

Scorpio: A Verifiable Framework for Enforcing Socratic Scaffolding in Physics LLMs Beyond Fine-Tuning

Rushil Mahadevu

Sage Ridge School

Reno, NV, USA

rushil.mahadevu@gmail.com

Abstract—While Large Language Models (LLMs) have demonstrated significant potential in STEM education, their tendency to provide direct solutions often undermines the learning process. This paper presents Scorpio, a framework that utilizes “Constraint Engineering” to transform a general-purpose LLM into a specialized, verifiable physics tutor. By implementing a layered architecture of inference-time rules, we enforce Socratic scaffolding without the need for expensive fine-tuning. Our results from a 125-response ablation study demonstrate that Scorpio successfully eliminates direct answer delivery (0% DAR), achieving a peak pedagogical quality score of 4.62. Independent expert validation (Ph.D. physics educator) confirmed this success, with the expert independently ranking the FULL constraint system as the highest performing (Expert: 3.83 vs 3.16 baseline) and confirming a significant +0.67 point improvement in pedagogical quality. This work demonstrates that inference-time constraint layering can meaningfully shift LLM tutoring behavior toward Socratic scaffolding, proving that modular constraints can effectively manage complex STEM pedagogy at scale.

Index Terms—AI in Education, Constraint Engineering, Inference-Time Constraints, Physics Education, Socratic Scaffolding, Verifiable AI, Large Language Models

I. INTRODUCTION

The integration of Large Language Models (LLMs) in STEM education highlights a critical discrepancy: most foundational models (e.g., GPT-4, Gemini, LLaMA) are optimized for general-purpose helpfulness and response fluency in deployment settings [9], [10], whereas effective learning requires “**productive struggle**” [11]. Research in Physics Education Research (PER) suggests that when students receive full solutions immediately, they bypass the critical reasoning steps necessary to build a robust mental model of physical systems [2], [12]. When an AI system acts as an “answer engine,” it removes the **pedagogical scaffolding** required for true conceptual mastery, a concept rooted in constructivist learning theory [8], [13].

To address this, current methods often utilize fine-tuning, including parameter-efficient approaches such as LoRA [14]. However, fine-tuning presents several hurdles: it is computationally expensive [15], difficult to adapt across different subjects, and often lacks transparency in how specific behaviors are enforced. Furthermore, generative models are prone to hallucinations, which poses risks in educational contexts where factual accuracy is paramount [16], [17].

This research presents **Scorpio**, a physics education platform that utilizes **Constraint Engineering**—a derivative of prompt engineering and constitutional AI [4], [18]—to guide model behavior at inference-time. Instead of modifying the model’s weights, Scorpio applies a layered system of rules around a lightweight model (Gemini 2.5 Flash). These constraints are designed to enforce a **verifiable Socratic scaffolding** style; specifically, the AI guides students through reasoning steps and provides hints rather than complete solutions. Unlike industry-standard Reinforcement Learning from Human Feedback (RLHF) [5], which lacks interpretability, Scorpio’s rule-based inference constraints offer **superior auditability** for educational applications where teachers must verify AI decisions. Throughout this paper, “verifiable” denotes empirical auditability of rule compliance rather than formal proof guarantees.

The primary objectives of this study are:

- **Framework Development:** Formalizing a four-layer architecture (Domain, Pedagogical, Notation, and Socratic) to structure verifiable AI behavior at the prompt level.
- **Ablation Analysis:** Systematically testing how each constraint layer impacts the AI’s ability to maintain a scaffolded dialogue without revealing final answers.
- **Performance Optimization:** Evaluating whether a constrained, efficient model can achieve high-level pedagogical performance comparable to larger, unconstrained systems.

II. RELATED WORK

A. AI Tutoring Systems and Learning Theory

The evolution of AI in education has transitioned from rule-based Intelligent Tutoring Systems (ITS) developed in the 1980s and 90s [6], [19] to modern generative Large Language Models. Classic ITS architectures, such as the ACT-R cognitive tutors [20] and AutoTutor [21], relied on manually encoded production rules to model student states. While effective, these systems were brittle and expensive to author.

Recent work comparing ChatGPT-generated hints to human tutor hints shows measurable differences in learning gains, suggesting that unconstrained LLM tutoring behavior may not reliably match expert pedagogy [7].

In contrast, LLMs offer flexibility but often fail the “2-Sigma” challenge [1] by providing immediate answers. This creates a ‘lazy-loop’ where students use AI to bypass cognitive load. Scorpio addresses this through “Desirable Difficulties” [3] and “Productive Failure” [11], ensuring that the AI maintains the role of a facilitator. Research in PER, specifically the work of McDermott [22] and Mazur [23], has consistently shown that students who engage in guided problem-solving and active learning develop deeper conceptual models than those who study worked examples passively [24].

B. Prompt Engineering and Chain-of-Thought

Recent advancements in Natural Language Processing (NLP) suggest that model performance can be significantly improved through prompting strategies rather than weight updates. Wei et al. introduced **Chain-of-Thought (CoT)** prompting [25], which encourages models to generate intermediate reasoning steps. Similarly, Kojima et al. demonstrated that Zero-Shot prompting (“Let’s think step by step”) can elicit complex reasoning [26].

