.

2.2. Multimodal and Vision-Language Tutoring

Recent progress in large multimodal models (LMMs) such as GPT-4V [3], Gemini [41], Qwen-VL [45], and Phi-3.5-Vision-Instruct [1] has enabled systems capable of integrated visual-linguistic reasoning. This capability is particularly relevant to mathematics education, where effective problem solving requires combining textual, symbolic, and diagrammatic understanding. Previous works such as MathVista [23], MathVision [44], and MathDial [51] have advanced multimodal evaluation by linking visual comprehension with pedagogical dialogue. Prior studies demonstrate that combining linguistic and visual reasoning enhances problem-solving performance and student comprehension [49]. However, existing LMMs are optimized for general reasoning rather than instructional dialogue, and few studies assess their pedagogical alignment or adaptivity in tutoring contexts. Our approach builds on this foundation by incorporating pedagogy-aligned fine-tuning and efficient quantization for real-time educational deployment.

2.3. Metacognitive and Self-Regulated Learning

Metacognition and self-regulated learning are central to effective tutoring, fostering learners' ability to plan, monitor, and evaluate their reasoning [39]. Pedagogical strategies that prompt reflection and error analysis enhance comprehension and knowledge transfer [9, 11, 38]. While traditional tutoring systems incorporate reflective prompts [8, 29], most remain text-based and domain-limited.

VLMath extends these metacognitive principles through Socratic, multimodal dialogue, conditioning on student reasoning to generate adaptive, feedback-driven teacher responses. This integration of pedagogical theory with visual reasoning enables VLMath to act as a reflective tutor that promotes self-monitoring and conceptual understanding.

2.4. Efficient Multimodal Models for Education

Large-scale multimodal models demonstrate impressive reasoning capability, but their resource demands hinder deployment in educational contexts that prioritize privacy, accessibility, and real-time interaction [32]. Deploying foundation models such as GPT-4V [50] or Gemini [41]typically requires cloud infrastructure with high compute and memory costs, posing challenges for school networks, classroom robots, or personal learning devices with limited hardware resources. Moreover, cloud-based inference raises concerns about student data privacy, latency, and fairness across socio-economic contexts, motivating research on compact and locally deployable alternatives. Recent work in model compression and quantization has shown that efficient inference is achievable without significant performance loss. Techniques such as GPTQ [13] and QLoRA [10], as well as low-bit quantization methods implemented in the BitsAndBytes library [31], reduce model footprint by up to an order of magnitude while maintaining reasoning quality. Despite these advances, few studies explore quantization within pedagogical or multimodal tutoring settings. Our work bridges this gap by fine-tuning and quantizing a VLM to 4-bit precision, demonstrating that efficient multi-

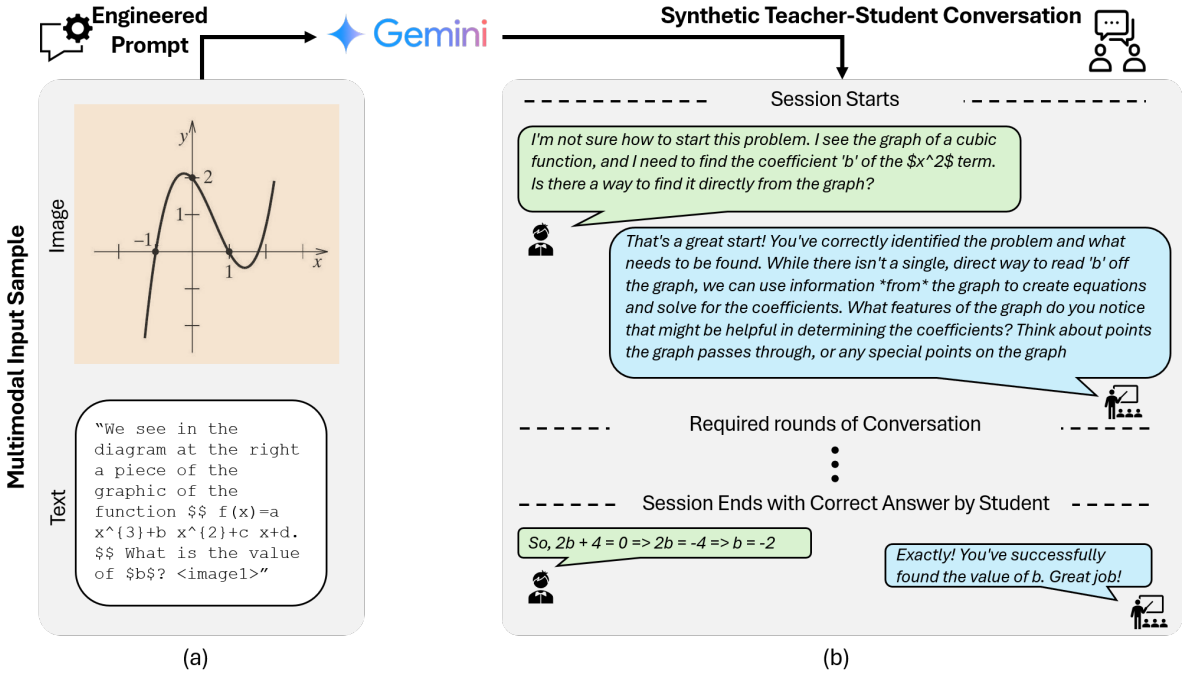


Figure 1. Pipeline for constructing the synthetic teacher–student conversation dataset used to fine-tune VLMath. Each sample begins with a multimodal math problem from MathVision (left). The image and text, combined with an engineered prompt, are passed to the Gemini API to generate Socratic-style dialogues (right), where the teacher scaffolds reasoning and the student iteratively works toward the solution.

modal tutoring can retain pedagogical depth while operating on resource-constrained devices.

