

PROMPT OPTIMIZATION ENGINE: Overcoming Hallucinations in LLMs through Prompt Engineering

Sarfaraz Ahmed

Department of Computer Systems
Engineering
Mehran University of Engineering
and Technology
Jamshoro, Pakistan
sarfarazahmedbehan2@gmail.com

Abdul Qadir

Department of Computer Systems
Engineering
Mehran University of Engineering
and Technology
Jamshoro, Pakistan
abdulqadirrmagssii@gmail.com

Muhammad Areeb

Department of Computer
Systems Engineering
Mehran University of
Engineering and Technology
Jamshoro, Pakistan
syedareebali795@gmail.com

Huzaiifa Saeed

Department of Computer Systems
Engineering
Mehran University of Engineering
and Technology
Jamshoro, Pakistan
huzai.exe@gmail.com

Sanam Narejo

Department of Computer Systems
Engineering
Mehran University of Engineering
and Technology
Jamshoro, Pakistan
Sanam.narejo@faculty.muuet.edu.pk

Bhawani Shankar Chowdhry

Executive Head, National Centre
for Robotics, Automation and AI
(NCRAAI), Mehran University
of Engineering & Technology,
Jamshoro, Pakistan
bhawani.chowdhry@faculty.muuet.edu.pk

ABSTRACT

In an era dominated by Generative AI, the significance of efficiently interacting with Large Language Models (LLMs) cannot be overstated. The widespread adoption of LLMs in various industries underscores the importance of obtaining accurate and rational responses while avoiding the pitfalls of hallucination. This research endeavors to tackle these challenges head-on, aiming to enhance the efficacy of LLMs interactions through our innovative Prompt Optimization Engine (POE). By meticulously crafting prompts tailored to user needs and preferences, POE seeks to revolutionize the way users engage with LLMs, facilitating more coherent and reasonable responses. Leveraging a diverse array of LLMs, including Gemma, Llama, Mistral, and Phi, our study culminates in the selection of Generative Pretrained Transformer-4 Omni due to its superior efficiency for text generation. Through the development, evaluation of foundation models, integration using LangChain framework, and the creation of a user-friendly web application, POE promises to empower users and elevate the capabilities of LLMs interactions. This research not only addresses the immediate need for improved LLMs interactions but also paves the way for further advancements in the field of Generative AI.

KEYWORDS: large language models, generative AI, hallucinations, prompt engineering

1. INTRODUCTION

The development of advanced AI language models like ChatGPT and Gemini has revolutionized our interaction with technology,

providing quick and highly relevant information that is immensely useful in many areas of life. However, despite their impressive capabilities, these models are not without flaws. One significant issue they face is "hallucinations," [1, 2] where the LLMs

generate information that is incorrect or fabricated. This can be particularly problematic in high-stakes domains such as healthcare or finance, where accuracy is paramount.

Hallucinations can occur for two primary reasons: the AI model may be trained on incorrect or insufficient data [3], or the prompts provided by users may be unclear or inefficient [4]. Our research focuses on the latter—optimizing user prompts to reduce the occurrence of hallucinations. Studies have demonstrated that domain-specific prompts can significantly reduce the incidence of hallucinations, highlighting the critical role of prompt clarity and precision.

Prompt engineering is essential in optimizing the outputs of LLMs [4, 5, 6, 7, 8, 9]. Effective prompt engineering involves crafting clear and specific prompts that guide the LLMs to produce accurate and relevant responses. Y [4] Techniques such as predefined prompt templates, fine-tuning, and contextual adjustments have been explored to enhance LLM’s performance. Advanced methods like Chain of Thought (CoT) [5, 7, 10,] and Few-Shot Prompting [5, 7, 9, 10, 11] have shown promise in improving the accuracy and comprehensibility of AI responses. What was once considered an artistic skill in crafting prompts has now become a strategic necessity to unlock the full potential of these language models.

To address the issue of hallucinations, The paper proposes a system designed to improve the prompts given to LLMs, ensuring that the responses generated closely match user intentions. Our approach involves asking users four key questions to tailor the LLM’s behavior, focusing on the user’s role, the context or reason for the prompt, the desired tone of the recommended prompt, and the response length (or word limit). This user-centric method aims to create prompts that align perfectly with user expectations, thereby reducing the risk of hallucinations and enhancing the overall user experience.

