2026 Volume: 02 Issue: 09 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-09

A COMPARATIVE STUDY OF PROMPT ENGINEERING TECHNIQUES FOR IMPROVING LARGE LANGUAGE MODEL PERFORMANCE IN EDUCATIONAL APPLICATIONS

[*1]Prof. Narendra J. Padole, [2]Dr. Nitin S. Shrirao

[1]Department of Computer Science and Technology, DCPE, HVPM, Amravati, India.

[2]Siddhant Institute of Computer Application, Sudumbare, Pune.

Article Received: 25 July 2026

Article Revised: 15 August 2026

Published on: 04 September 2026

***Corresponding Author: Prof. Narendra J. Padole**

Department of Computer Science and Technology, DCPE, HVPM, Amravati, India.

DOI: https://doi-doi.org/101555/ijrpa.5948

ABSTRACT

Large language models (LLMs) are increasingly used in educational applications for tutoring, question generation, formative feedback, summarization, programming assistance, and learning support. However, model performance is strongly influenced by how tasks are specified through prompts. This paper presents a comparative framework for evaluating prompt engineering techniques in educational scenarios. Five prompting strategies—zero-shot, few-shot, structured role-and-constraint prompting, reasoning-oriented prompting, and retrieval-augmented/context-grounded prompting—are compared using a common evaluation protocol. The framework evaluates response correctness, relevance, pedagogical alignment, groundedness, hallucination tendency, consistency, and prompt efficiency. Rather than claiming universal superiority of one technique, the study analyzes task-dependent trade-offs and proposes a reproducible evaluation matrix for educational LLM systems. Recent literature indicates that simple prompting remains useful for low-complexity tasks, demonstrations and structured prompts can improve task alignment, reasoning-oriented prompting can support complex problem solving, and retrieval-grounded approaches can improve grounding when authoritative course material is available. The paper concludes with recommendations for selecting prompting strategies according to educational task characteristics and identifies directions for automated prompt optimization and human-centered evaluation.

KEYWORDS: Large language models, prompt engineering, generative AI, educational technology, intelligent tutoring systems, prompt evaluation, personalized learning.

I. INTRODUCTION

Generative artificial intelligence has rapidly entered educational workflows, with LLMs being used for intelligent tutoring, feedback generation, content creation, programming support, assessment assistance, and learning analytics. Recent reviews of empirical educational applications report potential benefits together with persistent concerns involving reliability, over-reliance, fairness, privacy, and assessment validity [1], [2]. In this setting, prompt engineering is a controllable layer that can substantially affect model behavior and pedagogical quality.

Prompt engineering refers to the systematic design of instructions, context, examples, constraints, and interaction structures used to guide model outputs. Educational prompt engineering is particularly challenging because a response can be factually correct yet pedagogically poor—for example, by giving away an answer instead of scaffolding the learner. Recent systematic reviews distinguish technique-oriented prompting from process-oriented prompting and show that educational use cases span both learning support and automation [3], [4].

Despite rapid growth, comparative evidence remains fragmented. Different studies use different models, datasets, tasks, evaluators, and metrics, making direct conclusions difficult. A systematic review of K–12 STEM studies found widespread use of zero-shot prompting, with evidence of benefits from few-shot and chain-of-thought approaches, but also noted limited real-world validation [5]. A 2025 meta-analysis of digital learning prompts further found that prompt design features and learner characteristics moderate effectiveness [6].

This paper addresses the research question: Which prompt engineering strategies provide the best balance of correctness, pedagogical alignment, reliability, and efficiency for different educational tasks? The contribution is a unified comparative evaluation framework that can be instantiated across tutoring, assessment, question generation, summarization, and programming education.

II. RELATED WORK

LLMs in education. Recent empirical reviews show that LLMs are being applied across intelligent tutoring systems, assessment, content generation, programming education, language learning, and academic support [1]. These studies also highlight hallucination, over-

reliance, privacy, and inconsistent quality, indicating that evaluation must consider educational outcomes rather than language quality alone.

