Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL33980

The Impact of Prompt Strategies on the Quality of Generative AI Text: A Conceptual and Experimental Framework

Qianbin Dong

Faculty of Science, the University of Melbourne, Melbourne, Australia
dqbschoolemail@163.com

Abstract. The rapid development of large language models has made generative artificial intelligence increasingly common in writing, summarization, education, journalism, and knowledge production. However, the quality of AI-generated text does not depend only on the model itself; it is also shaped by the way users formulate prompts. This paper examines how different prompt strategies may influence the quality of generated text, especially in Chinese text generation tasks, such as news summarization and instruction-following writing. Drawing on recent research on prompt engineering, reasoning-oriented prompting, self-refinement, retrieval-augmented generation, and LLM-based evaluation, this study proposes a multidimensional framework for comparing prompt strategies. The main strategies discussed include zero-shot direct prompting, structured instruction prompting, few-shot prompting, plan-then-write prompting, self-refinement prompting, and retrieval-augmented prompting. Text quality is defined through task performance, linguistic fluency, coherence, factuality, instruction compliance, overall preference, robustness, and cost. This paper argues that different prompt strategies are likely to improve different aspects of text quality rather than produce one universally superior outcome. Its contribution is to provide a clear conceptual and methodological framework for evaluating prompt strategies in a way that is understandable for non-technical users and useful for future empirical research.

Keywords: generative AI, large language models, prompt engineering, text quality, LLM-as-a-judge

1. Introduction

Generative artificial intelligence has become an important tool in contemporary digital communication. Large language models (LLMs), such as GPT-series systems and Qwen models, can generate summaries, reports, essays, explanations, translations, and other forms of written content from natural language instructions. These systems are trained on large-scale data and generate text by predicting contextually appropriate sequences of words or tokens. Although model capacity is important, users mainly interact with these models through prompts. A prompt is the instruction, question, context, example, or constraint given to the model before it generates an answer.

Prompt design is therefore not a minor technical detail. A simple request such as "summarize this article" may produce a different result from a more specific instruction, such as "act as a professional news editor and summarize the article in no more than 150 Chinese characters,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL33980

covering the main event, participants, cause, and result, without adding unsupported information." The second prompt gives the model a role, task scope, length limit, content requirement, and factual constraint. This example shows why prompt engineering has become a significant research and practical field. Recent surveys describe prompt engineering as a way to guide model behavior without changing model parameters [1, 2]. Technical reports on GPT-4 and Qwen2.5 also show that modern LLMs are powerful, but their performance must still be understood through actual use cases and task settings [3, 4].

Existing research faces three problems. First, prompt effects are often unstable: small changes in wording, punctuation, or format may change output quality. Second, open-ended text is difficult to evaluate because it has no single correct answer. Third, studies often use different models, datasets, and standards, making conclusions hard to compare. These problems require a controlled and multidimensional framework.

This paper asks how different prompt strategies affect the quality of AI-generated text. It focuses on six strategies: zero-shot direct prompting, structured instruction prompting, few-shot prompting, plan-then-write prompting, self-refinement prompting, and retrieval-augmented prompting. Rather than claiming that one strategy is always best, the paper argues that different strategies may improve different dimensions of text quality. The aim is to build a conceptual and experimental framework that can guide future empirical studies, especially in Chinese text generation tasks such as news summarization and instruction-following writing.

2. Literature review

Prompt engineering has rapidly developed into a major area of LLM research. The Prompt Report provides a broad taxonomy of prompting techniques and demonstrates that prompt design includes many forms, such as role assignment, examples, decomposition, reasoning prompts, tool use, and self-correction [1]. Sahoo et al. similarly review prompt engineering across multiple applications and emphasize its value as a low-cost way to adapt model behavior without retraining [2]. Chen et al. further discuss how prompt engineering can release the potential of LLMs in practical applications [5]. These studies show that prompt strategies can be classified, but they do not fully compare their effects on multidimensional text quality under the same conditions.