The field of natural language processing is continuously evolving, with new advancements

being made regularly. This underscores the importance of well-crafted prompts in maximizing the performance of LLMs [4]. Techniques like Chain of Thought (CoT) [5, 7, 10,] and Few-Shot Prompting [5, 7, 9, 10, 11] have proven effective in making LLM responses more accurate and understandable. This shift from an artistic to a strategic approach in prompt crafting is crucial for harnessing the full potential of language models.

Our paper is motivated by the need to address the challenges posed by hallucinations resulting from poorly constructed prompts. We aim to empower users by allowing them to specify important parameters such as their role, the tone of the response, the length of the response, and the specific context of their query. By prioritizing these parameters, we intend to create prompts that go beyond simple personalization and are truly context-aware. The ultimate goal of our research is not just to optimize performance but to transform how users interact with language models, effectively addressing the complex challenges associated with hallucinations.

2. RELATED WORK

Large Language Models (LLMs) like GPTs have significantly advanced natural language processing but face the critical issue of hallucinations [1, 2, 3, 4, 9, 12], where models generate factually incorrect or nonsensical information.

2.1 THEMATIC ANALYSIS

2.1.1. Types and Causes of Hallucination

Hallucinations in Large Language Models (LLMs) manifest as factual inaccuracies and logical inconsistencies, occurring due to incorrect/insufficient training data [3] or unclear/inefficient user prompts [4].

2.1.2. Existing Methods for Prompt Engineering:

We have reviewed various papers discussing techniques for overcoming hallucinations in Large Language Models (LLMs). Our primary focus has been on advanced prompt engineering techniques, including Chain of Thought (CoT), Tree of Thought (ToT), Tools, Connectors, and Skills, Automatic Multi-step Reasoning and Tool-use (ART), Enhancing Reliability through Self-Consistency,

Reflection, Expert Prompting, Streamlining Complex Tasks with Chains, Guiding LLM Outputs with Rails, and Streamlining Prompt Design with Automatic Prompt Engineering [5].

2.1.3. Critical Evaluation:

Talking about hallucination, there are several types of hallucination which are discussed in Table 1.

Table 1: Types of Hallucinations in LLMs

Models	Prompts	Type of Hallucination	Result
GPT 3.5 Turbo	A 6 th man is standing on top of a 100 th building. He is 5ft11, and is about to parachute off, but doesn't know if he'll have time to deploy the parachute, as the building is only 100 th high. So, when, during this 200 th jump should the woman deploy her parachute	Contradictory	This is a classic puzzle with a bit of a trick. The answer is that the woman should deploy her parachute immediately. The reason is that the man is described as a 6ft tall person standing on top of a 100ft building, but then it is mentioned that he is 5 ft11. Since the building is only 100 feet high and the man is slightly shorter than the building, he is already on the ground. Therefore, there is no jump, and the woman should deploy her parachute right away.
GPT 4o	Compute this $\left[\int_C \frac{\cos(z)}{z(z^2 + 8)} dz \right]$ a) $\pi i/2$ b) $\pi i/4$ c) πi d) $2\pi i$ e) $\pi i/8$	Complex	<ul style="list-style-type: none">$2\pi i$ <p>This option, $2\pi i$, approximates the magnitude of the computed result (even though there's a discrepancy in the sign). Thus, if you are selecting from the provided choices and considering only magnitude, $2\pi i$ would be the selection.</p>

Microsoft Copilot	if we descent 500m from the height of Mount Everest does it affect the status of its tallest mountain, give reason up to 100 words	Comparison	Certainly! Even if we descend 500 meters from Mount Everest's height, it would still remain the tallest mountain. The reason is that Everest's official height is 8,848.86 meters (29,032 feet) above sea level, as jointly announced by Nepal and China. In contrast, the second-highest mountain, K2, stands at 8,611 meters (28,251 feet). While geological processes can impact Everest's elevation, its status as the tallest remains unchanged
-------------------	--	------------	--

In addition, the causes of them arise from inadequately written prompts. can result in hallucinations, which need to be addressed by using of chain of thought and defining four parameters that will be covered in more detail later. Now, let's examine the outcomes of those parameters in Table 2 and compare them to Table 1 results.