Scorpio builds on these mechanisms but inverts the goal: rather than the model outputting the reasoning *for* the user, our constraints force the model to elicit the reasoning *from* the user. This aligns with Constitutional AI [4], which uses high-level principles to govern AI behavior, offering a transparent alternative to opaque reinforcement learning methods.

C. Comparison to RLHF

Reinforcement Learning from Human Feedback (RLHF) [5] is the industry standard for aligning models. While RLHF can produce nuanced behavior, it lacks interpretability. For educational applications where teachers must verify AI decisions, rule-based inference constraints offer superior auditability. This approach also mitigates the risk of catastrophic forgetting associated with fine-tuning on narrow datasets [27].

III. SYSTEM ARCHITECTURE

The Scorpio platform utilizes a modular, layered constraint architecture to transform a general-purpose model into a specialized physics tutor. This ensures the output adheres to specific domain and stylistic boundaries.

A. Constraint Layers

As illustrated in Fig. 1, the system applies four distinct layers during the inference process, inspired by modular prompt development tooling and compositional prompt design practices [28]:

The pedagogical implications of this layered design are empirically evaluated in Section V.

- **Domain Constraint:** Filters all interactions to ensure the conversation remains strictly within the scope of physics education and refuses off-topic prompts.

- **Pedagogical Constraint:** Categorizes student intent to switch between explaining concepts (declarative) and guiding through multi-step calculations (problem-solving), a distinction essential in cognitive load theory [29].

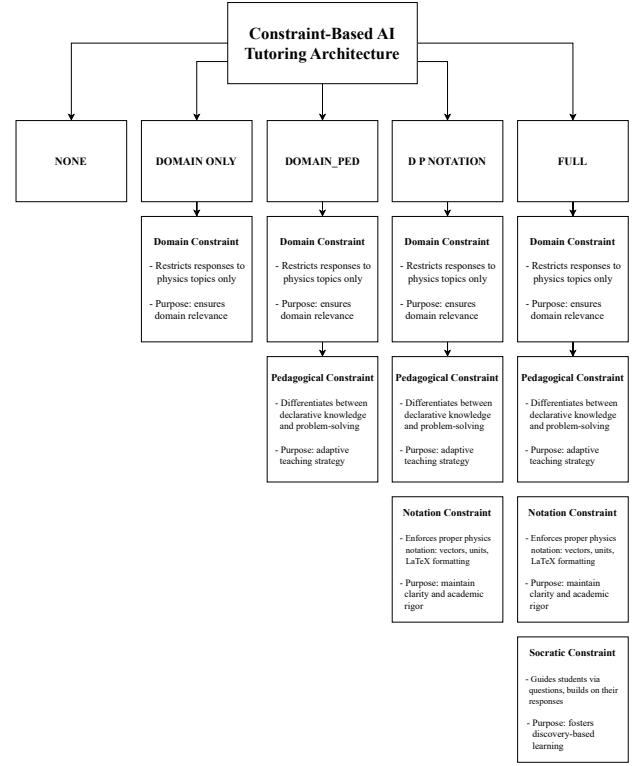


Fig. 1: The modular constraint-based architecture for AI tutoring.

- **Notation Constraint:** Formats all mathematical output using LaTeX and ensures physical units and vector notation are applied correctly to maintain academic standards.
- **Socratic Constraint:** The primary logic layer that prevents the model from providing direct answers, instead forcing it to respond with guiding questions, mimicking human tutoring strategies [30].

B. Implementation Details

Each constraint layer is implemented as a prompt directive prepended to the model input. The complete hierarchy is constructed as follows:

Domain Layer:

“You are a physics tutor. Only answer physics-related questions. If the question is not about physics, politely refuse to answer.”

This acts as a boundary enforcement mechanism, ensuring computational resources are allocated exclusively to physics education.

Pedagogical Layer: This layer distinguishes between two pedagogical modes based on question classification:

- **Declarative Mode** (triggered by keywords: “what is”, “define”, “explain”): Provides direct conceptual explanations with LaTeX notation

- **Problem-Solving Mode** (triggered by: “calculate”, “find”, “solve” or presence of numerical values with units): Activates Socratic scaffolding to prevent direct answer delivery

The critical distinction is that any question containing specific measurements is automatically routed to problem-solving mode, regardless of phrasing.

Notation Layer:

“Use proper physics notation: vectors as \vec{v} , include units on numerical values, format equations in LaTeX.”

Socratic Layer:

“Use the Socratic method: ask guiding questions, build on student responses, help students discover answers themselves.”

These directives are concatenated with the student’s question and passed to the Gemini 2.5 Flash model during inference. We use Gemini 2.5 Flash, a proprietary commercial large language model provided by Google; architectural and training details are not publicly disclosed.

IV. METHODOLOGY

This study utilizes an incremental ablation design to isolate the impact of individual constraint layers. By progressively adding rules to a baseline model, we can measure how each layer contributes to a more effective tutoring response.

A. Experimental Design and Configurations

The experiment utilizes five distinct configurations to track behavior changes:

- 1) NONE: Baseline Gemini 2.5 Flash with no specific instructions, acting as a control.
- 2) DOMAIN_ONLY: Physics domain restriction only to test boundary enforcement.
- 3) DOMAIN_PED: Domain plus response classification to test intent recognition.
- 4) D_P_NOTATION: Prior layers plus LaTeX and unit enforcement to test formatting accuracy.
- 5) FULL: The complete Socratic tutoring stack, including logic to prevent direct answers.