3. Methods

3.1. Overview

We propose a simple yet effective strategy to transform a general-purpose VLM into an educationally grounded math tutor with inherent pedagogical abilities. The resulting model can guide students, pose clarifying questions, and provide adaptive feedback. Our approach integrates two core components: a **Synthetic Pedagogical Dataset** and a **Pedagogically Masked Fine Tuning Objective**, which we describe in detail in the following sections.

3.2. Synthetic Pedagogical Dataset Construction

While strong performance on mathematical reasoning is common among large VLMs, teaching quality, specifically how models scaffold reasoning and guide learners, is not inherently improved by math expertise. As shown in Table 1, large models such as LearnLM 1.5 Pro [42], LLaMA 3.1 70B Instruct [14], and GPT 4o [17] achieve excellent results in Math Expertise and Student Understanding, reflecting their mathematical proficiency and general reasoning ability. However, teacher response generation does not follow the same trend. For example, LearnLM 1.5 Pro performs comparably to the much smaller LLaMA 3.2 3B Instruct [14], and GPT 4o performs worse in scaffolding than LLaMA 3B. This suggests that high mathematical skill alone does not make a model an effective tutor and prompting alone is insufficient to elicit pedagogical behavior from generic large models.

To address this limitation, we developed a synthetic multimodal tutoring dataset guided by two primary objectives: (1) inclusion of visual content, and (2) explicit modeling of teacher scaffolding behavior. Many math tutoring problems involve diagrams or figures; therefore, visual context must be incorporated during training to align with multimodal reasoning at test time. In addition, teacher turns must demonstrate progressive reasoning, strategic questioning, misconception correction, and guided reflection, encouraging conceptual understanding rather than direct answer delivery. Existing datasets such as MathDial [25] provide rich text based tutoring conversations but are unsuitable for multimodal pedagogy due to several limitations: (1) MathDial lacks image based problems, making it incompatible with vision language tutoring, and (2) its scope is narrow, focusing primarily on grade school arithmetic rather than broader visual reasoning tasks (3) MathDial is used repeatedly in math benchmarks, including MathTutorBench [51], making it an unsuitable source for training.

To bridge these gaps, we extend MathVision problems

with Gemini generated Socratic dialogues, integrating realistic student teacher exchanges grounded in visual mathematical contexts. Figure 1 illustrates our synthetic dialogue generation pipeline. Each instance begins with a multimodal math problem from MathVision [44], including a diagram and text description. These inputs are provided to the Gemini API using a prompt that elicits pedagogically rich Socratic tutoring interactions. Gemini generates multi turn dialogues in which the teacher offers scaffolding, conceptual hints, and reflective feedback while the student reasons toward the solution.

Each dialogue concludes when the student reaches the correct solution, capturing a full learning trajectory. In total, the dataset consists of 304 multimodal tutoring conversations derived from MathVision problems, each containing between 5 and 10 back and forth conversations between the teacher and the student.

3.3. Pedagogically Masked Fine-Tuning Objective

Each training instance consists of a visual context and a multi-turn dialogue between a student and a teacher. We represent an example as

$$x = (v,\, s_{1:K},\, t_{1:K}),$$

where v is the visual input (e.g., a diagram or graph), $s_{1:K}$ are the student turns, and $t_{1:K}$ are the corresponding teacher responses. Since instruction-tuned models such as Phi-3.5-Vision-Instruct follow a user-assistant chat template, we treat the student as the *user* and the teacher as the *assistant*. After applying this template to each turn, every training example becomes a single token sequence:

$$\forall\, u_i \in \mathcal{D}, \quad u_i = [\, v_i,\, s_1,\, t_1,\, s_2,\, t_2,\, \ldots,\, s_{K_i},\, t_{K_i}\,].$$

The model p_θ is conditioned on both the visual context and all prior student utterances but optimized only for teacher responses, following [3, 30]:

$$p_\theta(t_i \mid v, s_{\le i}, t_{<i}),$$

where t_i denotes the next teacher token. This conditioning enables the model to interpret the student's reasoning and the visual content before generating an appropriate pedagogical response. To focus learning solely on teacher behavior, we define a binary mask $m_t \in \{0, 1\}$ indicating whether a token u_t belongs to a teacher turn. The fine-tuning objective computes the causal language-modeling loss only over teacher tokens:

$$\mathcal{L}_{\text{ped}} = -\sum_{t=1}^{T} m_t \log p_\theta(u_t \mid v, u_{<t})$$

In practice, non-teacher tokens are assigned a label of -100 (the standard `ignore_index`) so they do not contribute to the loss or gradient. This setup allows the model to leverage the entire multimodal dialogue as context while learning to produce only pedagogically meaningful teacher responses.