One influential group of methods is reasoning-oriented prompting. Chain-of-thought prompting asks models to produce intermediate reasoning steps before giving a final answer and has been shown to improve performance on complex reasoning tasks [6]. Self-consistency extends this idea by sampling multiple reasoning paths and selecting the most consistent answer [7]. Tree of Thoughts allows models to explore several possible reasoning trajectories before choosing a solution [8]. Although often developed for mathematics and logic, the same principle can be adapted to writing. A plan-then-write prompt may ask the model to identify key information before generating the final text, potentially improving organization, completeness, and coherence.

Another important line of work concerns iterative improvement. Self-Refine asks a model to generate an initial output, provide feedback on that output, and then revise it [9]. This resembles human writing, where a draft is reviewed before final submission. Self-refinement may improve clarity, completeness, and overall quality, but it requires additional model calls and token usage. Therefore, it should be evaluated by both quality and cost.

Factuality is another central issue in generative AI. LLMs may produce hallucinations, meaning fluent but unsupported statements. SelfCheckGPT proposes a black-box approach for detecting hallucination risk by checking consistency across multiple model outputs [10]. Retrieval-augmented generation (RAG) responds to this problem by providing external evidence and asking the model to

DOI: 10.54254/2755-2721/2026.GL33980

generate text based on that evidence. In news summarization, factuality is crucial because a fluent summary is unacceptable if it adds unsupported information. RAG may improve faithfulness, but it depends on evidence quality.

Evaluation remains a major challenge. ROUGE is a classic automatic metric for summarization, measuring overlap between a generated summary and a reference summary [11]. BLEURT attempts to provide a learned metric that is more robust for evaluating generated text [12]. More recently, LLM-as-a-judge methods have become popular. G-Eval uses GPT-based evaluation to assess natural language generation with better alignment to human judgment [13]. However, automatic LLM evaluators may also show biases, such as length preference and position effects. Length-controlled AlpacaEval shows the need to control length bias in automatic evaluation [14]. For this reason, a reliable evaluation design should combine automatic metrics, LLM-based judging, and human evaluation.

For Chinese text generation, dataset selection is important. CNewSum is a large-scale Chinese news summarization dataset with human summaries and annotations related to adequacy and deducibility [15]. IFEval evaluates whether models follow verifiable instructions, such as length limits, required keywords, or output formats [16]. Together, these task types allow researchers to examine both content quality and formal compliance.

3. Research questions and methodological framework

This paper proposes five research questions: RQ1 asks whether different prompt strategies produce significant differences in AI-generated text quality. RQ2 asks which dimensions of quality are most affected. RQ3 asks whether the effects are consistent across different model families, such as closed-source and open-source models. RQ4 asks how consistent LLM-as-a-judge evaluations are with human evaluations. RQ5 asks how different strategies compare in terms of quality improvement and cost.

Based on existing literature, five hypotheses can be proposed. First, structured instruction prompts are expected to improve instruction compliance and readability compared with zero-shot direct prompts, but they may have a limited effect on factuality. Second, few-shot prompts may improve task alignment and information coverage, but they may also increase output length and stylistic uniformity. Third, plan-then-write prompts may improve organization and coherence, especially when the task requires information selection and structuring. Fourth, self-refinement prompts may improve overall quality and human preference, but at the cost of additional model calls and token consumption. Fifth, retrieval-augmented prompts may improve factuality and verifiability when reliable evidence is available, but may reduce coherence if the evidence is noisy or conflicting.

A suitable design is a controlled repeated-measures experiment. The same input texts should be given to the same model under different prompt strategies. This reduces input-difficulty effects because each text is compared with itself across prompt conditions. The resulting outputs can then be evaluated with the same metrics.

