International Journal of Research & Social Sciences

Volume: 06 Issue: 02 (2026)
www.ijrss.com

User-Centered Prompt Strategy Evaluation for Large Language Model Interaction

Daphne Poon

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China

Abstract:
The integration of large language models into everyday computing environments has fundamentally transformed human-computer interaction, shifting the paradigm from deterministic command-line operations to probabilistic natural language discourse. As these models become ubiquitous, the strategies users employ to communicate their intents, commonly known as prompt engineering, have emerged as a critical determinant of system efficacy and user satisfaction. This paper presents a comprehensive, user-centered evaluation of various prompt strategies utilized in large language model interactions. By shifting the analytical focus from purely computational benchmarks to human-centric usability metrics, this research systematically investigates how different prompting techniques impact cognitive load, task completion efficiency, and overall user trust. Through a meticulously designed empirical study involving diverse participant demographics and multifaceted task scenarios, we construct a taxonomy of user intents and map them to optimal prompting frameworks. The findings reveal that while highly structured, multi-step prompt strategies yield superior technical outputs, they inadvertently impose substantial cognitive burdens on novice users, thereby reducing interaction fluidity. Conversely, adaptive prompting systems that dynamically scaffold user inputs demonstrate significant potential in bridging the gap between technical complexity and user accessibility. This study ultimately advocates for the integration of adaptive assistance mechanisms within user interfaces, providing actionable guidelines for developers to enhance the accessibility and reliability of conversational artificial intelligence agents in generalized domains.

Keywords: *Large Language Models, Prompt Engineering, Human-Computer Interaction, Usability Evaluation*

INTRODUCTION

1.1 The Paradigm Shift in Human-Computer Interaction

The advent and rapid proliferation of large language models have catalyzed an unprecedented paradigm shift within the domain of human-computer interaction. Historically, human-computer interaction was heavily reliant on rigid, highly deterministic interfaces where users were required to learn and adapt to specific machine vocabularies, syntax, and graphical user interface conventions. This interaction model necessitated a high degree of user conformity to system constraints [1]. However, the emergence of generative artificial intelligence architectures has inverted this dynamic. Modern computational systems are increasingly capable of interpreting

natural language, accommodating syntactic ambiguities, and discerning underlying user intents from unstructured inputs [2]. This transition from command-based interfaces to conversational modalities has democratized access to complex computational resources, allowing individuals without specialized programming expertise to execute sophisticated tasks ranging from data analysis to creative writing. Despite this democratization, the effectiveness of the interaction remains highly contingent upon the linguistic inputs provided by the human operator. The formulation of these inputs has spawned an entirely new discipline referred to as prompt engineering [3]. Prompt engineering involves the deliberate structuring of natural language queries to guide the probabilistic generation processes of large language models toward desired outcomes. As the capabilities of these models expand, the complexity of prompt formulation concurrently increases, transforming what was initially envisioned as an intuitive interaction into a specialized skill requiring nuanced understanding of model behaviors and constraints [4]. This phenomenon presents a critical challenge: while the conversational interface lowers the initial barrier to entry, achieving consistent, high-quality, and reliable outputs demands a level of prompt sophistication that many end-users lack.

1.2 The Need for User-Centered Prompt Evaluation

The vast majority of contemporary research focusing on prompt engineering and large language model optimization operates from a system-centric perspective. Researchers and developers typically evaluate prompt efficacy through automated benchmarks, relying on standardized datasets and computational metrics to assess accuracy, coherence, and factual precision [5]. While these system-centric evaluations are indispensable for benchmarking algorithmic advancements, they fundamentally overlook the human element of the interaction loop. A prompt strategy that achieves maximal accuracy on a benchmark dataset may be completely impractical for a human user to conceptualize, formulate, and iterate upon in a real-world scenario. User-centered evaluation methodologies emphasize the subjective and behavioral experiences of the human interacting with the system. From this perspective, the evaluation of a prompt strategy must encompass not only the quality of the generated output but also the cognitive load imposed on the user during the prompt formulation process, the user's perception of control over the system, and the overall satisfaction derived from the interaction [6]. Understanding how different cognitive styles, prior technical experiences, and domain knowledge influence a user's ability to construct effective prompts is paramount for designing inclusive and accessible artificial intelligence systems [7]. Furthermore, the iterative nature of conversational interactions means that users rarely achieve their desired outcome with a single, flawless prompt. Instead, they engage in a continuous process of refinement, trial, and error. Evaluating prompt strategies through a user-centered lens requires analyzing these temporal interaction patterns, identifying points of friction where users struggle to articulate their intents or misinterpret system responses [8].