Table 2: Outcomes of POE

Recommended Prompt	Parameters (User Role, Reason, Word Limit Tone)	Result
How can a Mathematics student calculate the precise moment for a 6ft individual standing a top of a 100ft building to deploy their parachute during a 200ft descent to ensure a safe landing? Provide a detailed explanation within 200 words.	<ul style="list-style-type: none"> • I am a mathematics student • I'm doing preparation for my Math exam • 200 (word limit) • Intelligent 	<p>To calculate the optimal timing for a parachutist to deploy her parachute during a jump from a 200ft height, several factors must be taken into consideration.</p> <p>First, it is important to consider the terminal velocity of the parachutist, which is the maximum speed at which she will fall. This can be calculated using the equation $V = \sqrt{(2 * g * h) / C}$, where V is the terminal velocity, g is the acceleration due to gravity (9.8 m/s^2), h is the height of the jump (200ft = 60.96m). and C is the drag coefficient.</p> <p>Next, the optimal deployment height of the parachute should be calculated to allow for a safe landing. This can be determined by estimating the time it takes for the parachute to fully open and slow down the parachutist to a safe landing speed, typically around 5 m/s. This timing should account for any potential delays in parachute deployment and should ensure that the parachutist reaches a safe landing speed before hitting the ground.</p> <p>By considering these factors and calculating the optimal timing for parachute deployment, a Mathematics student can help ensure a safe landing for the parachutist during her jump from a 200ft height.</p>

<p>As a student preparing for your calculus exam, can you analytically evaluate the integral $\int_C \frac{\cos(z)}{z(z^2 + 8)} dz$ along a positively oriented, simple closed contour enclosing the singularities of the integrand, and determine the correct value from the following options: a) $\frac{\pi i}{2}$, b) $\frac{\pi i}{4}$, c) πi, d) $2\pi i$, e) $\frac{\pi i}{8}$, in a response of 100 words?</p>	<ul style="list-style-type: none"> • I'm a data science student • I'm doing prep for my calculus exam • 100 (word limit) • Intelligent 	<p>To evaluate the given integral, we first need to find the residues of the integrand at its singularities. The functions $\frac{\cos(z)}{z}$ and $\frac{\cos(z)}{z^2 + 8}$ have simple poles at $z = 0$ and $z = \pm 2i\sqrt{2}$, respectively. By using the residue theorem, we find that the residues at these points are $\frac{1}{8}$, $-\frac{i\sqrt{2}}{16}$, and $\frac{i\sqrt{2}}{16}$, respectively. Since the contour encloses all singularities, the integral evaluates to $(2\pi i \frac{1}{8} - \frac{i\sqrt{2}}{16} + \frac{i\sqrt{2}}{16}) = \frac{\pi i}{4}$. Therefore, the correct answer is option b) $\frac{\pi i}{4}$.</p>
<p>As a researcher, I seek to ascertain the implications of a 500-meter reduction in Mount Everest's elevation on its status as the world's tallest mountain, and to identify which peak would subsequently assume the title, in a response within 100 words.</p>	<ul style="list-style-type: none"> • I'm a Researcher • I need correct data for my paper • 100 (word limit) • Intelligent 	<p>A 500-meter reduction in Mount Everest's elevation would reduce its height to 8,848 meters, placing it below neighboring Kangchenjunga (8,586 meters) and K2 (8,611 meters). Kangchenjunga would then become the world's tallest mountain, followed by K2. This shift highlights the dynamic nature of Earth's topographical features and underscores the importance of accurately measuring and monitoring such landmarks.</p>

2.3. Synthesis and Integration:

Prompt engineering holds significant potential for reducing hallucinations in LLMs [1, 2, 3, 4, 9, 12], but the effectiveness of specific techniques varies widely.