The first strategy, zero-shot direct prompting, serves as the baseline. It gives the model a simple task instruction without examples or additional structure. The second strategy, structured instruction prompting, includes a role, target audience, task goal, constraints, and output format. The third strategy, few-shot prompting, provides two or three input-output examples before asking the model to complete a new task. The fourth strategy, plan-then-write prompting, asks the model to identify important information before generating the final text. The fifth strategy, self-refinement prompting, asks the model to draft, evaluate, and revise its output. The sixth strategy, retrieval-augmented

DOI: 10.54254/2755-2721/2026.GL33980

prompting, gives the model external evidence and instructs it to generate text based only on that evidence.

Two types of tasks are especially appropriate. The first is Chinese news summarization using CNewSum. This task evaluates whether prompt strategies help models produce accurate, concise, and informative summaries. The second is instruction-following text generation based on IFEval. These tasks can include verifiable requirements such as word count limits, paragraph structure, required keywords, or forbidden expressions. The study may compare a strong closed-source model such as GPT-4o with an open or reproducible model such as Qwen2.5 Instruct, because prompt effects may vary across model families.

Text quality should be evaluated across several dimensions. Task quality refers to whether the output completes the required task. Linguistic quality refers to fluency, coherence, readability, and naturalness. Factuality refers to whether the output is faithful to the source and avoids unsupported claims. Instruction compliance refers to whether the output follows explicit constraints. Overall preference refers to human or model-based judgment about which output is better. Robustness refers to whether quality remains stable when the prompt undergoes non-semantic changes, such as punctuation or formatting changes. Cost refers to model calls, token usage, time, and financial expense.

The evaluation should combine automatic metrics, LLM-based judging, and human review. Automatic metrics may include ROUGE, BLEURT, hallucination-risk indicators, and rule-based compliance checks. LLM-based evaluation can use a rubric similar to G-Eval, rating outputs on relevance, coherence, factuality, coverage, and expression. Human evaluation should be conducted on a sample of outputs to calibrate automatic and LLM-based scores. Statistical analysis can use repeated-measures methods. If the data are not normally distributed, the Friedman test can detect overall differences among strategies, followed by Wilcoxon signed-rank tests for pairwise comparison. If normality assumptions are satisfied, repeated-measures ANOVA or linear mixed-effects models may be used.

4. Expected results and discussion

Because this paper proposes a framework rather than reporting completed experiments, the following claims should be understood as expected patterns based on existing literature. First, structured instruction prompting is expected to improve readability and instruction compliance. By specifying role, audience, format, and constraints, the prompt reduces ambiguity. However, it may not strongly improve factuality without reliable evidence.

Second, few-shot prompting is expected to improve task alignment. Examples help the model understand the desired style and structure, but may also introduce bias. The model may imitate example length, tone, or structure too strongly, reducing diversity.

Third, plan-then-write prompting is expected to improve organization and coherence. By first identifying key information, the model may omit fewer important points. To avoid excessive length, the prompt can require planning internally while outputting only the final text.

Fourth, self-refinement is expected to produce high overall quality because revision can correct omissions, improve fluency, and adjust structure. However, it requires additional calls and token usage, so it is better understood as a quality-oriented strategy than a universally efficient one.

Fifth, retrieval-augmented prompting is expected to improve factuality. When a model relies only on provided evidence, it may be less likely to hallucinate. However, if evidence is incomplete or contradictory, the output may still be inaccurate or incoherent.

DOI: 10.54254/2755-2721/2026.GL33980

The sixth expected result concerns robustness. Prompt perturbation, such as changing punctuation, line breaks, or bullet formatting without changing meaning, may reveal instability in model outputs. Recent work on prompt sensitivity suggests that non-semantic changes can affect model performance [17]. A reliable strategy should be judged by both average score and stability.

Overall, this framework suggests that it is unlikely to be the best prompt strategy for all situations. Different strategies serve different goals. Structured prompts may be most practical for ordinary users who need clearer and more controlled outputs. Few-shot prompts may help when a specific style must be imitated. Plan-then-write prompts may help with complex writing. Self-refinement may be best when quality is more important than cost. Retrieval-augmented prompts may be best when factuality is the main concern.

