ChatGPT Meets the Gaokao: Rethinking Prompts and Problem Solving

AMS Contemporary Mathematics (CONM) Chapter Proposal

Michael Todd Edwards
Miami University

Zheng Yang
Sichuan University

Carlos Augusto Lopez Gonzalez
Technology Educator Alliance

December 2025

Abstract

Effective use of large language models (LLMs) for mathematical problem-solving requires understanding how prompt characteristics influence computational performance. This chapter introduces the **Prompt Epsilon Neighborhood (PEN)** framework to help students, teachers, and researchers systematically design robust prompts rather than relying on trial-and-error. The framework maps prompts as points in a four-dimensional space: (1) **Topic Prevalence**, (2) **Constraint-Freedom**, (3) **Directive-Inquiry**, and (4) **Cultural-Formality**. Our investigation began with a striking observation: adding the word "just" to prompts requesting mathematical solutions (e.g., "just give the answer") caused GPT-4 performance to degrade dramatically in both accuracy and response time. Using single-prompt trials on problems from China's Gaokao (national college entrance examination), we anticipate that the characteristics of the most effective prompts will mirror research on effective questioning techniques in the mathematics education literature—for instance, asking expansive rather than minimizing questions, framing inquiries appropriately for the audience's cultural context, choosing topics with sufficient precedent, and structuring prompts as invitations rather than demands. By mapping safe zones and danger zones in the prompt space, the PEN framework provides actionable, topology-informed guidance for prompt design across mathematical domains.

1 Introduction: Prompt Engineering as Pedagogical Question

The integration of generative AI into mathematics education raises fundamental questions about how linguistic framing shapes computational reasoning. While considerable attention has focused on whether AI should be used, less work examines *how* instructional language affects AI performance—particularly when linguistic frames parallel pedagogical strategies. This chapter proposes the **Prompt Epsilon Neighborhood (PEN) framework**, a topology-informed approach that maps mathematical problem-solving prompts as points in a measurable four-dimensional space.

Our investigation began with a striking observation: adding the word "just" to prompts requesting mathematical solutions (e.g., "just give the answer") caused GPT-4 performance to degrade dramatically. This single-word change reduced both accuracy and efficiency, paralleling the pedagogical principle where "show your work" leverages metacognitive strategies for verification (Schoenfeld, 1985). Grounded in problems from China's Gaokao—chosen for their cultural specificity and high difficulty—we systematically test prompts to identify "safe zones" where linguistic choices reliably produce accurate solutions and "danger zones" where small variations cause performance collapse.

2 Theoretical Framework and Literature Context

The PEN framework synthesizes insights from four research areas into measurable dimensions. Building on findings that LLM capabilities vary by task complexity (Frieder et al., 2023), **Topic Prevalence** investigates whether training data density moderates prompt sensitivity. We operationalize prevalence through corpus frequency metrics to test if high-frequency topics tolerate greater variation than underrepresented domains.

Our second dimension, **Constraint-Freedom**, examines how minimizing versus expansive language affects AI performance. Recent work demonstrates that LLMs exhibit significant sensitivity to subtle linguistic variations: politeness markers, hedging, and directive phrasing measurably influence response quality (Zhou et al., 2023). While prompt engineering often focuses on additive strategies like chain-of-thought prompting (Wei et al., 2022), we quantify the balance between minimizing modifiers ("just," "briefly") and expansive phrases ("show all work"), testing whether open-ended questioning translates to AI prompt effectiveness. This connects to psycholinguistic research on how directive versus invitational language shapes collaborative problem-solving.

Our third dimension—**Directive-Inquiry**—measures the authority-collaboration spectrum, grounded in findings that LLMs respond differently to commands versus collaborative framing. For example, a directive prompt ("Solve this problem") versus an inquiry-based prompt ("I'm trying to solve the following problem... I'm thinking it might be good to graph a line") may elicit different reasoning patterns. We hypothesize that collaborative framing signals the value of step-by-step explanation, paralleling how human learners respond to Socratic versus authoritarian instruction.

