Impact Of Prompt Design on The Accuracy and Reliability of Generative AI Responses

Siddhi Khilari[1], Tanishka Said[2]
[1,2]*Department of Computer Science, MAEER'S MIT Arts, Commerce and Science College, Alandi (D), Pune, India*
doi.org/10.64643/ijirt.208486-459

***Abstract*—Generative Artificial Intelligence (GenAI) has become an increasingly important technology for applications such as information retrieval, content generation, education, problem-solving and decision support. The effectiveness of these systems is strongly influenced by the way users formulate instructions, commonly referred to as prompts. Even minor changes in prompt structure, context, specificity or examples can produce significant differences in the quality of generated responses. This research investigates the impact of prompt design on the accuracy, relevance, consistency and reliability of responses generated by Generative AI systems. The study examines and compares different prompting approaches, including zero-shot, one-shot, few-shot and structured prompting, across selected tasks, drawing on existing literature to evaluate these approaches using criteria such as factual accuracy, relevance to the given task, completeness, consistency and reliability. The research also analyzes the influence of contextual information, explicit instructions, examples and constraints on the generated output.**

Through a systematic review and comparison of different prompt designs, the study aims to identify prompting strategies that can produce more accurate and dependable responses. The findings are expected to provide practical guidelines for designing effective prompts and improving user interaction with Generative AI systems.

The study may be particularly useful for students, educators, researchers and professionals who increasingly rely on AI-generated information and content. Overall, this research seeks to contribute to a better understanding of prompt engineering and encourage more effective, reliable and responsible utilization of Generative AI in academic and practical environments.

***Index Terms*—Generative AI; Prompt Engineering; Prompt Design; Zero-Shot Prompting; Few-Shot Prompting; Chain-of-Thought; AI Reliability**

I. INTRODUCTION

Generative Artificial Intelligence (GenAI) systems, particularly Large Language Models (LLMs) such as GPT-4, Claude and Gemini, have rapidly transformed the way individuals search for information, generate content, solve problems and make decisions [1]. Unlike earlier rule-based or narrowly trained systems, these models are capable of producing fluent, human-like responses across an extremely wide range of tasks without task-specific retraining [1], [2]. Their growing accessibility has led to widespread adoption in academic, professional and everyday contexts, with recent surveys of higher-education institutions reporting adoption rates exceeding eighty percent within a short span of time following the public release of tools such as ChatGPT [10], [11].

Despite this rapid adoption, the quality of the output generated by a GenAI system is not solely determined by the underlying model's capability. It depends heavily on how the task is communicated to the model, that is, on the design of the prompt [2], [3]. A prompt is the natural language instruction, question, or set of instructions and examples that a user provides to elicit a desired response from the model. Because LLMs generate text by predicting the most probable continuation of a given input, the phrasing, structure, context and examples embedded within a prompt directly shape the probability distribution over possible outputs [3], [7]. Consequently, two prompts that appear semantically similar to a human reader can produce responses that differ substantially in accuracy, completeness and reliability [7], [9].

This sensitivity to prompt design has given rise to the discipline of prompt engineering, which is concerned with systematically crafting inputs that reliably elicit high-quality outputs from GenAI systems [1], [2].

Several prompting paradigms have emerged in the literature, ranging from simple zero-shot prompting, where the model is given only a task description, to one-shot and few-shot prompting, where one or more demonstrations are included in the prompt, to more advanced structured techniques such as chain-of-thought prompting, which explicitly guides the model through intermediate reasoning steps [3], [4]. Each of these approaches carries distinct trade-offs in terms of accuracy, consistency, reliability and the amount of contextual information required from the user.

At the same time, GenAI systems remain prone to hallucination, that is, the generation of fluent but factually incorrect or unsupported content, which continues to undermine the trustworthiness of these systems in domains that demand factual accuracy [5], [6]. Understanding how prompt design can mitigate or, conversely, exacerbate such failures is therefore of considerable practical importance, particularly as students, educators, researchers and professionals increasingly rely on AI-generated information for academic and practical tasks [10], [11].