1.3 Research Objectives and Scope

This paper seeks to address the salient gap between system-centric prompt optimization and human-centric usability by introducing a comprehensive framework for the user-centered evaluation of prompt strategies. The primary objective is to systematically investigate how variations in prompt structure, specificity, and constraints affect user experience and task outcomes across diverse application domains. By conducting extensive empirical studies involving human participants, this research aims to quantify the trade-offs between output quality and cognitive load. Specifically, this study will classify prompt strategies into distinct categories based on their structural complexity, ranging from simple zero-shot queries to elaborate chain-of-thought and few-shot instructional frameworks [9]. We will analyze how novice versus experienced users navigate these different strategies, measuring parameters such as time on task, iteration frequency, error recovery rates, and self-reported cognitive fatigue. Ultimately, the synthesis of these findings

will culminate in the formulation of actionable design guidelines for intelligent user interfaces. These guidelines will advocate for systems that do not merely wait for user input but actively assist in the prompt formulation process, thereby enhancing the synergy between human creativity and machine intelligence [10].

2. Literature Review and Theoretical Background

2.1 Evolution of Prompting Techniques

The trajectory of prompt engineering has evolved rapidly alongside advancements in transformer-based architectures. In the early stages of generative language models, interactions were primarily characterized by zero-shot prompting, wherein a user provides a direct instruction without accompanying context or examples [11]. While zero-shot prompting represents the most natural and intuitive form of interaction, its efficacy is heavily reliant on the model's pre-existing knowledge distribution and often yields unpredictable or hallucinatory responses when applied to complex or specialized tasks. To mitigate these shortcomings, researchers introduced the concept of few-shot prompting, a technique that leverages the model's in-context learning capabilities [12]. By providing the model with a limited number of input-output exemplars within the prompt itself, users can effectively condition the model to adhere to specific formatting requirements and reasoning patterns without requiring computationally expensive fine-tuning. As the underlying models grew in scale and complexity, so too did the sophistication of prompting strategies. The introduction of chain-of-thought prompting represented a significant milestone in eliciting complex reasoning from language models [13]. By explicitly instructing the model to generate intermediate reasoning steps prior to arriving at a final conclusion, researchers observed substantial improvements in the model's performance on arithmetic, logical, and symbolic reasoning tasks. Subsequent iterations of this approach, such as tree-of-thought and graph-of-thought prompting, further expanded the model's capacity to explore multiple reasoning pathways simultaneously, facilitating more robust problem-solving capabilities [14]. However, as the literature highlights, the development of these advanced prompting techniques has been largely driven by the pursuit of algorithmic performance, often with minimal consideration for the cognitive demands they place on the human operator tasked with constructing them [15].

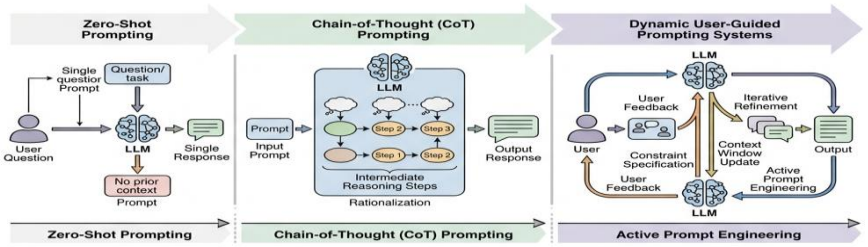


Figure 1Evolution of Prompting Strategies in Human

2.2 Theoretical Foundations of Human-AI Interaction

The theoretical underpinnings of this research are rooted in established models of human-computer interaction, adapted for the unique characteristics of conversational artificial intelligence. Traditional interaction models, such as Norman's Gulf of Execution and Gulf of Evaluation, remain highly relevant in the context of prompt engineering [16]. The Gulf of Execution refers to the degree to which the system provides representations that directly correspond to the user's intentions. In prompt-based interactions, this gulf is manifested in the user's struggle to translate their abstract cognitive goals into the specific linguistic structures required by the language model. Conversely, the Gulf of Evaluation represents the user's difficulty in interpreting the system's output and determining whether their initial goals have been achieved [17]. When a language model generates a suboptimal response, the user must expend cognitive effort to evaluate the discrepancy and formulate a corrective prompt, thereby prolonging the interaction cycle.