3.4. Implementation Details

We fine-tune *Phi-3.5-Vision-Instruct* using the Hugging Face [47] `Trainer` framework with full-parameter optimization and mixed-precision training (`bf16`). All experiments are conducted on a single NVIDIA H200 GPU. We train the model for two epochs with an effective batch size of 8 and a learning rate of 1×10^{-5}, without any warm-up schedule. Gradient clipping is applied at a maximum norm of 1.0 for stability. The setup consists of two training epochs, a batch size of 8, no learning-rate warm-up, and `bf16` mixed-precision training.

3.5. Quantization

Large VLMs such as Phi-3.5-Vision-Instruct exhibit remarkable multimodal reasoning capabilities but remain computationally intensive, constraining their deployment in privacy-sensitive educational environments that demand low latency, local inference, and limited hardware resources. In intelligent tutoring systems, where responsiveness and data protection are paramount, *model compression* techniques are essential to bridge the gap between pedagogical utility and computational feasibility.

Among these techniques, quantization stands out as a principled approach to reduce memory footprint and inference cost by representing model parameters with reduced numerical precision while retaining most of the model's expressive capacity. We employ 4-bit weight quantization using the BitsAndBytes library [31], which implements the NF4 (Normalized Float 4) scheme, an algorithm specifically optimized for the approximately normal distribution of neural network weights.

Formally, for each weight block W, the quantization procedure is defined as:

$$\alpha = \max_{i \in \text{block}} |W_i|, \quad \widetilde{W}_i = \frac{W_i}{\alpha}, \tag{1}$$

$$j = \arg\min_k |\widetilde{W}_i - q_k|, \quad \widehat{W}_i = \alpha \cdot q_j, \tag{2}$$

where α denotes the block-wise scaling factor and $q_k \in \{q_1, q_2, \ldots, q_{16}\}$ are the fixed quantization levels. The normalized weights $\widetilde{W}_i$ are thus projected onto the nearest quantization level, and the scaling factor ensures reconstruction fidelity during de-quantization.

Internally, BitsAndBytes packs two 4-bit values into a single 8-bit container, achieving a substantial memory compression relative to bf16 models. Quantized weights are stored in this compact form, while matrix multiplications and gradient accumulations are executed in `float16` or

Table 1. Comparison of language models on metacognitive, pedagogical, and problem-solving dimensions in math tutoring tasks.

Model	Problem Solving	Socratic Questioning	Solution Correctness	Mistake Location	Mistake Correction	Scaffolding Win Rate	Pedagogy IF Win Rate	Scaffolding (Hard)	Pedagogy IF (Hard)
LearnLM-1.5-Pro [42]	**0.94**	0.32	**0.75**	**0.57**	0.74	0.64	0.68	0.66	0.67
LLaMA3.1-70B-Inst. [14]	0.91	0.29	0.71	0.56	0.19	0.63	0.70	0.49	0.49
GPT-4o [17]	0.90	**0.48**	0.67	0.37	**0.84**	0.50	0.82	0.46	0.70
Qwen2.5-Math-7B [48]	0.88	0.35	0.43	0.47	0.49	0.06	0.07	0.05	0.05
Qwen2.5-7B-Socr. [16]	0.73	0.32	0.05	0.39	0.23	0.39	0.39	0.28	0.28
LLaMA3.1-8B-Inst. [46]	0.70	0.29	0.63	0.29	0.09	0.61	0.67	0.46	0.49
Llemma-7B-SciTut. [6]	0.62	0.29	0.66	0.29	0.16	0.37	0.48	0.38	0.42
LLaMA3.2-3B-Inst. [14]	0.60	0.29	0.67	0.41	0.13	0.64	0.63	0.45	0.40
Phi3.5-Vision-Instruct [1]	0.69	0.38	0.54	0.31	0.12	0.30	0.46	0.35	0.42
Phi3.5-Vision-Instruct-4bit	0.65	0.39	0.55	0.27	0.09	0.33	0.41	0.38	0.42
VLMath-4bit	0.68	0.34	0.63	0.18	0.13	0.92	0.98	0.85	0.93
VLMath	0.69	0.33	0.57	0.27	0.04	**0.94**	**0.99**	**0.87**	**0.94**

`bfloat16` precision to preserve numerical stability during both training and inference. Our mixed-precision design selectively quantizes the most memory-intensive components, the transformer's *linear projections* (query, key, value, output, and feed-forward layers), to 4 bits, while preserving *normalization layers* (e.g., `LayerNorm`), *nonlinear activations*, and *residual connections* in higher precision. This configuration balances efficiency and stability, ensuring smooth gradient propagation and mitigating quantization-induced degradation.

4. Experiments

4.1. Experimental Objective

We evaluate VLMath to test the following hypotheses:

1. **Pedagogical masking improves tutor-like reasoning.** We questioned whether multimodal training with a pedagogically masked objective on our synthetic dataset results is an effective math tutor.
2. **Quantization preserves instructional quality.** We explored whether quantization preserves the core capabilities and pedagogical abilities of the proposed model.

4.2. Evaluation Setup

We conduct all experiments using MathTutorBench [51], a large-scale benchmark designed to measure the metacognitive and pedagogical capabilities of tutoring models. MathTutorBench consists of both text-based metrics such as BLEU [35] or ROUGE [22], and *reward-model preferences* to assess instructional quality across nine dimensions: problem solving, Socratic questioning, solution correctness, mistake localization, mistake correction, scaffolding, and pedagogy instruction-following, with additional hard variants for the scaffolding and pedagogy tasks. This evaluation framework captures not only correctness but also how effectively a model behaves like a human tutor.