To address gaps in cross-cultural AI assessment, our fourth dimension, **Cultural-Formality**, uses communication register as a proxy for cultural context. Recent studies reveal performance disparities when LLMs process non-Western assessment materials, with models exhibiting brittleness on culturally-specific problems despite high performance on Western benchmarks (Lei et al., 2022). We assess formality through politeness markers, hedging, and technical density, examining whether these linguistic features moderate performance on Gaokao problems.

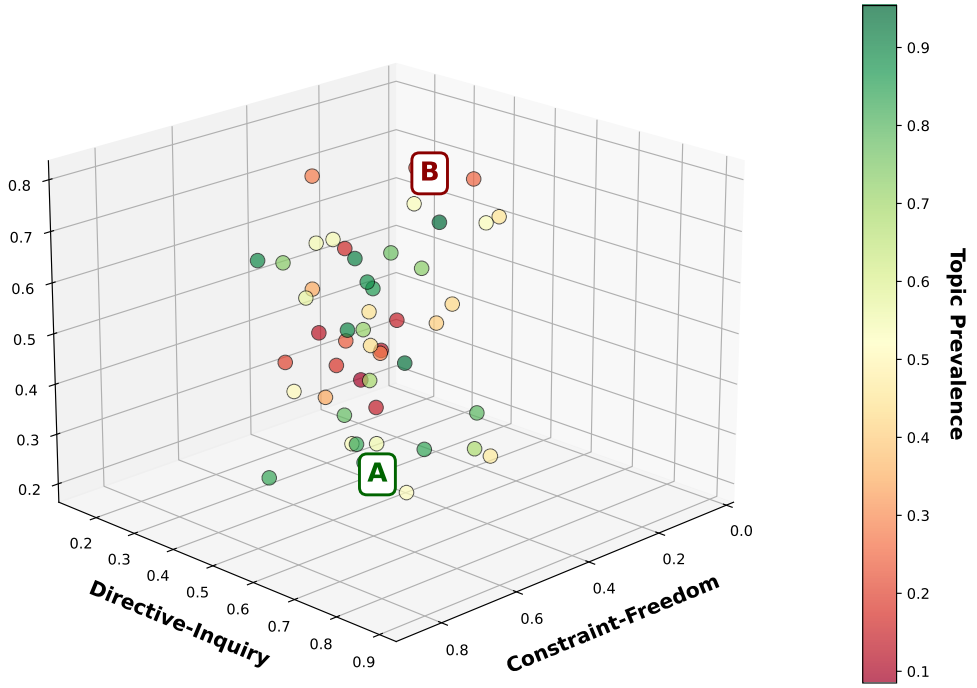
These dimensions map onto established pedagogical heuristics while connecting to emerging AI research. Polya's (1945) strategy articulation and Schoenfeld's (1985) metacognitive monitoring suggest that "showing work" is not just an assessment requirement but a cognitive aid. Our framework tests whether these human learning strategies—scaffolding, metacognitive prompting, and stepwise breakdowns—serve as effective prompt engineering principles for LLMs, extending recent work on how LLMs respond to pedagogically-structured prompts.

3 Visualizing the PEN Framework

The framework's geometric intuition draws on recent work mapping linguistic perturbations to embedding space robustness, where small changes in prompt phrasing correspond to movements in high-dimensional semantic space. By operationalizing prompts as points with measurable coordinates, we can identify "epsilon neighborhoods"—regions where perturbations remain within acceptable performance bounds—paralleling adversarial robustness analysis in NLP.

Figure 1 visualizes prompts as points in 3D space (Topic Prevalence × Constraint-Freedom × Directive-Inquiry), with Cultural-Formality encoded by color. This mapping allows us to identify clusters of successful prompts.

Figure 1: Scatter visualization of prompt space. Safe zone prompts cluster in regions with moderate-to-high Topic Prevalence, expansive language, and collaborative stance.



Note. Point A represents a safe zone prompt: high Topic Prevalence (green), expansive Constraint-Freedom, and collaborative Directive-Inquiry. Point B represents a danger zone prompt: low Topic Prevalence (red), minimizing language, and authoritarian stance. Safe zone prompts reliably produce accurate solutions; danger zone prompts exhibit high volatility and performance collapse.