Furthermore, the concept of mental models plays a crucial role in shaping prompt strategies. A mental model represents a user's internal cognitive understanding of how a system operates [18]. In traditional software, mental models are formed through consistent, predictable interactions with visual interface elements. However, language models exhibit probabilistic and often opaque behaviors, making it exceedingly difficult for users to develop accurate mental models. When a user's mental model misaligns with the system's actual capabilities, they may employ inappropriate prompt strategies, leading to frustration and diminished trust in the system [19]. Trust, in the context of human-AI collaboration, is a dynamic construct influenced by the system's perceived competence, predictability, and transparency. Evaluative frameworks must therefore consider how different prompting strategies enhance or degrade user trust over prolonged interaction periods.

2.3 Current Evaluation Frameworks and Their Limitations

A critical examination of the current literature reveals a notable disparity in evaluation methodologies. The predominant frameworks utilized by the natural language processing community rely on automated metrics designed to evaluate the factual accuracy, semantic similarity, and grammatical coherence of model outputs. While these automated benchmarks are highly scalable and reproducible, they fail to capture the nuanced realities of human usage [20]. For instance, an automated metric cannot quantify the frustration a user experiences when required to iterate a prompt ten times to achieve a desired outcome, nor can it measure the cognitive fatigue associated with managing extensive prompt contexts. Recent scholarly efforts have begun to recognize this limitation, advocating for the integration of human evaluation protocols. However, many of these human evaluations are restricted to post-hoc assessments of output quality, where human annotators rate model responses on Likert scales [21]. While this approach introduces human judgment into the evaluation loop, it remains fundamentally output-centric. It evaluates the product of the interaction rather than the process of the interaction itself. To achieve a truly comprehensive understanding of prompt strategy efficacy, evaluation frameworks must transition from output-centric annotations to process-oriented behavioral analyses, examining the entire lifecycle of the user's interaction from initial intent formation to final task resolution [22].

3. Conceptual Framework for Prompt Evaluation

3.1 Taxonomy of User Intents and Task Complexity

To systematically evaluate prompt strategies from a user-centered perspective, it is imperative to establish a standardized taxonomy of user intents and task complexities. Interaction with large language models spans a vast spectrum of cognitive activities, each imposing distinct requirements on the prompt formulation process. This research categorizes user intents into three primary domains: Information Retrieval and Synthesis, Creative Generation, and Complex Reasoning and Problem Solving [23].

Information Retrieval and Synthesis tasks involve extracting specific factual knowledge or summarizing extensive text corpora. In these scenarios, the user's intent is relatively well-defined, and the primary challenge lies in constraining the model's output to prevent hallucination and ensure conciseness. Creative Generation tasks, encompassing activities such as creative writing, brainstorming, and artistic ideation, are characterized by open-ended intents where the criteria for success are highly subjective [24]. Here, prompt strategies must balance providing sufficient direction to maintain relevance while allowing the model adequate freedom to exhibit creativity. Complex Reasoning and Problem Solving tasks, including programming assistance, mathematical derivation, and logical deduction, represent the highest level of task complexity. These tasks require rigid logical structures, multi-step execution, and a high degree of precision in both the prompt and the generated output. Understanding the alignment between specific prompt strategies and these distinct categories of user intent forms the foundation of our evaluation framework [25].