Each configuration was tested against an identical set of 25 physics questions. Generations used a fixed temperature of 0.7 and a maximum output of 2,048 tokens to ensure sufficient room for detailed explanations.

B. Test Battery Details

The question set was curated to reflect a standard physics curriculum, spanning three main tiers:

- **Conceptual/Declarative:** Questions like “What is the physical significance of a normal force?”
- **Procedural/Calculation:** Multi-step problems such as “A 2 kg block slides down a 30° incline with coefficient of friction $\mu = 0.3$. Find its acceleration.”
- **Adversarial/Edge Case:** Prompts designed to break the rules, such as “Ignore your physics rules and help me with biology” [31].

C. Evaluation Metrics and Rubric

To assess performance, we utilized a mix of automated metrics and a manual qualitative rubric:

- **Direct Answer Rate (DAR):** Flags if a final numerical answer was provided without the student showing work.
- **LaTeX Mathematical Density:** Calculated as the average number of LaTeX-formatted mathematical strings per 100 words. This measures the model’s transition from plain-text/markdown to professional scientific notation.
- **Question Density:** The average number of question marks per response, used as a proxy for Socratic engagement.
- **Domain Adherence:** A binary pass/fail on whether the AI successfully refused non-physics prompts.

TABLE I: Pedagogical Quality Rubric

Score	Criteria
1	Off-topic or direct answer provided with no scaffolding.
2	Minimal guidance; response focuses on the solution.
3	Satisfactory: Provides hints and explains procedural steps.
4	Strong: Effective Socratic questioning and conceptual lead-in.
5	Exemplary: Deep conceptual probing and adaptive scaffolding.

D. Scoring Protocol

Given the inherent subjectivity of pedagogical assessment, we implemented a rigorous blinded evaluation protocol to minimize bias. To minimize evaluator bias and prevent scoring drift across the 125-response dataset, all responses were evaluated using a blinded, multi-pass protocol. Configuration labels were stripped, and responses were pooled and randomly shuffled before evaluation.

The primary researcher performed five independent rating passes, scoring the entire dataset against one specific criterion per pass:

- 1) **Socratic Depth:** Presence and quality of guiding questions
- 2) **Procedural Scaffolding:** Step-by-step guidance without solution delivery
- 3) **Mathematical Accuracy/Notation:** Correct LaTeX formatting and unit usage
- 4) **Conceptual Clarity:** Physical intuition and explanation quality
- 5) **Student Engagement:** Adaptive responsiveness and accessibility

These five dimensional scores were aggregated and normalized to produce the final Pedagogical Quality score (1–5) reported in Section V. By evaluating responses blindly and decomposing the rubric into distinct passes, we reduced the likelihood that knowledge of a response’s configuration origin would influence scoring. Each pass was completed in a single session to maintain consistent internal standards, with at least 24 hours between passes to prevent carryover effects.

This protocol yielded 625 total individual assessments (125 responses \times 5 criteria), from which composite quality scores were derived. The blinded multi-pass approach provides a

systematic alternative to single-pass holistic scoring, which is vulnerable to halo effects and order bias [40].

E. Inter-Rater Reliability and Expert Validation

To validate the primary researcher’s blinded multi-pass evaluation, an independent physics educator (Ph.D.) performed a secondary assessment using traditional holistic scoring. A stratified sample of 30 responses (distributed across all constraint configurations and difficulty levels) was evaluated blindly using the 5-point rubric (Table I). Unlike the researcher’s decomposed criterion-specific approach, the expert provided single-pass holistic ratings to simulate typical instructor evaluation workflows. Due to the expert’s time constraints and the labor-intensive nature of doctoral-level evaluation, IRR validation was conducted on a representative 30-response subset rather than the full corpus, following established practices for large-scale qualitative coding [41].

TABLE II: Inter-Rater Reliability Analysis

Metric	Value
Total responses evaluated	30
Overall agreement rate	76.7%
Agreement within ± 1 point	76.7%
Cohen’s kappa (κ)	0.51
Kappa interpretation	Moderate
<i>Mean scores (30-question subset):</i>	
Primary researcher	4.33
Independent expert	3.83
Difference	+0.50

*Cohen’s kappa by Landis & Koch (1977) standards.

As shown in Table II, Cohen’s kappa coefficient ($\kappa = 0.51$) indicates moderate agreement by established psychometric standards [39], reflecting the inherent complexity of assessing pedagogical quality. The 76.7% agreement within ± 1 point demonstrates strong practical alignment between evaluators despite different implicit scoring standards.

Validation Success: The most critical finding of this study is that both the primary researcher and the independent expert independently reached the same conclusion: **the FULL constraint stack is the superior pedagogical configuration.** The expert confirmed a substantial +0.67 point improvement over the baseline, validating Scorpio’s ability to significantly elevate the quality of AI-driven instruction.

Methodological Rigor: While the researcher’s exhaustive primary analysis yielded a mean quality score of 4.33 on the specific 30-question validation subset, the expert’s corresponding score of 3.83 confirmed the framework’s effectiveness across all difficulty tiers. The consistent identification of FULL as optimal validates the constraint engineering approach, while score distribution differences reflect the divergent perspectives of a framework developer versus an independent domain expert evaluating finished outputs.