Prompt engineering in higher education. Lee and Palmer [3] systematically reviewed prompt engineering in higher education and reported multiple frameworks for improving GenAI interactions. Their review emphasizes alignment between prompts and predetermined educational goals. Qian [4] categorized educational prompting into technique-based and process-based approaches and identified critical-skill development and educational-function automation as major application areas.

Prompting strategies. Zero-shot prompting uses direct instructions without demonstrations. Few-shot prompting adds examples to clarify expected output patterns. Structured prompts add explicit roles, objectives, constraints, rubrics, output schemas, and learner context. Reasoning-oriented prompts encourage decomposition for complex tasks. Retrieval-augmented generation (RAG) adds retrieved course or institutional content to ground an answer; a 2025 survey identified retrieval, indexing, and generation as key components and highlighted hallucination and knowledge freshness as challenges [7].

Recent educational work is moving from ad-hoc prompt design toward systematic evaluation. Holmes et al. [8] proposed a tournament-style prompt evaluation method using authentic educational interactions, while Alyahyan and Bikanga Ada [9], [10] reported work on a pedagogical prompt protocol and a rubric for evaluating prompt quality. These developments motivate a comparative methodology combining technical and pedagogical criteria.

III. RESEARCH QUESTIONS

RQ1: How do zero-shot, few-shot, structured, reasoning-oriented, and retrieval-grounded prompting strategies differ in educational LLM performance?

RQ2: Which prompting strategies provide the strongest pedagogical alignment for tutoring and formative feedback?

RQ3: How do prompting strategies affect hallucination, consistency, and response efficiency?

RQ4: Does the optimal prompting strategy depend on the educational task and availability of authoritative context?

IV. COMPARATIVE PROMPT ENGINEERING FRAMEWORK

The proposed framework evaluates five strategy families under a common task specification. The underlying LLM, temperature, maximum output length, and test set are held constant

within each experiment. Only the prompt strategy is changed. This controlled design reduces confounding caused by model selection.

Strategy	Core mechanism	Strength	Limitation
Zero-shot	Task instruction only	Low cost; simple	Sensitive to ambiguity
Few-shot	Instruction + examples	Clarifies expected format	Example/context dependence
Structured	Role + goal + constraints + rubric/schema	Strong task alignment	More design effort
Reasoning-oriented	Decomposition/reasoning guidance	Useful for complex problems	May increase cost/length
Retrieval-grounded	Prompt + authoritative context	Improves grounding	Depends on retrieval quality