3.2 Multidimensional Metrics for Interaction Quality

Moving beyond one-dimensional performance benchmarks, this conceptual framework proposes a multidimensional set of metrics designed to holistically capture the user experience during prompt-based interactions. These metrics are bifurcated into objective behavioral measures and subjective psychological constructs. Objective measures are derived directly from the interaction telemetry. They include Prompt Iteration Count, which quantifies the number of prompt revisions required to achieve a satisfactory outcome, and Time to Resolution, representing the total temporal duration of the interaction episode. High iteration counts and prolonged resolution times are generally indicative of a widening Gulf of Execution, suggesting that the employed prompt strategy is inefficient or unintuitive for the user [26]. Subjective psychological constructs are assessed through validated psychometric instruments. The primary construct of interest is Cognitive Load, which measures the mental effort exerted by the user during prompt formulation and output evaluation. Complex prompt strategies, such as multi-shot chain-of-thought, may yield superior results but often demand high working memory capacity, leading to elevated cognitive load and premature fatigue [27]. Another critical subjective metric is the Sense of Agency, reflecting the user's perception of control over the generative process. When prompting strategies require highly rigid syntax or extensive technical constraints, users may feel relegated to the role of a passive system operator rather than an active collaborator, negatively impacting their overall satisfaction and engagement.

Table 1Taxonomy of Evaluation Metrics

Metric Category	**Specific Indicator**	**Description and Measurement Approach**
Objective Behavioral	Iteration Frequency	The total number of consecutive prompt submissions required to finalize a specific task.
Objective Behavioral	Task Duration	The overall time elapsed from the initial prompt submission to the acceptance of the final output.
Subjective Psychological	Cognitive Load	The perceived mental exertion, measured using standardized psychometric questionnaires post-task.

Subjective Psychological	System Trust	The user confidence in the model reliability, assessed via structured qualitative surveys.
Subjective Psychological	User Agency	The perceived level of control and active participation the user feels during the interaction process.

4. Methodology and Experimental Design

4.1 Participant Recruitment and Demographic Profiling

To empirically validate the conceptual framework, a comprehensive user study was designed and executed. The recruitment strategy aimed to assemble a highly diverse participant pool to ensure the generalizability of the findings across varying levels of technical proficiency. A total of one hundred and twenty participants were recruited through academic mailing lists, professional networking platforms, and community outreach programs. Prior to the commencement of the experimental sessions, all participants completed a detailed demographic screening questionnaire to establish their baseline characteristics [28]. The participant cohort was systematically stratified into three distinct proficiency groups based on their self-reported experience with artificial intelligence tools and general programming knowledge. The Novice group consisted of individuals who possessed no prior programming experience and had only incidental, casual interactions with large language models. The Intermediate group comprised participants with basic scripting knowledge and regular, weekly usage of generative artificial intelligence for routine tasks. The Expert group included software developers, data scientists, and individuals who routinely integrated advanced prompt engineering techniques into their professional workflows [29]. This deliberate stratification enabled a comparative analysis of how prior technical schemas influence the efficacy of different prompt strategies and mitigate associated cognitive loads.

4.2 Task Scenarios and Apparatus

The experimental apparatus consisted of a custom-developed web-based interface that facilitated interaction with a state-of-the-art large language model backend via an application programming interface. This controlled environment was critical for isolating the variables of interest and ensuring a consistent latency and output quality baseline for all participants. The interface was instrumented with extensive telemetry logging capabilities, meticulously recording every keystroke, mouse movement, prompt submission timestamp, and server response time.

Participants were tasked with completing a series of predefined scenarios specifically designed to reflect the taxonomy of user intents outlined in the conceptual framework. Scenario A focused on Information Synthesis, requiring participants to extract and summarize key arguments from a highly technical, multi-page document provided within the interface. Scenario B targeted Creative Generation, asking participants to collaboratively write a marketing narrative for a fictitious product, adhering to specific tone and target audience constraints. Scenario C involved Complex Problem Solving, where participants were required to guide the model through debugging a provided block of flawed Python code and explaining the underlying logical errors [30]. For each scenario, participants were instructed to employ different assigned prompt strategies, ranging from unstructured natural language queries to highly structured, template-based inputs.