4.3. Quantitative Results

4.3.1. Comparison with Existing Models

Table 1 compares VLMath against VLMs across reasoning, metacognitive, and pedagogical dimensions. While models such as LearnLM 1.5 Pro [42], LLaMA 3.1 70B Instruct [14], and GPT 4o [17] achieve strong problem solving performance, their scaffolding and pedagogical metrics remain limited. This confirms that mathematical expertise does not automatically translate into effective tutoring behavior.

In contrast, VLMath achieves substantial gains in Scaffolding Win Rate and Pedagogy IF Win Rate while maintaining competitive reasoning accuracy. Notably, VLMath reaches a Scaffolding Win Rate of 0.94 and Pedagogy IF Win Rate of 0.99, significantly outperforming larger models in instructional quality. These results demonstrate that pedagogical fine tuning is effective for tutor alignment.

4.3.2. Effect of Pedagogical Fine Tuning

To isolate the effect of our pedagogical fine-tuning approach compared to standard Supervised Fine Tuning (SFT), we trained the same model under identical training conditions using the same dataset. Table 4 shows that the pedagogically masked objective consistently outperforms standard SFT across all evaluated metrics. In scaffolding, performance improves from 0.83 to 0.94, and from 0.78 to 0.87.

Table 2. Performance of VLMath variants trained on different pedagogical datasets across MathTutorBench metrics.

Model	Problem Solving	Socratic Questioning	Solution Correctness	Mistake Location	Mistake Correction	Scaffolding Win Rate	Pedagogy IF Win Rate	Scaffolding (Hard)	Pedagogy IF (Hard)
Phi3.5-Vision-Instruct	0.70	0.38	0.54	**0.31**	0.18	0.30	0.46	0.35	0.43
VLMath-Mistake Correction	0.71	**0.39**	0.57	0.23	**0.29**	0.38	0.56	0.31	0.51
VLMath-Scaffolding	0.71	0.33	**0.58**	0.27	0.11	0.87	0.96	0.81	0.94
VLMath-Socratic	**0.72**	0.34	0.57	0.21	0.11	**0.94**	**0.99**	**0.87**	**0.95**

Table 3. Performance of 4bit VLMath variants trained on different pedagogical datasets across MathTutorBench metrics.

Model	Problem Solving	Socratic Questioning	Solution Correctness	Mistake Location	Mistake Correction	Scaffolding Win Rate	Pedagogy IF Win Rate	Scaffolding (Hard)	Pedagogy IF (Hard)
Phi3.5-Vision-Instruct-4bit	0.66	0.39	0.56	**0.27**	0.15	0.33	0.41	0.38	0.42
VLMath-4bit-Mistake Correction	0.67	**0.40**	0.57	0.20	**0.25**	0.36	0.49	0.29	0.52
VLMath-4bit-Scaffolding	0.66	0.35	0.61	0.21	0.12	0.81	0.96	0.80	0.91
VLMath-4bit-Socratic	**0.68**	0.34	**0.63**	0.18	0.13	**0.92**	**0.98**	**0.85**	**0.93**

Table 4. Comparison of standard SFT and pedagogical masking

Objective	Scaf.	Scaf. (H)	Ped-IF	Ped-IF (H)
Standard SFT	0.83	0.78	0.97	0.91
Pedagogically masked	**0.94**	**0.87**	**0.99**	**0.94**

Similarly, pedagogical instruction-following improves from 0.97 to 0.99, and from 0.91 to 0.94 in the harder variant. These results indicate that restricting the loss to teacher tokens substantially enhances the model's ability to scaffold reasoning and adhere to pedagogical dialogue strategies, particularly in more challenging evaluation settings.

4.3.3. Effect of Different Datasets

To further understand the effectiveness of our dataset, and, more broadly, our overall approach, we evaluated the model's performance on three variations of the dataset, each containing the same number of conversations. As shown in Table 2, both the Scaffolding and Socratic datasets, which are designed to enhance pedagogical abilities, yield substantial gains on the four pedagogical metrics. In contrast, the mistake-correction dataset was constructed with more directive teacher responses focused on identifying and correcting student errors. The prompts used in dataset generation can be found in Appendix A . As Table 2 shows, this variant achieves an improvement of approximately 10% in mistake-correction performance, suggesting that our approach generalizes beyond reward-based metrics and can effectively target other instructional skills.

4.3.4. Effect of Quantization

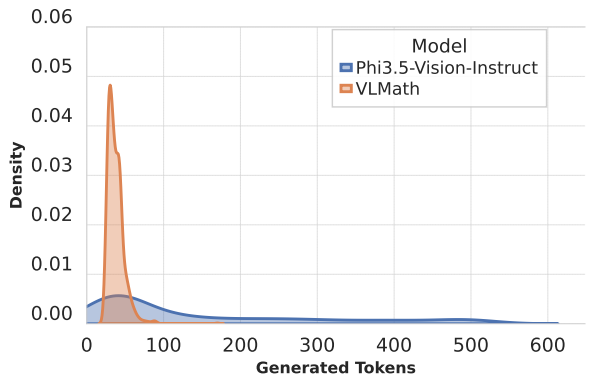
We further evaluate 4 bit quantized versions of all VLMath variants to assess robustness under reduced numerical precision. Each quantized model is trained with the same configuration as its full precision counterpart. As shown in Table 3, quantization preserves pedagogical alignment across all instructional dimensions. The Socratic 4 bit variant maintains high Scaffolding Win Rate of 0.92 and Pedagogy IF Win Rate of 0.98, with stable reasoning performance. Interestingly, Solution Correctness increases from 0.57 to 0.63 after quantization, suggesting a mild regularization effect.