This paper presents a literature-based investigation into the impact of prompt design on the accuracy and reliability of GenAI responses. Section II reviews related work on prompting techniques, hallucination and prompt sensitivity. Section III presents a comparative analysis of zero-shot, one-shot, few-shot and structured prompting approaches with respect to accuracy, relevance, consistency and reliability. Section IV discusses the role of contextual information, explicit instructions, examples and constraints in shaping GenAI output. Section V distills practical guidelines for designing effective prompts. Section VI outlines challenges and future research directions, and Section VII concludes the paper.

II. RESEARCH METHODOLOGY

This study adopts a literature-based, comparative-analytical methodology rather than an empirical experiment on live GenAI systems. The approach was chosen to allow a broad synthesis of findings already established across the prompt engineering literature, spanning foundational papers, systematic surveys, and recent empirical studies of prompt sensitivity, hallucination and educational adoption. Relevant literature was identified through targeted searches of academic and preprint repositories, including arXiv and peer-reviewed venues indexed through Google Scholar and ScienceDirect, using search terms such as "prompt engineering," "zero-shot prompting," "few-shot learning," "chain-of-thought reasoning," "prompt sensitivity," and "hallucination in large language models."

Sources were selected based on three criteria: relevance to the accuracy, consistency or reliability of GenAI output; methodological rigor, with preference given to peer-reviewed publications, well-cited preprints, and foundational papers introducing widely adopted techniques; and recency, with an emphasis on studies published between 2020 and 2026 to ensure the review reflects the current state of large language model capabilities. Foundational works that introduced core prompting paradigms, such as few-shot in-context learning and chain-of-thought prompting, were retained regardless of publication date given their continued influence on the field.

The reviewed literature was organized thematically into six areas: the evolution of prompt engineering as a discipline; zero-shot, one-shot and few-shot prompting; structured and chain-of-thought prompting; hallucination and factual reliability; prompt sensitivity; and the practical adoption of GenAI in academic and professional contexts. Findings from each theme were then synthesized into a qualitative comparative analysis of prompting approaches across four evaluation criteria, namely accuracy, relevance, consistency and reliability, as reported in the reviewed studies. This synthesis forms the basis for the practical guidelines proposed in Section V. Because the study relies on secondary evidence rather than original experimentation, the analysis prioritizes convergent findings replicated across multiple independent studies over isolated results from any single source.

III. LITERATURE REVIEW

A. Evolution of Prompt Engineering

Prompt engineering has emerged as a pivotal discipline for optimizing the performance of large language models by structuring inputs to enhance coherence, accuracy and task alignment, without requiring any modification of the underlying model parameters [1], [2]. Rather than updating model weights through fine-tuning, prompting allows a

single pre-trained model to be adapted to a wide variety of downstream tasks purely through the natural language instructions and, optionally, demonstrations supplied at inference time [1]. This shift has made LLMs accessible to a much broader base of users, including those without a background in machine learning, since effective use of these systems increasingly depends on the ability to phrase instructions clearly rather than on programming or model-training expertise [2], [3]. Comprehensive surveys in this area have catalogued dozens of distinct prompting techniques, organizing them according to the NLP tasks and application domains in which they have proven effective [2], [3].

B. Zero-Shot, One-Shot and Few-Shot Prompting

The foundational taxonomy of prompting strategies, zero-shot, one-shot and few-shot, was systematically formalized and evaluated in the work introducing GPT-3 [3]. In the zero-shot setting, the model is given only a natural language description of the task, with no demonstrations provided; in the one-shot setting, exactly one example of the task is included in the prompt; and in the few-shot setting, several demonstrations, typically ranging from ten to a hundred depending on context window constraints, are supplied [3]. Empirical results from this line of work consistently show that one-shot and few-shot performance is often substantially higher than true zero-shot performance, and that the performance gap between these settings tends to widen as model capacity increases, suggesting that larger models are more proficient at learning from in-context demonstrations [3]. For example, on closed-book question-answering benchmarks, accuracy has been shown to improve markedly moving from the zero-shot to the few-shot setting, in some cases matching or exceeding state-of-the-art fine-tuned systems operating under comparable conditions [3].