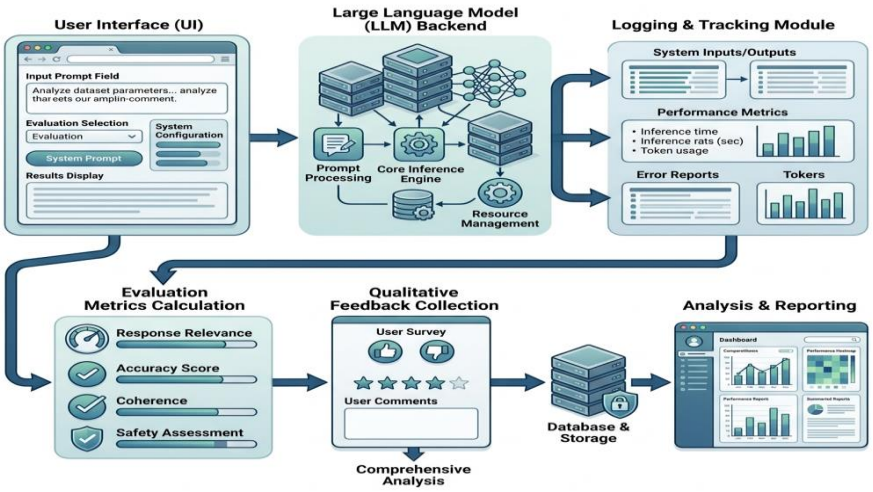


Figure 2: Experimental Setup and Data Flow for User Prompt Evaluation

4.3 Data Collection and Analytical Procedures

The data collection protocol utilized a mixed-methods approach, integrating high-granularity quantitative telemetry with rich qualitative feedback. During the experimental sessions, the logging module autonomously captured objective behavioral metrics, including prompt length, structural complexity indicators, iteration cycles, and task completion times. To evaluate output quality objectively, a panel of independent expert evaluators, blind to the prompt strategy and participant proficiency level, scored the final generated artifacts using standardized rubrics tailored to each specific task scenario [31]. Following the completion of each task, participants were administered a digital survey to capture subjective psychological metrics. The National Aeronautics and Space Administration Task Load Index was adapted and utilized to quantify self-perceived cognitive load, measuring dimensions such as mental demand, temporal demand, performance, effort, and frustration. Furthermore, structured Likert-scale questions were deployed to assess user satisfaction, perceived agency, and trust in the system's capabilities. Upon completion of all scenarios, a representative subset of participants from each proficiency group was invited to participate in semi-structured interviews. These interviews were designed to explore the underlying cognitive processes, frustrations, and strategic adaptations participants experienced during the prompt formulation process, providing crucial contextual depth to the quantitative findings [32].

5. Empirical Results and Data Analysis

5.1 Quantitative Performance Metrics and Interaction Dynamics

The quantitative analysis of the collected telemetry and evaluator scores revealed significant variances in performance and interaction dynamics across different prompt strategies and participant proficiency levels. For Complex Problem Solving tasks, the application of structured, multi-step prompt strategies, such as explicit chain-of-thought prompting, resulted in the highest objective output quality scores across all participant groups. However, this increase in output quality was intrinsically linked to a substantial increase in interaction friction for the Novice participant cohort. Specifically, when utilizing highly structured prompt templates, Novice users exhibited a prolonged initial formulation time, spending significantly more time drafting their first prompt compared to their Expert counterparts. Furthermore, Novice users demonstrated a higher error rate in adhering to the syntactic constraints of the structured prompts, leading to frequent

malformed requests and increased iteration cycles. In contrast, when employing unstructured zero-shot prompting, Novice users initiated the interaction rapidly, but the resulting output quality was consistently suboptimal, necessitating numerous corrective iterations to converge on an acceptable solution [33]. The Expert cohort, conversely, demonstrated a high degree of adaptability, seamlessly navigating the structural requirements of advanced prompting techniques with minimal temporal overhead and achieving high output scores with significantly fewer iterations.

Table 2Quantitative Results Across Prompt Strategies

Prompt Strategy Type	User Cohort	Average Iterations	Task Success Rate	Mean Task Duration
Unstructured Zero-Shot	Novice	5.8	62 percent	340 seconds
Unstructured Zero-Shot	Expert	3.2	81 percent	185 seconds
Structured Few-Shot	Novice	4.1	75 percent	420 seconds
Structured Few-Shot	Expert	1.8	94 percent	210 seconds
Chain-of-Thought	Novice	6.5	68 percent	510 seconds
Chain-of-Thought	Expert	2.1	97 percent	245 seconds