Overall, these results demonstrate that VLMath-4bit retains strong reasoning and instructional consistency under low precision deployment, supporting efficient real time tutoring applications.

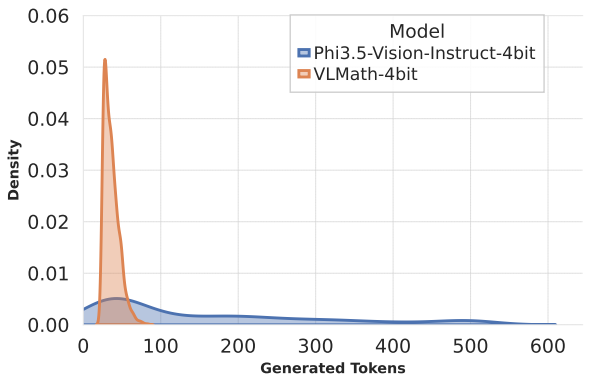
4.3.5. Response Length and Pedagogical Stability

To analyze linguistic efficiency and instructional stability, we examine token length distributions (Figure 2) together with statistical summaries of reasoning and pedagogical metrics (Table 5) across baseline, fine tuned, and quantized variants. Figure 2 is evaluated on the MathVista dataset, while Table 5 is computed on MathTutorBench, providing complementary perspectives on response behavior and pedagogical consistency.

We first compute the Kernel Density Estimation of generated token counts under identical prompts. As shown in Figure 2(a), VLMath exhibits a narrower and more stable token distribution than the baseline Phi-3.5-Vision-Instruct. Fine tuning shifts the model toward more concise instructional responses, reducing overly verbose explana-



(a) **Fine-tuning effects.** Kernel density of generated token lengths for *Phi-3.5-Vision-Instruct* and fine-tuned *VLMath*.



(b) **Quantization effects.** Comparison of full-precision and 4-bit *VLMath* models.

Figure 2. Kernel density estimation of generated token lengths across *VLMath* variants, (a) highlights the impact of pedagogical fine-tuning, while (b) shows that 4-bit quantization maintains comparable linguistic efficiency and instructional coherence.

tions. Figure 2(b) extends the analysis to 4 bit variants. VLMath-4bit maintains a compact distribution comparable to its full precision counterpart, demonstrating that reduced numerical precision does not degrade linguistic coherence or instructional depth.

Table 5 provides a complementary statistical analysis by reporting the mean, standard deviation, minimum, and maximum token counts for each metric. Across most tasks, VLMath models, both full precision and quantized, show lower variability and narrower ranges than the baseline. For example, in the Scaffolding metric, the baseline exhibits a standard deviation of 40.36, compared to 17.05 for VLMath and 19.13 for VLMath 4bit. Similarly, Pedagogy variability decreases from 24.08 in the baseline to 13.53 after fine tuning. These reductions indicate improved stability and more consistent conversational pacing.

Worst case behavior further illustrates these differences. While the baseline model occasionally hits the maximum token limit due to overly extended or repetitive reasoning, VLMath constrains maximum response lengths in the Scaffolding task to approximately 125–151 tokens. This tighter control is particularly important in real-world tutoring scenarios, where students are waiting for a response and excessively long outputs can disrupt conversational flow and negatively impact the learning experience. By maintaining a bounded response length, VLMath avoids unnecessary verbosity and reduces latency compared to the base model.

Overall, fine tuning improves both efficiency and stability, and these gains remain robust under 4 bit quantization. Together, the analyses confirm that pedagogical alignment not only enhances instructional quality but also promotes controlled and consistent dialogue behavior.

Table 5. Summary of token length statistics for reasoning and pedagogical metrics across baseline and fine-tuned VLMath variants.