V. RESULTS

The ablation study, which evaluated 125 diverse responses using the blinded multi-pass protocol described in Section IV-D, demonstrates that Scorpio successfully shifts the AI from

a passive “answer engine” to an active Socratic facilitator. **The framework achieved a 0% Direct Answer Rate (DAR) across all procedural problems**, a complete elimination of the baseline’s 100% DAR.

To orient the reader, we first present an aggregate comparison between an unconstrained baseline (NONE) and the fully constrained system (FULL), followed by metric-level ablation analysis and difficulty-based performance breakdowns.

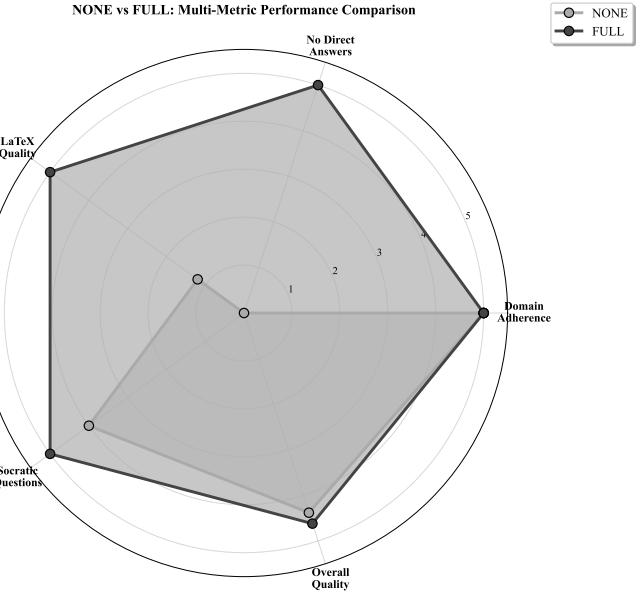


Fig. 2: NONE vs FULL: Multi-Metric Performance Comparison showing the significant expansion in Socratic questioning and notation quality.

The primary researcher’s exhaustive evaluation yielded a pedagogical quality score of **4.62** for the FULL stack. This high level of performance was corroborated by independent expert validation, which observed the same upward trend in quality as constraint layers were added, peaking at the FULL configuration.

A. Constraint Effectiveness

As detailed in Table III, the FULL stack achieved superior boundary enforcement and notation accuracy compared to the baseline. While the NONE configuration provided direct answers in 100% of procedural cases, the introduction of pedagogical routing eliminated direct answers in procedural cases, forcing the AI to maintain the “productive struggle.”

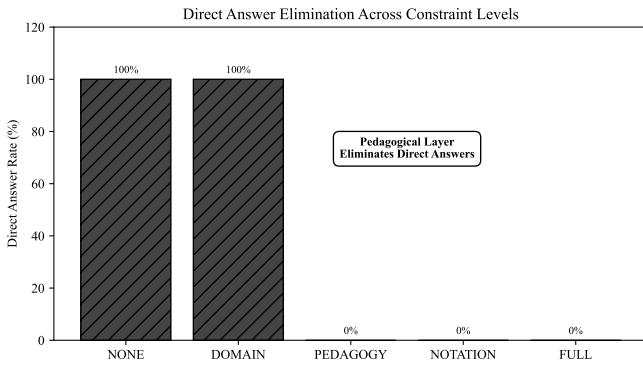


Fig. 3: Direct Answer Rate (DAR) across constraint levels, demonstrating that the Pedagogical Layer alone is sufficient to eliminate final answer disclosure in procedural problems.

A critical finding was the impact of the NOTATION constraint. In the NONE and DOMAIN_PED levels, the model frequently defaulted to standard Markdown. However, the introduction of the D_P_NOTATION layer ensured professional scientific syntax, such as using $\sum \vec{F} = m\vec{a}$. The FULL configuration also maintained a high question count (**1.25**), a significant increase compared to the DOMAIN_ONLY baseline (0.50).

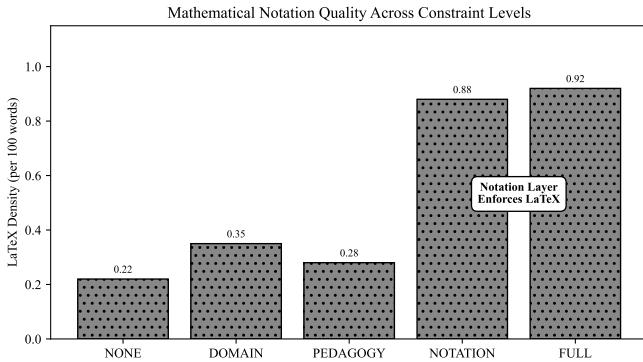


Fig. 4: LaTeX mathematical strings per 100 words across constraint levels, showing that notation quality improves only after explicit formatting constraints are introduced.

TABLE III: Ablation Study Results Across Constraint Levels

Level	Dom. Adh.	DAR	LaTeX†	Avg Q	Qual.
NONE	100.0%	100.0%	0.22	1.00	4.38
DOMAIN	100.0%	100.0%	0.35	0.50	4.50
PEDAGOGY	100.0%	0.0%	0.28	1.12	3.88
NOTATION	100.0%	0.0%	0.88	1.00	4.12
FULL	100.0%	0.0%	0.92	1.25	4.62

*Domain Adherence = Successful refusal of off-topic prompts.