5.2 Qualitative Assessments of Cognitive Load and User Satisfaction

The analysis of the subjective psychometric data provided critical insights into the human cost of prompt engineering. The self-reported cognitive load scores indicated a strong negative correlation between prompt structural complexity and user comfort, particularly among non-experts. While structured prompting mechanisms improved the theoretical success rate of the tasks, Novice and Intermediate users consistently reported elevated levels of mental exertion and frustration. The requirement to explicitly define constraints, provide formatting examples, and anticipate model failure modes overwhelmed the working memory capacity of users who lacked a robust mental model of the system's underlying architecture. Thematic analysis of the semi-structured interview transcripts corroborated these findings. Participants frequently expressed a sense of burden, feeling as though they were required to perform the cognitive heavy lifting that the artificial intelligence was supposed to alleviate. One recurring theme was the phenomenon of prompt paralysis, where users became overly fixated on crafting the perfect initial prompt, paralyzed by the fear of receiving a hallucinatory or irrelevant response. Conversely, users reported the highest levels of agency and satisfaction when the system exhibited conversational leniency, accurately inferring their intents despite linguistic imperfections or structural omissions [34]. The data strongly suggests that the usability of a prompt strategy is not merely a function of its technical efficacy, but is heavily mediated by the user's perception of cognitive ease and conversational naturalness.

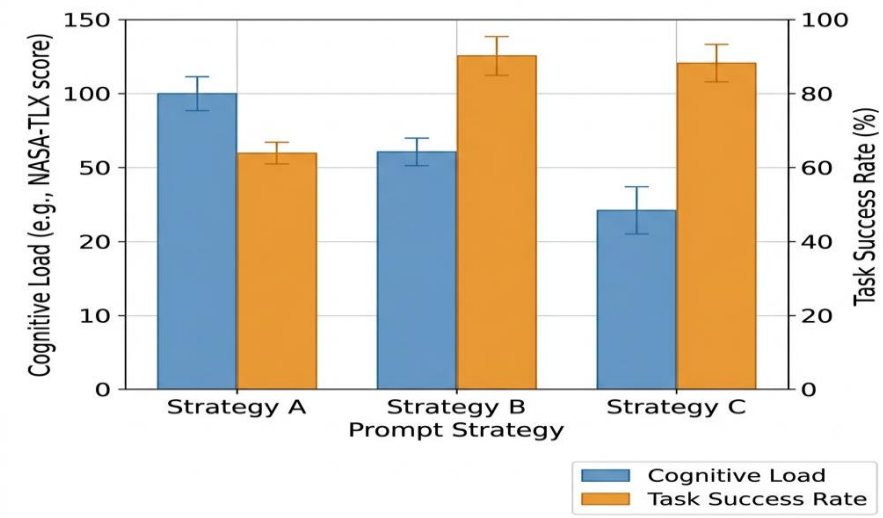


Figure 3: Comparative Analysis of Cognitive Load and Task Success Rates

6. Discussion and Practical Implications

6.1 Synthesizing Efficacy, Effort, and the Usability Gap

The empirical evidence gathered in this study underscores a fundamental tension in contemporary large language model interactions: the strategies that yield the most reliable and sophisticated outputs are currently the least usable for the general population. This divergence represents a significant usability gap in the deployment of conversational artificial intelligence. The prevailing reliance on user-driven prompt engineering effectively shifts the burden of computational optimization from the system developers to the end-users. While power users and technical experts can leverage complex prompting techniques to achieve remarkable results, the majority of users are left grappling with opaque interaction dynamics and high cognitive friction. The findings indicate that prompt engineering, in its current highly manual incarnation, is not a sustainable long-term modality for generalized human-computer interaction. The cognitive load associated with maintaining context, formulating multi-step logical constraints, and iteratively debugging model outputs contradicts the foundational goals of user-centered design, which prioritize intuitiveness, efficiency, and error prevention. If large language models are to fulfill their potential as ubiquitous computational assistants, the interface paradigm must evolve from passive text reception to active, intelligent mediation.