These findings have been extended and refined by subsequent research. It has been shown that few-shot in-context learning capability is not restricted to extremely large, unidirectional models but can also be observed in smaller and bidirectional architectures under appropriate conditions [3]. Other work has systematically studied the effect of example ordering within few-shot prompts, finding that permutations of the same set of demonstrations can produce large swings in task accuracy, and that the optimal ordering identified for one model often fails to transfer to another [7]. This indicates that the benefits of few-shot prompting are not solely a function of the information content of the examples, but also of their presentation, an observation with direct implications for prompt design practice.

C. Structured and Chain-of-Thought Prompting

Beyond simple demonstration-based prompting, structured techniques have been developed to explicitly guide a model's internal reasoning process. Chain-of-thought (CoT) prompting was introduced as a technique in which a small number of exemplars in the prompt include not only the final answer but also the intermediate reasoning steps leading to it [4]. This encourages the model to generate its own step-by-step reasoning before producing a final answer, which has been shown to substantially improve performance on arithmetic, commonsense and symbolic reasoning tasks [4], [22]. Notably, the benefits of CoT prompting have been found to emerge primarily in sufficiently large models; smaller models tend to produce fluent but logically inconsistent reasoning chains that do not translate into improved accuracy [4]. Subsequent work demonstrated that chain-of-thought reasoning need not depend on few-shot demonstrations at all: simply appending an instruction such as "Let's think step by step" to a zero-shot prompt (zero-shot CoT) has been shown to meaningfully improve reasoning performance, indicating that structured prompting is, at least in part, a property of instruction design rather than solely of demonstration design [27].

D. Hallucination, Factual Accuracy and Reliability

A persistent limitation of GenAI systems is their tendency to hallucinate, producing content that is fluent and syntactically well-formed but factually inaccurate or unsupported by any verifiable evidence [5], [6]. Because hallucination directly undermines the reliability and trustworthiness of GenAI output, it has become a central concern in domains requiring high factual accuracy, such as healthcare, law and education [5], [32]. Existing surveys categorize mitigation strategies broadly into model-centric approaches, which involve architectural or training-based interventions, and prompting- or retrieval-based approaches, which operate purely at inference time [5]. Among the latter, retrieval-augmented techniques that ground model responses in external, verifiable

sources have been found to meaningfully improve factual accuracy, while purely prompting-based interventions, such as explicitly instructing the model to cite sources, express uncertainty, or verify its own reasoning, offer a lighter-weight but generally less robust alternative [5], [6]. The literature broadly agrees that no single technique fully eliminates hallucination, and that hybrid approaches combining careful prompt design with retrieval and model-level safeguards currently offer the most promising path toward more reliable GenAI output [5].

E. Prompt Sensitivity

A growing body of empirical research demonstrates that LLM outputs can be highly sensitive to seemingly minor, meaning-preserving changes in prompt formatting, wording, punctuation and example ordering [7], [8], [9]. This phenomenon, often referred to as prompt sensitivity, has been shown to produce large and sometimes disproportionate swings in task accuracy: variations in label choice or phrasing have been found to shift performance by tens of percentage points, and changes in prompt formatting alone have produced accuracy variations of comparable magnitude across benchmark tasks [7]. Similar sensitivity has been documented in domain-specific contexts, including healthcare-related question answering, where minor rewordings of otherwise equivalent prompts were shown to alter model outputs in ways that could have practical consequences if deployed without safeguards [9]. In the context of code generation, it has similarly been observed that the degree of specification detail and structural richness of a prompt, rather than wording alone, influences the correctness of generated code, and that under certain conditions under-specified prompts can outperform highly detailed ones, suggesting that the relationship between prompt detail and output quality is not strictly monotonic [8]. Collectively, this literature underscores that prompt design is not a peripheral concern but a first-order determinant of GenAI reliability, motivating systematic study of which design choices most consistently improve accuracy and consistency.