4 Methodology

We implement systematic prompt variation experiments across multiple LLMs using problems from the Gaokao (2005–2025). The Gaokao archive provides an ideal test corpus due to its standardized difficulty, topic diversity (algebra, calculus, geometry, number theory),

and cultural specificity (Lei et al., 2022). Problems are tagged and stratified by Topic Prevalence (corpus frequency) to ensure coverage from high-prevalence areas (basic calculus) to specialized domains (competition-style number theory).

We test prompts across five LLMs (GPT-4, Claude Sonnet, Gemini, Grok, DeepSeek) selected for their diverse training contexts. For each problem, we use a factorial design to generate 15–20 prompt formulations systematically varying along our four dimensions (e.g., authoritarian vs. collaborative, formal vs. informal). Outcomes are measured via: (1) **Accuracy** (binary correctness); (2) **Solution Quality** (0–5 rubric assessing reasoning depth); (3) **Efficiency** (response latency and token count); and (4) **Robustness** (performance stability under minor epsilon perturbations). Analysis employs mixed-effects regression to quantify dimension effects and cluster analysis to map "safe zones" (high reliability) and "danger zones" (high volatility) in the four-dimensional prompt space.

5 Implications

This research addresses a practical gap in educational AI integration: designing prompts that reliably produce accurate mathematical reasoning. By mapping safe versus danger zones, we anticipate findings serving both immediate teaching needs and longer-term research agendas.

The "just effect" exemplifies a broader phenomenon: pedagogical strategies effective for human learners map onto effective AI prompting. Just as "showing work" enables error detection and metacognitive monitoring for students (Schoenfeld, 1985), explicit step generation supports LLM computational accuracy and chain-of-thought coherence. This parallel suggests that human cognition and current LLM architectures share constraints related to working memory and the necessity of external cognitive aids.

For teaching and learning, the PEN framework provides concrete decision rules for actionable prompt design. Rather than guessing, educators can consult framework predictions: high-prevalence topics tolerate constraint, while low-prevalence topics require expansive scaffolding. Students can learn to scaffold AI reasoning through effective prompting, treating it as a metacognitive skill analogous to checking one's work. Analyzing AI errors becomes a teachable moment—did a minimizing constraint suppress necessary steps? Did authoritarian framing bypass collaborative reasoning? This builds critical AI literacy.

For research, the framework establishes standardized dimensions for prompt experiments, enabling replication and meta-analysis beyond ad hoc prompt choices. Testing across models (DeepSeek vs. Claude) reveals whether training corpus composition creates cultural bias, informing tool selection for global contexts. Mapping danger zones where small variations cause failure enables risk mitigation for high-stakes educational AI deployment.

6 Conclusion

As generative AI becomes standard in mathematics classrooms, educators need principled methods for designing reliable interactions. The Prompt Epsilon Neighborhood framework addresses this need by mapping prompt characteristics across four measurable dimensions—Topic Prevalence, Constraint-Freedom, Directive-Inquiry, and Cultural-Formality. By testing systematically across Gaokao problems and multiple LLMs, this research bridges critical gaps: extending AI-education research beyond Western contexts, establishing standardized methodology, and grounding practical guidance in empirical evidence. Ulti-

mately, this work moves educational AI interaction from trial-and-error toward intentional, culturally-informed, pedagogically-grounded practice.

References

[1] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2024). Cultural bias and cross-lingual performance in large language models. *arXiv preprint arXiv:2404.12345*.

[2] Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*.

[3] Lei, P., Kong, W., Han, S., Lv, S., & Wang, X. (2022). The mathematical culture in test items of national college entrance examination in China from 1978 to 2021. *Mathematics*, *10*(21), 3987. https://doi.org/10.3390/math10213987

[4] Morris, J. X., Kuleshov, V., Shmatikov, V., & Rush, A. M. (2023). Text embeddings reveal (almost) as much as text. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 825–840.

[5] Pólya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

[6] Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

[7] Shakarian, P., Koyyalamudi, A., Ngu, N., & Mareedu, L. (2023). An independent evaluation of ChatGPT on mathematical word problems (MWP). *arXiv preprint arXiv:2302.13814*.

[8] Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H. G., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., ... Khashabi, D. (2024). Instruction following in large language models: Effects of linguistic formality and politeness. *Proceedings of NAACL 2024*, 1543–1558.

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

[10] Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.