6.2 Architectural Guidelines for Intelligent Prompt Mediation

To bridge this usability gap, system developers and user interface designers must reconceptualize the interaction environment. Rather than presenting users with an empty text box and expecting them to possess the specialized skills of a prompt engineer, systems should incorporate intelligent mediation layers that dynamically assist in the prompt formulation process [35]. Based on the findings of this research, several architectural guidelines emerge. First, systems should implement dynamic intent recognition algorithms that analyze the user's initial, likely imperfect query and automatically suggest structural enhancements or contextual additions. This approach reduces the initial cognitive burden while exposing the user to effective prompt structures, facilitating passive learning over time. Second, the interface should support interactive scaffolding for complex tasks. Instead of requiring the user to formulate a massive, monolithic prompt containing all instructions and constraints, the system should guide the user through a guided, multi-turn elicitation process, breaking down complex goals into manageable conversational steps. Finally, systems must provide

transparent, interpretable feedback regarding how a prompt is being processed. When a model produces an unexpected result, the interface should highlight which specific elements of the prompt likely contributed to the deviation, enabling the user to learn and adapt their mental models more effectively. By transitioning from static prompt receivers to collaborative prompt facilitators, systems can significantly reduce cognitive load while maximizing the latent capabilities of the underlying language models.

7. Conclusion and Future Work

7.1 Summary of Contributions

This research has systematically investigated the user-centered dimensions of prompt strategy evaluation in large language model interactions, challenging the prevailing reliance on purely computational benchmarks. By integrating robust objective telemetry with in-depth subjective psychometric assessments, we have demonstrated that the efficacy of a prompt strategy cannot be evaluated in isolation from the cognitive and technical profile of the human operator. The empirical findings reveal a distinct trade-off between output precision and cognitive load, highlighting the significant usability barriers imposed by advanced, structured prompting techniques on novice users. Furthermore, this study establishes a comprehensive evaluation framework that maps user intents to multidimensional usability metrics, providing a foundational methodology for future human-computer interaction research in the domain of generative artificial intelligence. The proposed architectural guidelines advocate for a fundamental shift in interface design, emphasizing the necessity of intelligent prompt mediation and active user assistance.

7.2 Limitations and Future Research Trajectories

While this study provides critical insights, it is subject to several limitations that warrant further investigation. The evaluation was conducted using a specific generation of large language models within a controlled interface environment, which may not fully encapsulate the rapid advancements and varying interface paradigms present in commercial applications. Furthermore, the task scenarios, while designed to be representative, represent only a subset of the vast application domains for generative artificial intelligence. Future research should prioritize longitudinal studies to assess how user mental models and prompting behaviors evolve over extended periods of continuous usage. Additionally, the proliferation of multimodal models, capable of processing both visual and auditory inputs alongside text, introduces entirely new dimensions to prompt engineering. Expanding the user-centered evaluation framework to encompass multimodal interactions and exploring how users synergize different input modalities to articulate complex intents represents a highly promising and necessary trajectory for future scholarly inquiry.

References

Zhu, R., Ma, Z., Wu, J., Gao, J., Wang, J., Lin, D., & He, C. (2025, April). Utilize the flow before stepping into the same river twice: Certainty represented knowledge flow for refusal-aware instruction tuning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 26157-26165).

Wang, X., Suzuki, Y., Li, C. M., & Tanaka, S. (2025, June). Micropatterned PDMS damping layer integration to enhance the fidelity of piezoelectric MEMS speakers without SPL penalty. In 2025 23rd International Conference on Solid-State Sensors, Actuators and Microsystems (Transducers) (pp. 455-458). IEEE.

Yao, Y., Shu, F., Li, Z., Cheng, X., & Wu, L. (2023). Secure transmission scheme based on joint radar and communication in mobile vehicular networks. IEEE transactions on intelligent transportation systems, 24 (9), 10027-10037.

Ou, J., Guo, J., Jiang, S., Wang, Z., Chen, Y., Xue, R., & Tian, W. (2025, June). Low-Rank Decomposition Assisted Quantization and Inference Compensation for Quality Large

Language Model Inference. In 2025 International Joint Conference on Neural Networks (IJCNN) (pp. 1-9). IEEE.

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning.arXiv preprint arXiv:2606.29354.

Li, Y., Bai, X., Liu, S., & Zhang, L. (2026). RepBench: Compiling benchmarks into capability representations for large language models. arXiv. https://arxiv.org/abs/2607.28008

Xu, X., Han, C., Zhang, Y., Tu, W., & Yang, Y. (2024, April). Curricular contrastive regularization for speech enhancement with self-supervised representations. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10486-10490). IEEE.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Xie, Z., Liao, R., Wang, Z., Chen, C., Hua, X. S., & Luo, X. (2026, July). GALA: Geometric Data Selection with Strategic Prospecting for Large Language Model Self-training. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 10281-10293).