F. Generative AI Adoption in Academic and Practical Contexts

The practical significance of prompt design is amplified by the scale at which GenAI tools are now used in educational and professional settings. Recent survey-based studies report that a large majority of university students use GenAI tools regularly in their academic work, with adoption rates in some institutions exceeding eighty percent within a few years of the public release of conversational AI systems [10], [11]. Students and educators alike increasingly rely on these tools for tasks ranging from information retrieval and explanation generation to drafting and problem-solving [10], [11]. Given this scale of reliance, even modest improvements in prompting practice, achieved through greater awareness of how prompt structure influences output quality, have the potential to meaningfully improve the accuracy and dependability of AI-assisted academic and professional work.

IV. COMPARATIVE ANALYSIS OF PROMPTING APPROACHES

Drawing on the literature reviewed in Section II, this section synthesizes a comparative analysis of the four prompting approaches considered in this study, zero-shot, one-shot, few-shot and structured/chain-of-thought prompting, with respect to four evaluation criteria: accuracy, relevance, consistency and reliability. Table I summarizes the qualitative characteristics of each approach as reported across the reviewed studies.

TABLE I. Qualitative Comparison of Prompting Approaches

Prompting Approach	Accuracy	Consistency	Contextual Dependence	Typical Use Case
Zero-shot	Moderate; declines on complex/reasoning tasks	Low to moderate; sensitive to wording	Relies solely on model's pre-trained knowledge	Simple, well-known tasks
One-shot	Improved over zero-shot	Moderate	Single example anchors output format	Tasks needing format guidance
Few-shot	Highest among basic methods; approaches fine-tuned performance	Higher, but order/selection of examples matters	Strongly dependent on quality and order of examples	Classification, structured NLP tasks

Structured / Chain-of-Thought	Highest on multi-step reasoning tasks	Generally high when reasoning steps are explicit	Requires explicit instructions, constraints, and step decomposition	Arithmetic, logical, and multi-step reasoning tasks

Zero-shot prompting, while requiring the least effort from the user, generally yields the lowest and most variable accuracy among the approaches considered, particularly on tasks that require multi-step reasoning or domain-specific knowledge not easily inferred from a bare task description [3]. Its performance is entirely dependent on the model's internal, pre-trained knowledge and its ability to correctly infer the intended task from limited instruction, which makes it especially vulnerable to prompt sensitivity effects [7].

One-shot and few-shot prompting improve accuracy and relevance by anchoring the model's output format and reasoning pattern to concrete examples [3]. Few-shot prompting in particular has been shown to approach the performance of fine-tuned models on several benchmark tasks, but this benefit is conditional on the quality, diversity and ordering of the examples provided; poorly chosen or inconsistently ordered demonstrations can degrade rather than improve performance [7]. This makes few-shot prompting more powerful but also more demanding in terms of prompt construction effort compared to zero-shot prompting.

Structured and chain-of-thought prompting achieves the strongest results on tasks involving multi-step reasoning, arithmetic or logical inference, because it explicitly decomposes the task into intermediate steps rather than requiring the model to produce a final answer in a single inference step [4]. This approach tends to yield higher consistency on reasoning-heavy tasks, since errors can often be traced to a specific step in the reasoning chain, but it requires either well-constructed demonstrations or carefully worded instructions, and its benefits are most pronounced in sufficiently large models [4], [22].