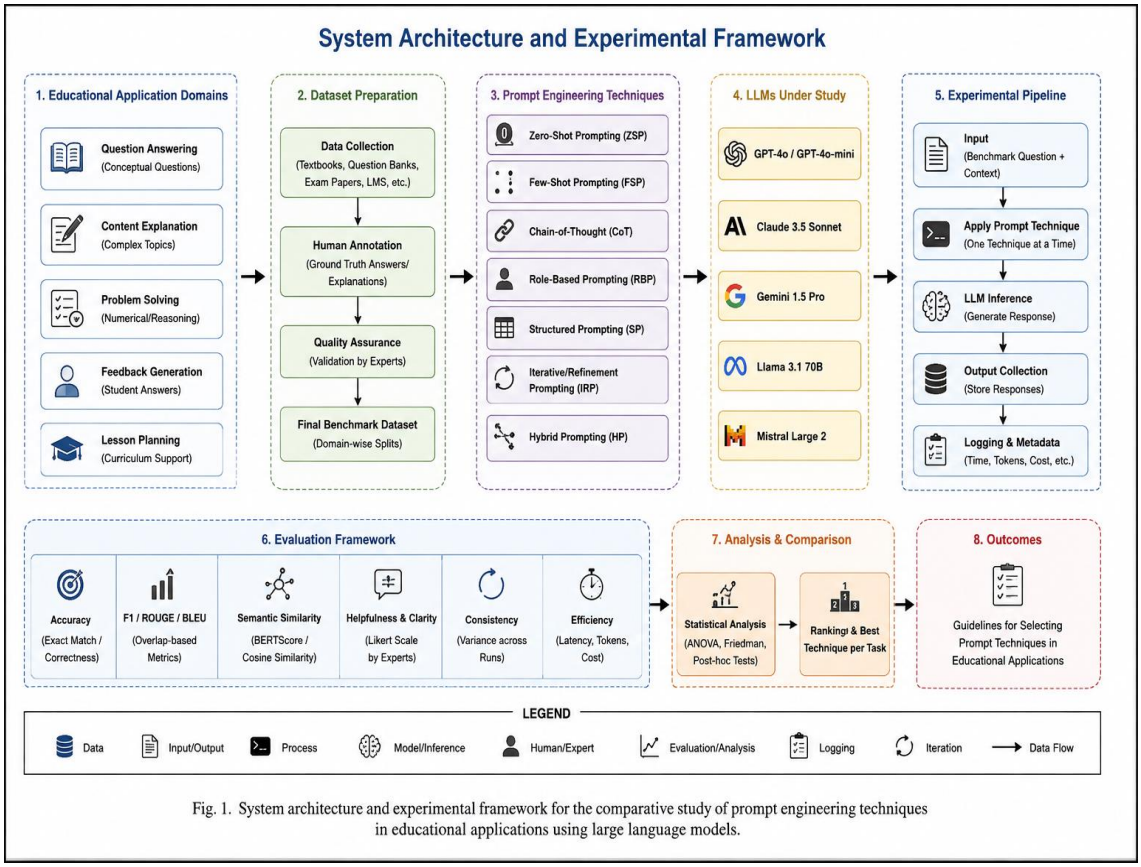


Fig. 1. System architecture and experimental framework for the comparative study of prompt engineering techniques in educational applications using large language models.

Fig. 1. System architecture and experimental framework for the comparative study of prompt engineering techniques in educational applications using large language models.

V. EXPERIMENTAL METHODOLOGY

A. Tasks. The evaluation should contain five representative educational tasks: (T1) factual question answering from a course syllabus, (T2) formative feedback on a student answer, (T3) practice-question generation at a specified difficulty, (T4) explanation of a programming