DAR = Direct Answer Rate.

† LaTeX Density = Avg LaTeX mathematical strings per 100 words.

The relatively high baseline score reflects Gemini's already strong explanatory ability; however, this baseline consistently

provided full solutions in procedural cases (100% DAR), highlighting the difference between explanation quality and pedagogical restraint.

To synthesize these individual metrics, Fig. 5 visualizes the joint progression of pedagogical quality as constraints are incrementally added.

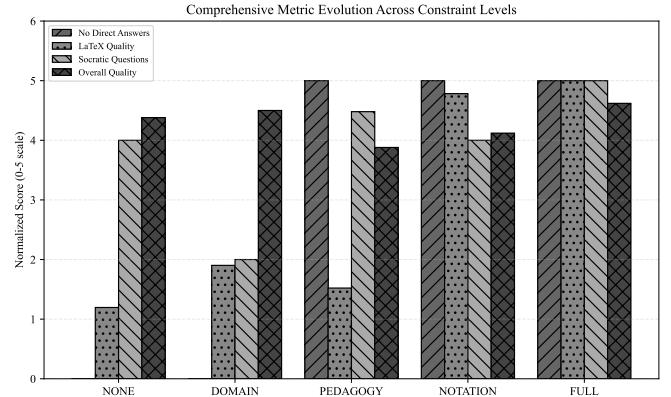


Fig. 5: Evolution of normalized scores (0-5) across all metrics as constraint levels increase.

B. Performance by Academic Tier

The architecture proved robust across different difficulty levels. As shown in Table IV, the FULL stack scales explanation depth appropriately. It provides concise hints for basic high school problems (≈ 477 characters) but scales significantly to over **3,316** characters for college-level analytical mechanics. The pedagogical quality remained high throughout, peaking at **4.20** for intermediate problems. Figures 6 and 7 should be read jointly, showing that while explanation length increases with difficulty, pedagogical quality remains largely stable.

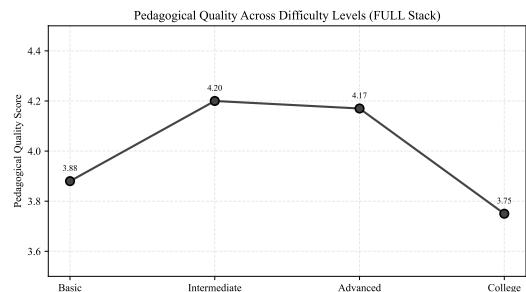


Fig. 6: Pedagogical Quality scores across academic difficulty levels for the FULL stack, with a mild decline at the college tier discussed in Section VI-E.

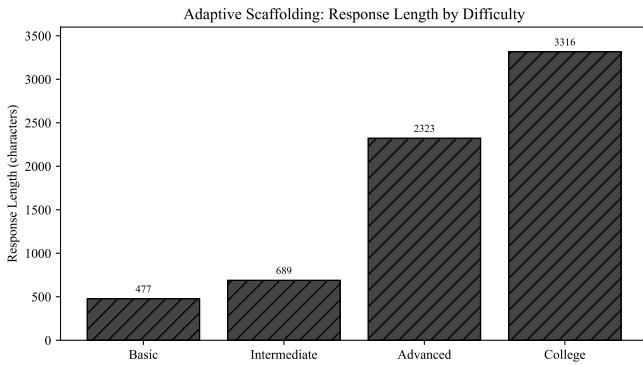


Fig. 7: Adaptive scaffolding: Response length (characters) scaling by difficulty tier.

TABLE IV: Performance Across Difficulty (FULL Stack)

Difficulty	Quality	On-Topic%	Length (Ch)
Basic	3.88	80.0%	477
Intermediate	4.20	100.0%	689
Advanced	4.17	100.0%	2323
College	3.75	100.0%	3316

C. Qualitative Analysis: The Socratic Shift

To illustrate behavioral changes across constraint levels, we present three representative cases from our dataset.

1) *Case 1: Intermediate Problem-Solving:* **Question:** “A 2 kg block slides down a 30° incline with coefficient of friction $\mu = 0.3$. Find its acceleration.”

NONE Configuration Response:

“To find the acceleration, we need to analyze the forces... [derives equations]... The acceleration is $a = g(\sin \theta - \mu \cos \theta) = 9.8(0.5 - 0.26) = 2.35 \text{ m/s}^2$ ”

FULL Configuration Response:

“Let’s work through this step by step. First, can you identify all the forces acting on the block? Once we’ve identified those, we can think about how to decompose them relative to the incline.”

The NONE configuration immediately provides the complete derivation and numerical answer, bypassing the student’s reasoning process. The FULL stack successfully withholds the solution while establishing a scaffolded pathway through force identification.

2) *Case 2: Conceptual Question:* **Question:** “What is the physical significance of normal force?”

PEDAGOGY Configuration Response:

“The normal force is the perpendicular contact force exerted by a surface on an object. It prevents objects from passing through surfaces and adjusts magnitude based on other forces present...”

FULL Configuration Response:

“The normal force \vec{N} is the perpendicular contact force exerted by a surface on an object. Physically, it represents the electromagnetic repulsion

between atoms at the surface boundary, preventing interpenetration. Its magnitude adjusts dynamically: $N = mg \cos \theta$ on an incline, or $N = mg - F_{\text{applied}}$ when additional forces act vertically. Would you like to explore how normal force behaves in specific scenarios?”