Overall, the literature suggests a general trend in which accuracy and consistency improve as more contextual structure, whether in the form of examples or explicit reasoning steps, is added to the prompt, but with diminishing and task-dependent returns, and with an increased sensitivity to how that structure is presented [7], [8]. This motivates the discussion of contextual information and constraints in the following section.

V. DISCUSSION: INFLUENCE OF CONTEXTUAL INFORMATION, INSTRUCTIONS, EXAMPLES AND CONSTRAINTS

A. Contextual Information

The amount and relevance of contextual information supplied in a prompt has a direct bearing on output accuracy. When a prompt provides sufficient background, such as the intended audience, domain, or the specific format required, the model is better able to narrow the space of plausible responses to those that are actually useful to the user [2], [8]. Conversely, prompts that omit necessary context force the model to make implicit assumptions, which increases the likelihood of responses that are technically fluent but misaligned with the user's actual intent [8], [45].

B. Explicit Instructions

Explicit instructions, including directives about the desired reasoning process, tone, length or level of certainty, have been shown to materially influence model output. Instructions that explicitly request step-by-step reasoning, for instance, have been associated with substantial gains in accuracy on reasoning-intensive tasks, effectively functioning as a lightweight, zero-shot substitute for full chain-of-thought demonstrations [27]. At the same time, instructions that are ambiguous, contradictory, or overly terse increase the likelihood of prompt sensitivity effects, in which superficially similar instructions elicit divergent outputs [7], [9].

C. Examples (Demonstrations)

As discussed in Section III, the inclusion of examples in a prompt, whether one or several, generally improves the model's ability to match the expected output format and reasoning style. However, the literature emphasizes that not all examples are equally beneficial: the number, diversity, order and relevance of examples all independently affect performance, and poorly curated demonstrations can introduce misleading patterns that the model then reproduces in its output [3], [7]. This suggests that prompt designers

should treat example selection as a deliberate design decision rather than an incidental addition.

D. Constraints

Explicit constraints, such as word limits, required output formats, prohibitions on speculation, or instructions to cite sources, act as guardrails that narrow the model's output space and can improve both relevance and reliability [5], [6]. Constraints that require the model to ground its answer in verifiable information, in particular, have been associated with reduced hallucination rates in the reviewed literature, although no single constraint has been found to eliminate factual errors entirely [5]. This reinforces the conclusion that reliable GenAI use depends on a combination of well-designed prompts and realistic expectations about residual error rates.

VI. PRACTICAL GUIDELINES FOR EFFECTIVE PROMPT DESIGN

Based on the synthesis of literature presented above, the following practical guidelines are proposed for designing more accurate and reliable prompts when interacting with GenAI systems.

1. State the task explicitly rather than relying on implicit context; ambiguous task descriptions increase the risk of misaligned or inconsistent output [2], [8].
2. Provide relevant examples for tasks that require a specific output format or reasoning pattern, and pay attention to the order and diversity of the examples chosen, since these factors independently affect performance [3], [7].
3. For tasks involving arithmetic, logic, or multi-step reasoning, explicitly request step-by-step reasoning, either through worked examples or through a direct instruction such as "explain your reasoning step by step" [4], [27].
4. Where factual accuracy is critical, include constraints that require the model to indicate uncertainty, avoid speculation, or reference verifiable information, and independently verify any output before relying on it [5], [6].
5. Avoid unnecessary variation in prompt wording and formatting across repeated uses of the same task, since even minor, meaning-preserving changes have been shown to produce substantial variation in output quality [7], [9].
6. Where possible, test a prompt on a small number of representative cases before relying on it at scale, given the documented sensitivity of GenAI systems to prompt phrasing [7], [42].