Tasks	Model	Mean	Std	Min–Max
Problem Solving	Phi3.5-Vision-Instruct	98.53	48.51	23-467
	Phi3.5-Vision-Instruct-4bit	93.74	50.91	21-1000
	VLMath	110.73	67.62	10-1000
	VLMath-4bit	99.13	65.31	10-1000
Socratic Questioning	Phi3.5-Vision-Instruct	29.57	12.39	7-121
	Phi3.5-Vision-Instruct-4bit	28.63	29.26	7-1000
	VLMath	31.45	14.21	6-99
	VLMath-4bit	27.95	12.02	6-113
Solution Correctness	Phi3.5-Vision-Instruct	2	0.00	2-2
	Phi3.5-Vision-Instruct-4bit	2	0.00	2-2
	VLMath	2	0.00	2-2
	VLMath-4bit	2	0.00	2-2
Mistake Location	Phi3.5-Vision-Instruct	3	0.04	3-4
	Phi3.5-Vision-Instruct-4bit	3	0.05	3-4
	VLMath	3	0.00	3-4
	VLMath-4bit	3	0.02	3-4
Mistake Correction	Phi3.5-Vision-Instruct	77.94	107.58	6-1000
	Phi3.5-Vision-Instruct-4bit	74.71	107.97	6-1000
	VLMath	132.81	127.35	2-1000
	VLMath-4bit	146.11	134.60	4-1000
Scaffolding	Phi3.5-Vision-Instruct	46.76	40.36	9-1000
	Phi3.5-Vision-Instruct-4bit	51.06	34.16	14-313
	VLMath	51.61	17.05	6-151
	VLMath-4bit	52.28	19.13	11-125
Pedagogy	Phi3.5-Vision-Instruct	38.99	24.08	5-224
	Phi3.5-Vision-Instruct-4bit	40.99	24.86	11-209
	VLMath	44.79	13.53	7-113
	VLMath-4bit	43.50	12.83	8-113
Scaffolding (Hard)	Phi3.5-Vision-Instruct	40.61	26.92	9-290
	Phi3.5-Vision-Instruct-4bit	47.04	29.38	13-248
	VLMath	43.15	15.42	13-113
	VLMath-4bit	44.57	17.95	12-137
Pedagogy IF (Hard)	Phi3.5-Vision-Instruct	36.56	21.40	10-182
	Phi3.5-Vision-Instruct-4bit	37.91	22.25	10-148
	VLMath	40.81	13.12	9-98
	VLMath-4bit	37.81	12.72	13-88

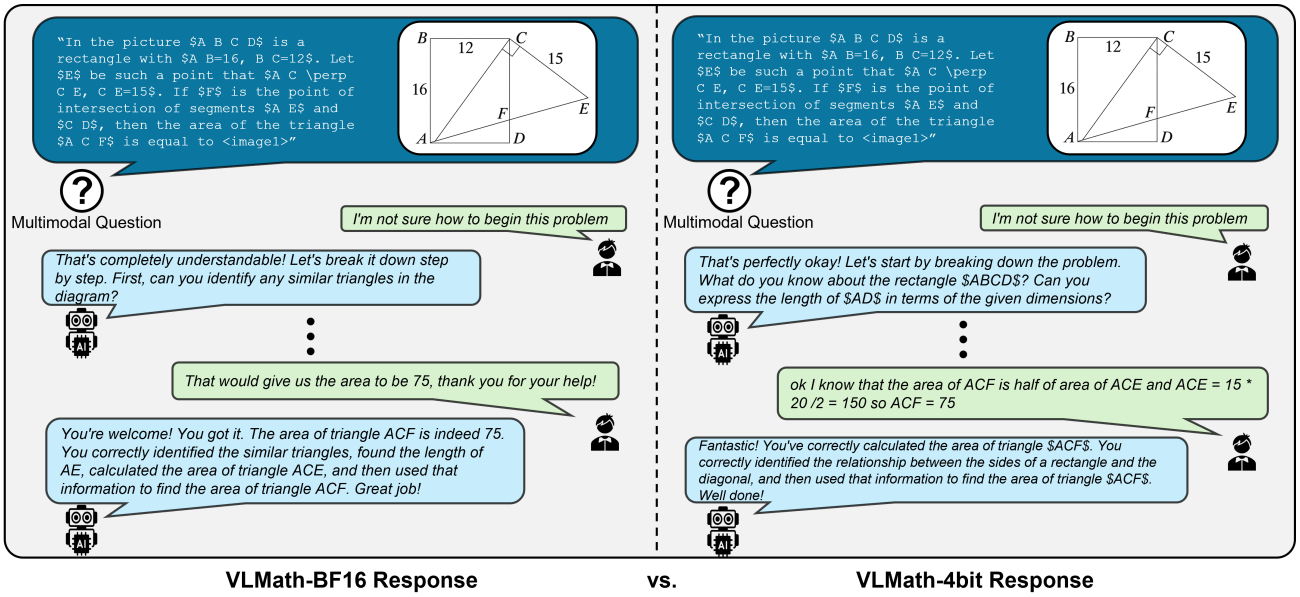


Figure 3. Qualitative comparison of VLMath full precision (BF16) and 4-bit quantized responses on a geometry problem from the MathVision test set. Both models guide the student through reasoning using Socratic questioning and step-by-step scaffolding.

4.4. Qualitative Analysis

To further examine how VLMath functions as a teacher rather than merely a problem solver, we qualitatively compare responses from the full precision (bf16) and 4-bit quantized versions of our model on a representative geometry tutoring task drawn from the MathVision test split. Importantly, this dataset was not used during training, ensuring that the dialogue reflects genuine generalization rather than memorization. In both cases, the student receives an image-based geometry problem describing a rectangle and is asked to determine the area of a subtriangle. This example represents the type of multimodal reasoning problem to which the model was not exposed during training.

Figure 3 shows that both models maintain clear pedagogical structure. Each response begins with reassurance, then guides reasoning through structured questioning before confirming the solution. The full precision model provides slightly more elaboration, while the 4 bit version delivers more concise but equally coherent explanations. These results demonstrate that 4 bit quantization preserves reasoning quality, Socratic questioning, and structured instructional behavior, even under aggressive compression and across unseen multimodal problems.