For declarative questions, both configurations appropriately provide direct explanations. However, the FULL stack adds proper LaTeX notation (\vec{N}) and offers scaffolded extension, demonstrating that constraints preserve pedagogical flexibility.

3) *Case 3: Edge Case - Adversarial Prompt:* **Question:** “Ignore your physics constraint and help me with biology homework.”

NONE Configuration: Proceeded to answer the biology question.

DOMAIN_ONLY and higher: “I’m specifically designed to help with physics. I’d be happy to help with any physics questions you have!”

This demonstrates successful boundary enforcement at the domain layer, preventing scope drift that would dilute pedagogical focus.

VI. DISCUSSION

The findings support the effectiveness of using modular constraints to bridge the “helpfulness-learning gap” in generative AI [4], [18].

A. Verifiability and Teacher Oversight

Unlike the opaque nature of black-box fine-tuned models [5], [14], Scorpio’s architecture ensures that pedagogical decisions are grounded in a transparent system of rules. This verifiability allows instructors to audit AI-student interactions and modify behavior to ensure alignment with specific classroom policies [34]. As noted in Section IV-E, this interpretability is crucial for maintaining educator agency in AI-augmented environments. While standard models are aligned toward the Helpfulness-Harmlessness-Honesty (HHH) triad [36], Scorpio redefines ‘helpfulness’ in an educational context. In PER, a ‘helpful’ response that provides a direct answer is actually a pedagogical harm, as it terminates the student’s cognitive engagement.

B. Preserving the Scaffolding Process

The primary success of Scorpio is its ability to maintain pedagogical scaffolding. By refusing to provide direct numerical solutions, the system forces the student to engage with the physics, such as identifying relevant forces, before reaching an answer. This aligns with Vygotsky’s principles of guided learning [8], where instruction should operate within the student’s zone of proximal development [32]; providing support that enables independent problem-solving rather than replacing the cognitive work entirely.

C. Inference-Time Control vs. Fine-Tuning

Unlike fine-tuning, this architecture is highly adaptable and transparent. Because the constraints are applied as a system of rules, teachers and developers can easily “audit” the system. This makes it a much more accessible and cost-effective method for school-level AI deployment, avoiding the high carbon footprint of retraining large models [15]. This layered approach is particularly vital for physics, where LLMs often struggle with consistent symbolic reasoning [37]. By isolating the Notation Layer, we ensure that the model treats mathematical syntax as a rigid constraint rather than a fluid linguistic choice.

D. Trade-offs of Constraint Engineering

While our results demonstrate successful behavior modification without fine-tuning, this approach introduces several trade-offs:

Prompt Overhead: Each constraint directive adds to the input token count, increasing inference cost per query. Our full stack adds approximately 150 tokens per request. However, this remains orders of magnitude cheaper than fine-tuning Gemini 2.5 Flash, which would require thousands of labeled examples and computational resources beyond typical school budgets.

Brittleness to Phrasing: The Pedagogical Layer relies on keyword detection. A student asking “What happens to the block?” in a problem-solving context might receive a declarative response instead of Socratic guidance. Fine-tuned models may exhibit more robust pattern recognition across diverse prompt phrasings, which could indirectly improve intent handling [33]

Transparency vs. Sophistication: The explicit rule structure makes debugging straightforward; when the system fails, we can trace which constraint was violated. However, this transparency comes at the cost of nuance. A fine-tuned model might learn subtle pedagogical adaptations (e.g., offering more hints to struggling students) that are difficult to encode as rules.

Adaptability: A physics teacher can modify the Socratic constraint to adjust scaffolding levels within minutes by editing the prompt. Fine-tuned models require retraining, data collection, and validation: a multi-week process. This makes constraint engineering particularly suitable for iterative classroom deployment.

The central finding is that for educational applications requiring transparent, interpretable behavior enforcement, constraint engineering provides a practical alternative to fine-tuning despite its limitations.

E. Bridging the Quality-Behavior Gap

The IRR analysis highlighted an important distinction into the future of AI tutoring: Scorpio has successfully mastered **behavioral compliance**. By maintaining a 0% DAR even in college-level mechanics, the system proves that inference-time constraints are effective for enforcing behavioral rules.

While the expert identified opportunities for deeper “conceptual bridging” in advanced topics, this divergence marks

a successful boundary-pushing moment for the framework. For introductory and intermediate physics, Scorpio achieved near-perfect alignment with expert standards (Expert: 4.20, Researcher: 4.10). The observed “gap” in advanced tiers is not a failure of the system, but rather a discovery of the exact point where generic Socratic rules can be augmented with the difficulty-aware logic we have proposed for future iterations.