VII. CHALLENGES AND FUTURE SCOPE

Despite the growing body of research on prompt engineering, several challenges remain. First, the field currently lacks a fully unified theoretical understanding of why prompt sensitivity occurs; while empirical patterns are well documented, mechanistic explanations remain an active area of research [43]. Second, the effectiveness of a given prompting strategy is often model- and task-dependent, meaning that guidelines derived from one model or benchmark do not always transfer reliably to another [7], [8]. Third, hallucination mitigation through prompting alone remains only partially effective, and further work is needed to combine prompt-level interventions with retrieval-based and model-level safeguards in a way that is accessible to non-expert users [5], [6]. Future research could usefully focus on developing standardized, empirically validated prompting frameworks for specific domains such as education and healthcare, on building lightweight tools that help users detect when a prompt is likely to be under-specified or ambiguous, and on longitudinal studies examining how prompting practices evolve as users gain experience with GenAI systems [9], [10].

VIII. CONCLUSION

This paper has presented a literature-based examination of the impact of prompt design on the accuracy and reliability of Generative AI responses. The reviewed evidence consistently shows that prompt structure, including the presence and quality of examples, the explicitness of instructions, the amount of relevant context, and the use of constraints, plays a decisive role in determining the accuracy, relevance, consistency and reliability of GenAI output. Structured approaches such as few-shot and chain-of-thought prompting generally outperform simple zero-shot prompting on complex tasks, but their benefits are conditional on careful prompt construction, and all prompting strategies remain subject to sensitivity effects arising from seemingly minor changes in wording or formatting. Given the scale at which

GenAI tools are now used by students, educators, researchers and professionals, improving general awareness of effective prompt design practices, as summarized in the guidelines proposed in this paper, represents a practical and accessible means of improving the accuracy and dependability of AI-assisted work. Future research combining prompt-level interventions with retrieval-based and model-level safeguards is likely to further improve the reliability of Generative AI systems in both academic and practical settings.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science, MAEER'S MIT Arts, Commerce and Science College, Alandi (D), Pune, for the support and resources provided in the course of this research.

REFERENCES

[1] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different NLP tasks," *arXiv preprint arXiv:2407.12994*, 2024.

[2] "A systematic survey of prompt engineering in large language models: Techniques and applications," arXiv preprint arXiv:2402.07927, 2024.

[3] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020.

[4] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 24824–24837, 2022.

[5] A. Alansari and H. Luqman, "Large language models hallucination: A comprehensive survey," arXiv preprint arXiv:2510.06265, 2025.

[6] "Hallucination to truth: A review of fact-checking and factuality evaluation in large language models," arXiv preprint arXiv:2508.03860, 2025.

[7] S. Shevkari, P. S. Bhosale, and G. A. Mule, "Machine learning in cybersecurity: Cutting-edge approaches to threat detection and protection," International Journal of Research Publication and Reviews, vol. 6, no. 11, pp. 7090–7092, 2025.

[8] A. Akli, M. Cordy, M. Papadakis, and Y. Le Traon, "When prompt under-specification improves code correctness: An exploratory study of prompt wording and structure effects on LLM-based code generation," arXiv preprint arXiv:2604.24712, 2026.

[9] G. N. Taur, M. M. Vidhate, and S. S. Shevkari, "Deep learning meets biometrics: A comparative study of modern identification techniques," International Journal of Innovative Research in Technology, vol. 12, no. 2, pp. 545–552, 2025.

[10] Z. Contractor and G. Reyes, "Generative AI in higher education: Evidence from an elite college," arXiv preprint arXiv:2508.00717, 2026.

[11] S. Devikar, M. Hinge, and S. S. Shevkari, "Human vs. machine: A review of AI's impact on language learning interaction or replacing human interaction," International Journal for Multidisciplinary Research, vol. 7, no. 3, 2025, doi: 10.36948/ijfmr. 2025.v07i03.49459.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Technical Report, 2019.

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[14] S. Shevkari, M. M. Hinge, and S. C. Devikar, "DevOps and continuous integration: An investigation into the benefits and challenges," International Journal of Innovative Research in Technology, vol. 12, no. 2, pp. 3993–3998, 2025.

[15] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "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.

[16] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.