5. Implications

This study has both theoretical and practical implications. Theoretically, it contributes to prompt engineering research by treating text quality as multidimensional. Instead of asking only whether a prompt "works," the proposed framework asks what kind of improvement occurs, whether it is stable, and how much it costs. This approach can help move prompt engineering from informal practice toward systematic evaluation.

Practically, the study is valuable for non-technical users. Students, teachers, writers, journalists, and office workers can improve outputs by writing clearer prompts, specifying audience and format, providing examples, asking for planning, requesting revision, or supplying reliable source materials. The study also helps organizations make cost-sensitive decisions: self-refinement may suit high-stakes documents, while retrieval-augmented prompting may be essential for factual reports.

The study also has implications for AI evaluation. No single metric is sufficient. ROUGE may capture reference overlap but not factuality or readability. LLM judges may capture qualitative differences but may also show bias. Human evaluation is valuable but expensive. Therefore, triangulation across automatic metrics, LLM-based judging, and human review is necessary.

6. Conclusion

This paper examines how different prompt strategies may influence the quality of text generated by large language models. It argues that prompt design is a central factor in generative AI performance and should be studied systematically. The paper proposes a framework comparing zero-shot direct prompting, structured instruction prompting, few-shot prompting, plan-then-write prompting, self-refinement prompting, and retrieval-augmented prompting.

The main conclusion is that prompt strategies affect different dimensions of text quality in different ways. Structured instruction prompting may improve readability and compliance. Few-shot prompting may improve task alignment. Plan-then-write prompting may improve coherence and completeness. Self-refinement may improve overall quality but increase cost. Retrieval-augmented prompting may improve factuality but depends on evidence quality. Therefore, prompt strategy selection should be guided by task goals, quality requirements, and cost constraints.

Future research should implement this framework using real datasets such as CNewSum and instruction-following benchmarks. It should compare multiple models, include human evaluation, and test robustness under prompt perturbations. Such work would provide stronger empirical evidence for understanding how prompt strategies shape generative AI text quality.

DOI: 10.54254/2755-2721/2026.GL33980

References

[1] Schulhoff, S., et al. (2024). The prompt report: A systematic survey of prompting techniques. arXiv. https: //arxiv.org/abs/2406.06608
[2] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv. https: //arxiv.org/abs/2402.07927
[3] Achiam, J., et al. (2023). GPT-4 technical report. arXiv. https: //arxiv.org/abs/2303.08774
[4] Yang, A., et al. (2024). Qwen2.5 technical report. arXiv.
[5] Chen, B., et al. (2025). Unleashing the potential of prompt engineering for large language models. ScienceDirect.
[6] Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. arXiv. https: //arxiv.org/abs/2201.11903
[7] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain-of-thought reasoning in language models. arXiv. https: //arxiv.org/abs/2203.11171
[8] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv. https: //arxiv.org/abs/2305.10601
[9] Madaan, A., et al. (2023). Self-Refine: Iterative refinement with self-feedback. arXiv. https: //arxiv.org/abs/2303.17651
[10] Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheck GPT: Zero-resource black-box hallucination detection for generative large language models. arXiv. https: //arxiv.org/abs/2303.08896
[11] Lin, C.Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the ACL Workshop on Text Summarization Branches.
[12] Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. In Proceedings of the Association for Computational Linguistics.
[13] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. arXiv. https: //arxiv.org/abs/2303.16634
[14] Dubois, Y., et al. (2024). Length-controlled AlpacaEval: A simple way to debias automatic evaluators. arXiv. https: //arxiv.org/abs/2404.04475
[15] Wang, D., Chen, J., Wu, X., Zhou, H., & Li, L. (2021). CNewSum: A large-scale Chinese news summarization dataset with human-annotated adequacy and deducibility level. arXiv. https: //arxiv.org/abs/2110.10874
[16] Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., & Hou, L. (2023). Instruction-following evaluation for large language models. arXiv. https: //arxiv.org/abs/2311.07911
[17] Seleznyov, M., et al. (2025). When punctuation matters: A large-scale comparison of prompt robustness methods for LLMs. arXiv.