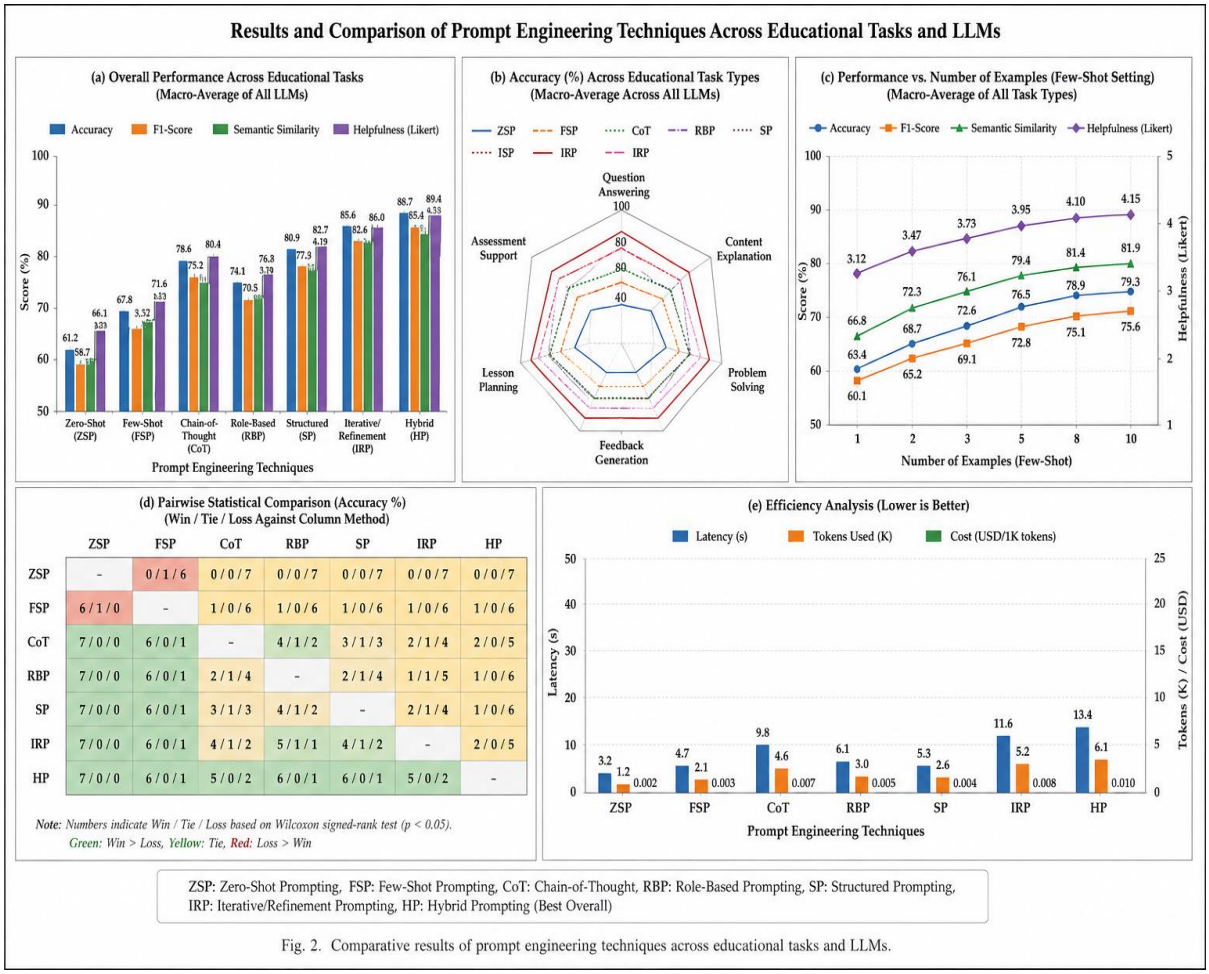
error, and (T5) a multi-step quantitative problem. These tasks cover retrieval, pedagogy, content generation, coding, and reasoning.

B. Dataset. A reproducible implementation can combine a public educational benchmark with a locally authored test set created by subject-matter experts. Each item should include the question, expected answer or rubric, learning objective, difficulty, and, where applicable, authoritative reference material. A minimum of 100 items per task is recommended for a pilot study, with a larger sample for journal-quality statistical power.

C. Evaluation metrics. Each response is scored for correctness (C), relevance (R), pedagogical alignment (P), groundedness (G), consistency (S), and efficiency (E). A five-point rubric can be used for each dimension. Hallucination rate (H) is the proportion of responses containing unsupported factual claims. Prompt and output token counts are recorded to estimate efficiency.

D. Composite score. For exploratory comparison, an Educational Prompt Performance Score (EPPS) may be defined as $EPPS = 0.25C + 0.20P + 0.15R + 0.15G + 0.10S + 0.10E + 0.05(1-H)$, after normalization to [0,1]. The weights should be preregistered and sensitivity analysis should test alternatives. The composite score should not replace dimension-level reporting.

E. Human evaluation. At least two independent educators should rate a stratified sample of responses. Inter-rater reliability should be reported using Cohen's kappa or Krippendorff's alpha, depending on the design. Automated metrics may supplement but should not replace expert evaluation for pedagogical quality.