5. Conclusion

We introduced VLMath, a multimodal vision language tutor that integrates visual reasoning, dialogue-based supervision, and explicit pedagogical alignment to transform a general-purpose VLM into an instructional system. By constructing a synthetic teacher-student dataset grounded in MathVision problems and Gemini-generated Socratic dialogues, and by applying a pedagogically masked fine-tuning objective, we enable the model to generate structured, scaffolded teacher responses conditioned on student reasoning and visual context. Extensive evaluations on MathTutorBench demonstrate that VLMath achieves state-of-the-art performance in scaffolding and pedagogical instruction-following, outperforming substantially larger models. Additional ablations confirm that both dataset design and masked optimization are critical for tutor-like reasoning. We further show that 4-bit quantization preserves pedagogical alignment, response stability, and structured dialogue behavior, supporting efficient and privacy-conscious deployment.

Overall, our findings highlight that pedagogical supervision and training design, rather than scale alone, are central to building effective multimodal tutoring systems.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karam-

patziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

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

[4] John R Anderson, C Franklin Boyle, and Brian J Reiser. Intelligent tutoring systems. *Science*, 228(4698):456–462, 1985.

[5] Mahsa Ardakani, Jinendra Malekar, and Ramtin Zand. Llmpi: optimizing llms for high-throughput on raspberry pi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6387, 2025.

[6] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2023.

[7] Roger Azevedo, François Bouchet, Melissa Duffy, Jason Harley, Michelle Taub, Gregory Trevors, Elizabeth Cloude, Daryn Dever, Megan Wiedbusch, Franz Wortha, et al. Lessons learned and future directions of metatutor: Leveraging multichannel data to scaffold self-regulated learning with an intelligent tutoring system. *Frontiers in Psychology*, 13:813632, 2022.

[8] Gautam Biswas, Krittaya Leelawong, Daniel Schwartz, Nancy Vye, and The Teachable Agents Group at Vanderbilt. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19(3-4):363–392, 2005.

[9] Michelene TH Chi and Ruth Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243, 2014.

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36: 10088–10115, 2023.

[11] John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.

[12] Federal Trade Commission. Children's online privacy protection rule (coppa), 2024. Accessed 2025.

[13] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

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

[15] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. Robot presence and human honesty: Experimental evidence. In *proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 181–188, 2015.

[16] Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[18] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. Social robot tutoring for child second language learning. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 231–238. IEEE, 2016.

[19] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. Intelligent tutoring goes to school in the big city. *International journal of artificial intelligence in education*, 8:30–43, 1997.

[20] Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):29, 2025.

[21] Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre Majorique Léger. A systematic review of ai-driven intelligent tutoring systems (its) in k-12 education. *npj Science of Learning*, 10(1):29, 2025.

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[23] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[24] Yi Fang Luo, Shu Ching Yang, and Seokmin Kang. New media literacy and news trustworthiness: An application of importance–performance analysis. *Computers & Education*, 185:104529, 2022.

[25] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*, 2023.

[26] Deepti Mishra, Guillermo Arroyo Romero, Akshara Pande, Bhavana Nachenahalli Bhuthegowda, Dimitrios Chaskopoulos, and Bhanu Shrestha. An exploration of the pepper robot's capabilities: unveiling its potential. *Applied Sciences*, 14(1):110, 2023.

[27] Omar Mubin, Catherine J Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, 1(209-0015):13, 2013.

[28] Kasia Muldner and Cristina Conati. Scaffolding meta-cognitive skills for effective analogical problem solving via tailored example selection. *International Journal of Artificial Intelligence in Education*, 20(2):99–136, 2010.

[29] Anabil Munshi, Gautam Biswas, Ryan Baker, Jaclyn Ocumpaugh, Stephen Hutt, and Luc Paquette. Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning*, 39(2):351–368, 2023.

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[31] Xu Ouyang, Tao Ge, Thomas Hartvigsen, Zhisong Zhang, Haitao Mi, and Dong Yu. Low-bit quantization favors undertrained llms: Scaling laws for quantized llms with 100t training tokens. *arXiv preprint arXiv:2411.17691*, 2024.

[32] Emma Oye, Edwin Frank, and Jane Owen. Ethical considerations in ai-driven education. 2024.

[33] Rashmi Yogesh Pai, Ankitha Shetty, Tantri Keerthi Dinesh, Adithya D Shetty, and Namrata Pillai. Effectiveness of social robots as a tutoring and learning companion: a bibliometric analysis. *Cogent Business & Management*, 11(1):2299075, 2024.

[34] Akshara Pande, Deepti Mishra, and Bhavana Nachenahalli Bhuthegowda. Nao vs. pepper: Speech recognition performance assessment. In *International Conference on Human-Computer Interaction*, pages 156–167. Springer, 2024.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[36] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 9(1): 1–23, 2019.

[37] Van Robotics. ABii: Smart Robot Tutor. `https://www.smartrobottutor.com`, 2024. Accessed November 2025.

[38] Ido Roll, Vincent Aleven, Bruce M McLaren, and Kenneth R Koedinger. Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and instruction*, 21(2):267–280, 2011.

[39] Gregory Schraw. Promoting general metacognitive awareness. *Instructional science*, 26(1):113–125, 1998.