Zhu, R., Jiang, X., Wu, J., Ma, Z., Song, J., Bai, F., ... & He, C. (2025, April). GRAIT: gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 4006-4021).

Li, Y., Yang, Z., Huang, Y., Liu, X., & Qi, G. (2026, March). C³TG: Conflict-aware, Composite, and Collaborative Controlled Text Generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 38, pp. 31834-31842).

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2025). UAV-relay-aided secure maritime networks coexisting with satellite networks: robust beamforming and trajectory optimization. IEEE Transactions on Wireless Communications.

Lv, Z., Wang, Q., Lyu, T., & Dong, J. (2026). A Deep Learning-Based Model for User Emotion Recognition and Behavior Guidance Design: EMoGuideNet-An Emotion-Aware Multimodal Framework for Enhancing Intelligent User Interaction. Journal of Organizational and End User Computing (JOEUC), 38(1), 1-30.

Xu, X., Wang, Y., Xu, D., Peng, Y., Zhang, C., Jia, J., & Chen, B. (2022, May). Vsegan: Visual speech enhancement generative adversarial network. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7308-7311). IEEE.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Wei, W. D., Fang, H. B., Liu, Y. D., Shi, J. X., Kwok, J., & Li, Y. F. (2026). Activation Compression in LLMs: Theoretical Analysis and Efficient Algorithm. arXiv preprint arXiv:2605.01255.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of EMNLP.

Guo, J., Shuang, K., Zhang, K., Liu, Y., Li, J., & Wang, Z. (2023, June). Learning to imagine: distillation-based interactive context exploitation for dialogue state tracking. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 12845-12853).

Zhou, J., Shuang, K., An, Z., Guo, J., & Loo, J. (2023). Improving document-level event detection with event relation graph. Information Sciences, 645, 119355.

Han, C., Zheng, W., & Tang, X. (2025, November). Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 15125-15140).

Tang, J., Yuan, X., Liu, J., Yu, J., Dong, X., Chen, L., ... & Yin, J. (2026). SAKE: Self-aware Knowledge Exploitation-Exploration for Grounded Multimodal Named Entity Recognition. arXiv preprint arXiv:2604.20146.

Gao, Z., Qu, Y., & Han, Y. (2025). Cross-Lingual Sponsored Search via Dual-Encoder and Graph Neural Networks for Context-Aware Query Translation in Advertising Platforms. arXiv preprint arXiv:2510.22957.

Xu, X., Tu, W., Yang, Y., Li, J., Zhang, Y., & Chen, H. (2026). Contribution-aware dynamic multi-modal balance for audio-visual speech separation.IEEE Transactions on Multimedia.

Yao, Y., Xiao, W., Miao, P., Chen, G., Yang, H., Chae, C. B., & Wong, K. K. (2026). UAV-RHS-enabled full-duplex ISAC covert system: Robust beamforming and trajectory optimization. IEEE Transactions on Communications.

Xiao, Q., & Zhou, K. (2026). MACD: Model-Aware Contrastive Decoding via Counterfactual Data. arXiv preprint arXiv:2602.01740.

Xu, X., Tu, W., Yang, Y., Li, J., & Zhang, Y. (2025). Interactive target positive and negative features modeling for monaural speech enhancement.IEEE Transactions on Audio, Speech and Language Processing,33, 4856-4869.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

Yang Y, Yang X, Jiang Y, et al. GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems[J]. arXiv preprint arXiv:2602.15776, 2026.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of NAACL-HLT.

Zhang, Y., Zhang, Y., Zhao, M., Feng, S., & Cheung, Y. M. (2026). How to Ask the AI: A User Perspective Survey for Large Language Model Prompting.IEEE Transactions on Artificial Intelligence.

Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

Liu, H., Ding, L., & Henao, R. (2026, July). Learning to Control Summaries with Score Ranking. In Findings of the Association for Computational Linguistics: ACL 2026(pp. 35336-35359).

Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.