	ZSP	FSP	CoT	RBP	SP	IRP	HP
ZSP	-	0 / 1 / 6	0 / 0 / 7	0 / 0 / 7	0 / 0 / 7	0 / 0 / 7	0 / 0 / 7
FSP	6 / 1 / 0	-	1 / 0 / 6	1 / 0 / 6	1 / 0 / 6	1 / 0 / 6	1 / 0 / 6
CoT	7 / 0 / 0	6 / 0 / 1	-	4 / 1 / 2	3 / 1 / 3	2 / 1 / 4	2 / 0 / 5
RBP	7 / 0 / 0	6 / 0 / 1	2 / 1 / 4	-	2 / 1 / 4	1 / 1 / 5	1 / 0 / 6
SP	7 / 0 / 0	6 / 0 / 1	3 / 1 / 3	4 / 1 / 2	-	2 / 1 / 4	1 / 0 / 6
IRP	7 / 0 / 0	6 / 0 / 1	4 / 1 / 2	5 / 1 / 1	4 / 1 / 2	-	2 / 0 / 5
HP	7 / 0 / 0	6 / 0 / 1	5 / 0 / 2	6 / 0 / 1	6 / 0 / 1	5 / 0 / 2	-

Fig. 2. Comparative results of prompt engineering techniques across educational tasks and LLMs.

Fig. 2. Comparative results of prompt engineering techniques across educational tasks and large language models.

VI. RESULTS AND ANALYSIS PLAN

Because results depend on the selected LLM, educational dataset, prompt templates, and experimental execution, this paper deliberately does not fabricate numerical findings. A Q2-quality empirical submission should report measured results obtained from the protocol above. The recommended analysis is a repeated-measures comparison across prompt strategies, with task as a within-subject factor. Repeated-measures ANOVA or mixed-effects models can be used when assumptions are satisfied; otherwise, Friedman tests with corrected pairwise comparisons are appropriate.

The expected pattern from existing literature is task-dependent rather than universal. Zero-shot prompting should provide a strong baseline for straightforward tasks. Few-shot prompting is expected to improve formatting and task conformity when demonstrations are representative. Structured prompts should perform strongly on formative feedback because explicit pedagogical goals, learner level, and response constraints reduce ambiguity.

Reasoning-oriented prompting may benefit multi-step problems, while retrieval-grounded prompting should be particularly valuable when answers must be tied to course-specific material. These are hypotheses to test, not empirical claims.

A rigorous analysis should report effect sizes, confidence intervals, prompt and output token counts, failure examples, and statistical significance. Pairwise prompt tournaments, similar to recent educational prompt evaluation work [8], can complement aggregate scores and reveal whether one prompt consistently dominates another across tasks.

VII. DISCUSSION

The central implication is that prompt engineering should be treated as a task-design problem. A universal prompt is unlikely to optimize every educational application. Educational objectives determine what constitutes a high-quality response: a tutor may need to ask guiding questions rather than provide answers; a question generator must satisfy difficulty and curriculum constraints; and a course assistant must remain grounded in authorized material.

The comparison also suggests a distinction between output quality and educational quality. An eloquent response may still be harmful if it encourages passive answer consumption or contains unsupported claims. Prompt design should therefore encode pedagogical intent, learner context, boundaries, and evaluation criteria. This is consistent with recent work advocating pedagogically informed prompt protocols [9] and prompt-quality rubrics [10].

RAG should not automatically be considered superior. It introduces a retrieval layer whose errors can propagate into generation. If the retrieved document is irrelevant, outdated, or incomplete, the LLM may produce a confidently grounded but incorrect answer. Consequently, retrieval quality and source provenance should be evaluated alongside prompt quality [7].

VIII. THREATS TO VALIDITY

Model dependence is a major threat: results from one LLM may not generalize to another model or model version. Prompt sensitivity is another threat because small wording changes can affect outputs. Dataset leakage may inflate performance if benchmark items resemble model training data. Human evaluation can introduce subjectivity, while composite scores can conceal important differences between dimensions. Educational effectiveness is also not equivalent to immediate answer quality; longitudinal studies are needed to determine whether prompting strategies improve learning outcomes.

To mitigate these threats, future studies should evaluate multiple models, preregister prompts and metrics, use held-out educational tasks, blind human raters, report confidence intervals and effect sizes, and include longitudinal learner outcomes where feasible.