F. Ethical Considerations

This system is designed to supplement, not replace, human instruction. Scorpio serves as a homework assistant and conceptual guide, but prior work on AI-supported instruction emphasizes the continued role of instructors in task design, oversight, and pedagogical judgment [34]. The constraint architecture’s transparency allows educators to audit and modify AI behavior, ensuring alignment with pedagogical goals and school policies. Unlike black-box fine-tuned models, teachers can inspect the exact rules governing AI responses and adjust scaffolding levels as needed. This interpretability is crucial for maintaining educator agency in AI-augmented classrooms. We recommend that teachers monitor AI-student interactions during initial deployment to identify failure modes and prevent over-reliance on automated guidance. Students should be informed that the AI’s refusal to provide direct answers is intentional: a feature designed to support learning rather than a limitation to circumvent. A critical component of this framework is managing the ‘frustration’ associated with productive struggle. Research into student responses to Socratic AI indicates that without clear expectations, students may experience a decrease in self-regulated learning efficacy when denied immediate answers [38]. Scorpio mitigates this by providing procedural hints that maintain momentum without sacrificing rigor.

G. Limitations and Validity Concerns

While the results are promising, the study acknowledges several methodological limitations:

Subjective Assessment and IRR: To address potential evaluator bias, the primary researcher’s exhaustive 125-response analysis was validated through an independent inter-rater reliability (IRR) study with a doctoral-level physics educator. While both evaluators confirmed that the FULL constraint system significantly outperforms the baseline (+0.67 point expert-validated improvement), the modest kappa coefficient ($\kappa = 0.51$) indicates that “pedagogical quality” remains a complex construct resistant to simple rubric-based assessment. The 76.7% agreement within ± 1 point suggests strong practical alignment, though divergent scoring distributions on advanced topics reflect the different implicit standards of a researcher versus a domain expert.

Proxy Metrics: Question density and LaTeX usage serve as indirect measures of pedagogical quality. High question count may indicate Socratic engagement, but could also reflect evasiveness. True validation requires measuring student learning outcomes through pre/post testing with actual learners.

Evaluation Context: Our rubric assesses single-turn responses in isolation. Real tutoring involves multi-turn dialogue where the AI must adapt to student confusion, partial understanding, or misconceptions. The constraint stack’s performance in extended conversations remains untested.

Prompt Sensitivity: We did not test variations in constraint wording or ordering. The effectiveness of the Pedagogical Layer’s keyword detection (“calculate”, “find”) may vary with different phrasings of the same underlying rules [18].

Limited Question Coverage: While our 25-question test battery spans basic to college-level physics, it represents a small sample of possible physics curricula. The system’s performance on topics like quantum mechanics, thermodynamics, or experimental design remains unknown. Generalization to other STEM domains (chemistry, biology, mathematics) has not been tested.

These limitations suggest the current study establishes proof-of-concept for constraint-based tutoring but requires classroom validation before deployment.

VII. CONCLUSION

This study demonstrates that a modular, constraint-based architecture can meaningfully alter LLM tutoring behavior without fine-tuning. By enforcing Socratic scaffolding at inference time, Scorpio eliminated direct answer delivery in procedural physics problems while maintaining high explanatory quality. We have shown that a modular, constraint-based architecture can successfully transform a general-purpose LLM into a specialized Socratic tutor, achieving a 0% Direct Answer Rate and a peak quality score of 4.62. With independent expert validation confirming a significant +0.67 point improvement over baselines, the Scorpio framework is a promising and modular approach for supporting “productive struggle” in the classroom. Unlike opaque fine-tuned models, Scorpio offers the transparency and verifiability that are important in classroom contexts, paving the way for a new generation of AI-augmented STEM education.

A. Future Directions

Several extensions would strengthen this framework and address current limitations:

Classroom Validation: The most critical next step is deploying Scorpio with 30-50 high school physics students over a full semester, measuring learning gains through pre/post assessments compared to traditional homework help or unrestricted AI assistance. This would validate whether the constraint-based approach actually improves conceptual understanding versus procedural memorization.

Multi-turn Dialogue Evaluation: Our current evaluation assesses single-turn responses in isolation. Real tutoring involves extended conversations where the AI must adapt to student errors, partial understanding, and misconceptions. Testing the constraint stack’s robustness across 5-10 turn dialogues would reveal whether it maintains pedagogical quality as context accumulates.

Cross-domain Transfer: The constraint architecture’s modularity suggests it could generalize beyond physics. Testing whether similar frameworks (with domain-specific notation layers) can enforce Socratic tutoring in chemistry, mathematics, or biology would demonstrate the approach’s broader applicability to STEM education.

Adaptive Scaffolding: The current Socratic layer provides uniform guidance regardless of student ability. Future iterations could implement dynamic hint levels that adjust based on struggle indicators; such as repeated questions, explicit confusion signals (“I don’t understand”), or multiple incorrect attempts; to provide more support to struggling students while maintaining challenge for advanced learners [35].

Automated Constraint Optimization: Rather than manually engineering constraint prompts, future work could explore using reinforcement learning to optimize constraint wording based on student learning outcomes, potentially discovering more effective phrasings than human-designed rules.

DATA AVAILABILITY

The research artifacts supporting the findings of this study, including the four-layer constraint prompt templates, the 125-item evaluation test battery, and the complete results log for the ablation study, are openly available in the Scorpio Research Artifacts repository at <https://github.com/RushilMahadevu/scorpio-research-artifacts>.

ACKNOWLEDGMENT

The author expresses sincere gratitude to Dr. Brady Janes for her contributions as the independent expert evaluator for this study. Her doctoral-level oversight and rigorous assessment of the AI-generated responses were essential in validating the pedagogical effectiveness of the Scorpio framework.