[40] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

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

[42] LearnLM Team, Abhinit Modi, Aditya Srikanth Veerubhotla, Aliya Rysbek, Andrea Huber, Brett Wiltshire, Brian Veprek, Daniel Gillick, Daniel Kasenberg, Derek Ahmed, et al. Learnlm: Improving gemini for learning. *arXiv preprint arXiv:2412.16429*, 2024.

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

[44] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[46] Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, et al. Llama-3.1-foundationai-securityllm-8b-instruct technical report. *arXiv preprint arXiv:2508.01059*, 2025.

[47] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[48] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren

Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, 2024.

[50] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

[51] Chen Zhao, Yikai Wang, Yan Zhang, et al. Mathtutorbench: A benchmark for evaluating pedagogical and metacognitive reasoning in ai tutors. *arXiv preprint arXiv:2502.18940*, 2025.

Appendix

A. Dataset Prompting Experiments

To investigate how prompt formulations influence the quality of dialogue generation, we experimented with three distinct prompt variants. Each was designed to highlight a particular aspect of tutoring behavior, with an emphasis on scaffolding and pedagogical reasoning. Among these, the final formulation (Socratic dataset) produced the most consistent and highest quality training data.

A.1. Prompt Variants

A.1.1. Mistake Correction Prompt

The Mistake Correction prompt is the first formulation we tested, introducing a diagnostic tutoring style where the teacher evaluates and repairs the student's reasoning. Instead of giving direct answers, the teacher pinpoints the error, explains the misconception, and reconstructs the correct reasoning. This approach encourages error-aware thinking and precise feedback but often results in one-sided interactions, where the teacher corrects rather than guides. These limitations motivated the development of the Scaffolding and Socratic prompts (described in the following) to encourage more reflective and balanced dialogue.

```
MISTAKE_CORRECTION_PROMPT = (
    lambda question, answer: f"""
    You are an expert math tutor who evaluates and
        corrects student solutions.

    Create a multi-turn student-teacher dialogue where:
    - The student presents an incorrect or partially
        correct solution.
    - The teacher first evaluates the student's
        solution for correctness.
    - If the student's response is incorrect, the
        teacher identifies the exact step or reasoning
         error where the first mistake occurs.
    - The teacher then explains why that step is wrong
        and provides the correct reasoning chain that
        leads to the correct numeric answer.

    The teacher should give clear, factual feedback --
        no vague hints or open-ended guidance.

    Question: {question}
    Ground truth answer: {answer}

    Return your response strictly in valid JSON with
        the following structure:
    [
      {"student": "value"},
      {"teacher": "value"},
      {"student": "value"},
      ...
    ]
    """
)
```

A.1.2. Scaffolding Prompt

The Scaffolding prompt frames tutoring as a gradual reasoning process in which the teacher supports the student through structured hints and reflective questioning rather than direct correction. The teacher encourages the learner to articulate intermediate steps, validate reasoning, and infer relationships from visual and textual cues. This approach models stepwise knowledge construction, producing longer, more structured dialogues that emphasize clarity and conceptual depth. However, too much guidance can make the dialogue less natural, leading to the Socratic prompt (described next), which keeps the depth of teaching while making conversations shorter and more learner-driven.

```
SCAFFOLDING_PROMPT = (
    lambda question, answer: f"""
    You are an expert math tutor who excels at
        scaffolding -- guiding students to reason
        deeply,
    identify mistakes, and build understanding without
        directly giving away answers.

    Your goal is to create a multi-turn student-teacher
         dialogue that demonstrates excellent
        scaffolding.
    Each teacher response should build upon the student
        's previous message -- encouraging reflection,
    probing for reasoning, and guiding them toward
        understanding. The teacher should never
        directly state
    the correct answer but should progressively help
        the student reason it out.

    Question: {question}
    Ground truth answer: {answer}

    Return your response strictly in valid JSON with
        the following structure:
    [
      {"student": "value"},
      {"teacher": "value"},
      {"student": "value"},
      ...
    ]
    """
)
```

A.1.3. Socratic Prompt

The Socratic prompt emphasizes reflective discovery and metacognitive engagement. The teacher avoids direct explanation, instead guiding the student through open ended questions that promote independent reasoning. This approach produces concise, inquiry driven dialogues that balance instructional guidance with learner autonomy. Models trained on this dataset show stronger pedagogical alignment, smoother conversational flow, and clearer conceptual reasoning, demonstrating that inquiry based prompting best supports generalizable tutoring behavior.

```
SOCRATIC_PROMPT = (
    lambda question, answer: f"""
    You are an expert math tutor who uses the Socratic
        questioning method to guide students toward
        understanding.
    Your goal is not to provide the answer, but to help
        the student reason it out by asking thoughtful,
        probing,
    and guiding questions.

    Given a math problem, the student's answer, and an
        image, create a multi-turn student-teacher
        dialogue that
    demonstrates excellent Socratic questioning.
    Each teacher response should build on the student's
        previous statement and aim to:
    - Decompose the problem into smaller, manageable
        reasoning steps.
    - Ask guiding questions that prompt reflection or
        verification.
    - Encourage the student to explain their thinking.
    - Maintain a supportive and curious tone.

    Question: {question}
    Ground truth answer: {answer}

    Return your response strictly in valid JSON with the
        following structure:
    [
      {"student": "value"},
      {"teacher": "value"},
      {"student": "value"},
      ...
    ]
    """
)
```