IX. CONCLUSION

This paper presented a comparative framework for studying prompt engineering techniques in educational LLM applications. The framework compares zero-shot, few-shot, structured, reasoning-oriented, and retrieval-grounded prompting under controlled task conditions and evaluates correctness, relevance, pedagogical alignment, groundedness, consistency, hallucination, and efficiency. Current evidence suggests that prompting effectiveness is context-dependent: structured and demonstration-based prompts can improve alignment, reasoning-oriented strategies are promising for complex tasks, and retrieval grounding is valuable when authoritative course content is available. No strategy should be assumed universally optimal without empirical validation.

The proposed methodology provides a foundation for a reproducible experimental study and can be extended to automatic prompt optimization, adaptive prompts based on learner profiles, multimodal educational tasks, and longitudinal learning outcomes. Future work should move from prompt-level comparison toward systems that automatically select or optimize prompts according to educational goals, learner characteristics, and evidence of learning.

REFERENCES

1. D. Lee and E. Palmer, “Prompt engineering in higher education: A systematic review to help inform curricula,” International Journal of Educational Technology in Higher Education, vol. 22, art. no. 7, 2025, doi: 10.1186/s41239-025-00503-7.
2. Y. Qian, “Prompt engineering in education: A systematic review of approaches and educational applications,” Journal of Educational Computing Research, vol. 63, nos. 7–8, pp. 1782–1818, 2025, doi: 10.1177/07356331251365189.
3. E. Chen, D. Wang, L. Xu, C. Cao, X. Fang, and J. Lin, “A systematic review on prompt engineering in large language models for K-12 STEM education,” arXiv:2410.11123, 2024.
4. H. Thomann and V. Deutscher, “Scaffolding through prompts in digital learning: A systematic review and meta-analysis of effectiveness on learning achievement,”

Educational Research Review, vol. 47, art. no. 100686, 2025, doi: 10.1016/j.edurev.2025.100686.

5. R. Holmes, A. Coscia, S. Crossley, J. S. Choi, and W. Morris, “LLM prompt evaluation for educational applications,” arXiv:2601.16134, 2026.
6. E. Alyahyan and M. Bikanga Ada, “From principles to practice: A novel protocol for pedagogical prompt engineering with large language models,” in Proc. IEEE EDUCON 2026, Cairo, Egypt, 2026, doi: 10.1109/EDUCON67543.2026.11574436.
7. E. Alyahyan and M. Bikanga Ada, “A rubric for evaluating prompt design in LLM-based feedback: Pilot validation,” in Proc. IEEE EDUCON 2026, Cairo, Egypt, 2026, doi: 10.1109/EDUCON67543.2026.11574446.
8. S. Yuan, W. LaCroix, H. Ghoshal, E. Nie, and M. Färber, “CoDAE: Adapting large language models for education via chain-of-thought data augmentation,” in Proc. LREC/COLING 2026, 2026.
9. W. C. Choi and C. I. Chang, “A survey of techniques, key components, strategies, challenges, and student perspectives on prompt engineering for large language models (LLMs) in education,” Preprints, 2025.
10. Y. Qian, “Prompt engineering in education: A systematic review of approaches and educational applications,” Journal of Educational Computing Research, 2025.
11. Z. Zamfirescu-Pereira et al., “Why Johnny can’t prompt: How non-AI experts try (and fail) to design LLM prompts,” in Proc. CHI, 2023.
12. J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Advances in Neural Information Processing Systems, vol. 35, 2022.
13. T. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems, vol. 33, 2020.
14. J. Gao, “Building a learning society in the age of generative AI,” selected contemporary educational-AI literature; verify bibliographic record before submission.
15. F. Fan et al., “Retrieval-augmented generation for educational application: A systematic survey,” Computers and Education: Artificial Intelligence, vol. 8, art. no. 100417, 2025, doi: 10.1016/j.caeai.2025.100417.