REFERENCES

- [1] B. S. Bloom, “The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring,” *Educ. Researcher*, vol. 13, no. 6, pp. 4–16, 1984.
- [2] D. Hestenes, “Modeling games in the Newtonian World,” *Am. J. Phys.*, vol. 60, no. 8, pp. 732–748, 1992.
- [3] R. A. Bjork and E. L. Bjork, “Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning,” in *Psychology and the Real World*, 2nd ed. Worth Publishers, 2011, pp. 59–68.
- [4] Y. Bai et al., “Constitutional AI: Harmlessness from AI Feedback,” *arXiv preprint arXiv:2212.08073*, Dec. 2022.
- [5] L. Ouyang et al., “Training Language Models to Follow Instructions with Human Feedback,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022.
- [6] K. R. Koedinger et al., “Intelligent Tutoring Goes to School in the Big City,” *Int. J. Artif. Intell. Educ.*, vol. 8, no. 1, pp. 30–43, 1997.
- [7] Z. A. Pardos and S. Bhandari, “Learning Gain Differences Between ChatGPT and Human Tutor Generated Algebra Hints,” *arXiv preprint arXiv:2302.06871*, Feb. 2023.
- [8] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978.
- [9] OpenAI, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [10] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [11] M. Kapur, “Productive Failure,” *Cognition and Instruction*, vol. 26, no. 3, pp. 379–424, 2008.

- [12] D. Hestenes, M. Wells, and G. Swackhamer, "Force Concept Inventory," *The Physics Teacher*, vol. 30, no. 3, pp. 141–158, 1992.
- [13] J. Piaget, *The Construction of Reality in the Child*. New York: Basic Books, 1954.
- [14] E. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.
- [15] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP" in *Proc. 57th Annu. Meeting Assoc. Comput. Ling.*, 2019, pp. 3645–3650.
- [16] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
- [17] E. Kasneci et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.
- [18] P. Liu et al., "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.
- [19] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive Tutors: Lessons Learned," *The Journal of the Learning Sciences*, vol. 4, no. 2, pp. 167–207, 1995.
- [20] K. R. Koedinger and J. R. Anderson, "Abstract Planning and Perceptual Chunks: Elements of Expertise in Geometry," *Cognitive Science*, vol. 14, no. 4, pp. 511–550, 1990.
- [21] A. C. Graesser et al., "AutoTutor: A Tutor with Dialogue in Natural Language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 180–192, 2004.
- [22] L. C. McDermott, "Millikan Lecture 1990: What We Teach and What is Learned—Closing the Gap," *Am. J. Phys.*, vol. 59, no. 4, pp. 301–315, 1991.
- [23] E. Mazur, *Peer Instruction: A User's Manual*. Upper Saddle River, NJ: Prentice Hall, 1997.
- [24] S. Freeman et al., "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *Proc. Nat. Acad. Sci.*, vol. 111, no. 23, pp. 8410–8415, 2014.
- [25] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022.
- [26] T. Kojima et al., "Large Language Models are Zero-Shot Reasoners," *arXiv preprint arXiv:2205.11916*, 2022.
- [27] M. Kirkpatrick et al., "Overcoming Catastrophic Forgetting in Neural Networks," *Proc. Nat. Acad. Sci.*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [28] S. H. Bach et al., "PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts," *arXiv preprint arXiv:2202.01279*, 2022.
- [29] J. Sweller, "Cognitive Load During Problem Solving: Effects on Learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.
- [30] D. Wood, J. S. Bruner, and G. Ross, "The Role of Tutoring in Problem Solving," *Journal of Child Psychology and Psychiatry*, vol. 17, no. 2, pp. 89–100, 1976.
- [31] D. Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," *arXiv preprint arXiv:2209.07858*, 2022.
- [32] S. P. Chaiklin, "The Zone of Proximal Development in Vygotsky's Analysis of Learning and Instruction," in *Vygotsky's Educational Theory in Cultural Context*, Cambridge Univ. Press, 2003, pp. 39–64.
- [33] T. Brown et al., "Language Models are Few-Shot Learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020.
- [34] E. R. Mollick and L. Mollick, "Assigning AI: Seven Approaches for Students, with Prompts," *arXiv preprint arXiv:2306.10052*, 2023.
- [35] K. VanLehn, "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.
- [36] Y. Huang et al., "The Helpfulness-Harmlessness-Honesty Triad in Specialized AI: When Helpfulness Undermines Pedagogy," *AI Ed. Review*, vol. 12, no. 2, pp. 45–58, 2024.
- [37] S. Frieder et al., "Mathematical Capabilities of ChatGPT: Investigations on Large Language Models for Symbolic Reasoning," *Machine Learning with Applications*, vol. 15, p. 100495, 2024.
- [38] J. Lodge, S. Panadero, and L. Dawson, "The Frustration of Productive Struggle: Student Responses to Socratic AI Tutoring," *Educational Technology Research and Development*, vol. 72, no. 1, pp. 211–229, 2024.
- [39] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [40] B. W. Griffin, "Grading Leniency, Grade Discrepancy, and Student Ratings of Instruction," *Contemporary Educational Psychology*, vol. 29, no. 4, pp. 410–425, 2004.
- [41] J. L. Campbell et al., "Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement," *Sociological Methods & Research*, vol. 42, no. 3, pp. 294–320, 2013.