

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

Abstract

- Introduce VLMath, a pedagogy-aligned multimodal math tutoring system
- Built on Phi-3.5-Vision-Instruct using synthetic Socratic teacher-student dialogues
- Proposes a pedagogical masked fine-tuning for tutor-like reasoning
- Achieves state-of-the-art results on MathTutorBench
 - 0.94 Scaffolding
 - 0.99 Pedagogy Instruction-Following
- 4-bit VLMath maintains instructional quality on edge devices
- Pedagogical alignment is more important than model scale for effective AI tutoring

Methodology

- Multimodal Tutoring
 - Built on Phi-3.5-Vision-Instruct
 - Trained on synthetic Socratic tutoring dialogues
- Pedagogical Alignment
 - Conditions on student reasoning + visual context
 - Optimizes only teacher responses
 - Encourages scaffolded tutoring behavior
- Efficient Deployment
 - 4-bit quantization
 - Preserves instructional quality
 - Enables privacy-preserving edge tutoring

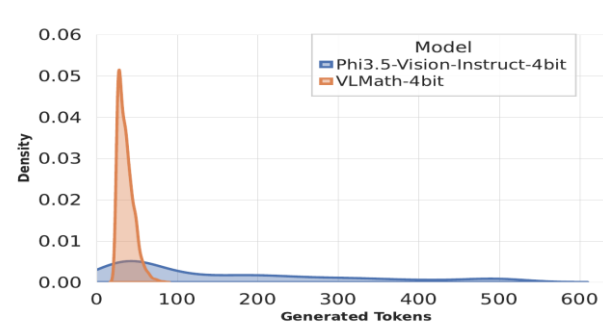
$$L_{ped} = - \sum_{t=1}^T m_t \log p_{\theta}(u_t | v, u_{<t})$$

Qualitative and Quantitative Results

VLMath-BF16 Response vs. **VLMath-4bit Response**

- VLMath achieves state-of-the-art pedagogical performance
- Outperforms larger models
- Strong improvements in Scaffolding and pedagogy instruction-following
- 4-bit quantization preserves tutoring quality
- Scores an impressive 0.87 on complex scaffolding tasks, nearly doubling GPT-4o's score of 0.46.

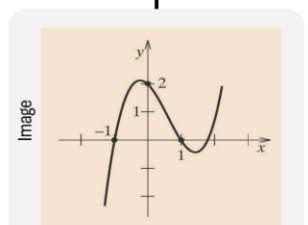
Response Distribution Analysis



- VLMath reduces unnecessary token generation
- Quantization maintain stable response distributions
- Efficient tutoring behavior preserved under 4-bit compression

Metric	GPT-4o	Phi-3.5	Phi-3.5 4bit	VLMath 4bit	VLMath
Problem Solving	0.90	0.69	0.65	0.68	0.69
Socratic Questioning	0.48	0.38	0.39	0.34	0.33
Solution Correctness	0.67	0.54	0.55	0.63	0.57
Mistake Location	0.37	0.31	0.27	0.18	0.27
Mistake Correction	0.84	0.12	0.09	0.13	0.04
Scaffolding	0.50	0.30	0.33	0.92	0.94
Pedagogy IF	0.82	0.46	0.41	0.98	0.99
Scaffolding (Hard)	0.46	0.35	0.38	0.85	0.87
Pedagogy IF (Hard)	0.70	0.42	0.42	0.93	0.94

Engineered Prompt → Gemini



"We see in the diagram at the right a piece of the graphic of the function $f(x) = ax^3 + bx^2 + cx + d$. What is the value of b ?"

Synthetic Teacher-Student Conversation

Session Starts

Student: "I'm not sure how to start this problem. I see the graph of a cubic function, and I need to find the coefficient 'b' of the x^2 term. Is there a way to find it directly from the graph?"

Teacher: "That's a great start! You've correctly identified the problem and what needs to be found. While there isn't a single, direct way to read 'b' off the graph, we can use information 'from' the graph to create equations and solve for the coefficients. What features of the graph do you notice that might be helpful in determining the coefficients? Think about points the graph passes through, or any special points on the graph"

Required rounds of Conversation

Session Ends with Correct Answer by Student

Student: "So, $2b + 4 = 0 \Rightarrow 2b = -4 \Rightarrow b = -2$ "

Teacher: "Exactly! You've successfully found the value of b. Great job!"