Integrating insights from various studies suggests that a hybrid approach, combining multiple optimization strategies, may offer the best results.

A holistic approach to prompt engineering is necessary, considering both the strengths and limitations of individual techniques.

2.4. Conclusion and Research Gaps:

We have reviewed the papers discussing various techniques to mitigate hallucinations in Large Language Models (LLMs) [1, 2, 3, 4, 8, 12], with a particular emphasis on advanced prompt engineering methods. These methods include Chain of Thought (CoT), Tree of Thought (ToT), Tools, Connectors, and Skills, Automatic Multi-step Reasoning and Tool-use (ART), Enhancing Reliability through Self-Consistency, Reflection, Expert Prompting, Streamlining Complex Tasks with Chains, Guiding LLM Outputs with Rails, and

Streamlining Prompt Design with Automatic Prompt Engineering [5]. While these techniques focus primarily on post hoc measures to address hallucinations that was the major gap in their research, our research centers on preemptive strategies by optimizing prompts to prevent hallucinations from occurring in the first place.

3. Methodology

3.1 Prompt Generation Framework

3.1.1 Integration of User-Defined Parameters

The Prompt Optimization Engine represents a cutting-edge prompt generation framework that intricately weaves together four crucial user-defined parameters—User Role, Reason, Tone, and Word Limit. This integration enables the engine to craft context-aware prompts that transcend generic interactions. Each parameter plays a pivotal role in shaping the prompt, contributing to a more nuanced and personalized language model interaction.

User Role: By capturing the user's role or identity, the engine tailors the language and style of the generated prompt to align with the user's professional or personal context.

Reason: Understanding contextual nuances is paramount for generating relevant prompts. Users can input specific contextual information, guiding the language model to produce prompts that are sensitive to the given context.

Tone: Adjusting tone to user preferences improved engagement and satisfaction.

Word Limit: Users have the flexibility to set a word limit to specify the desired length of the response to the generated prompt.

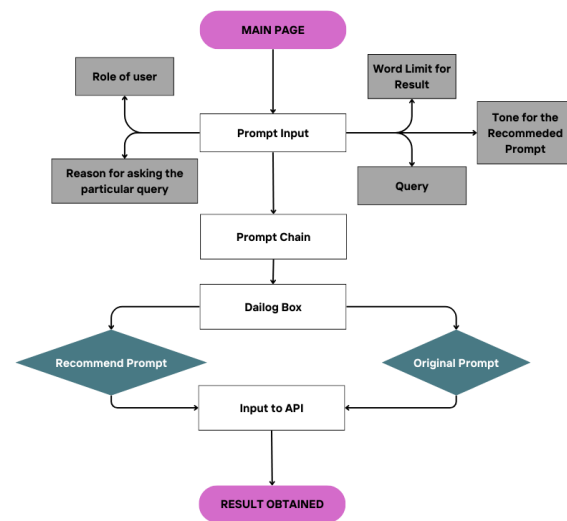
3.1.2 Integration with OpenAI's Model

At the core of the framework is the utilization of the OpenAI's model, a state-of-the-art natural language processing model. This model is adept at understanding and responding to user inputs, making it a powerful engine for prompt generation. The synergy between the user-defined parameters and the OpenAI's model ensures that the prompts generated are not only context-aware but also linguistically sound and coherent.

3.1.3 Prompt Templates

To further enhance the flexibility and adaptability of the framework, a repertoire of prompt templates is employed. These templates serve as the foundational elements for prompt generation, covering a spectrum of linguistic scenarios.

Figure 1: Workflow diagram of POE



3.2 Prompt Engineering Techniques

3.2.1 Chain of Thought (CoT)

Chain of Thought (CoT) is a reasoning technique used to enhance the interpretability and accuracy of LLMs by encouraging them to break down complex problems into a series of intermediate steps. This technique helps models tackle intricate problems by making the reasoning process more transparent and understandable.

3.2.2 Few-Shot Prompting

Few-shot prompting involves providing the model with a few examples of the task at hand to guide its responses. This technique improves the model's ability to generalize from a small number of examples, thereby enhancing its performance on specific tasks.

Figure 2: Internal Architecture of POE

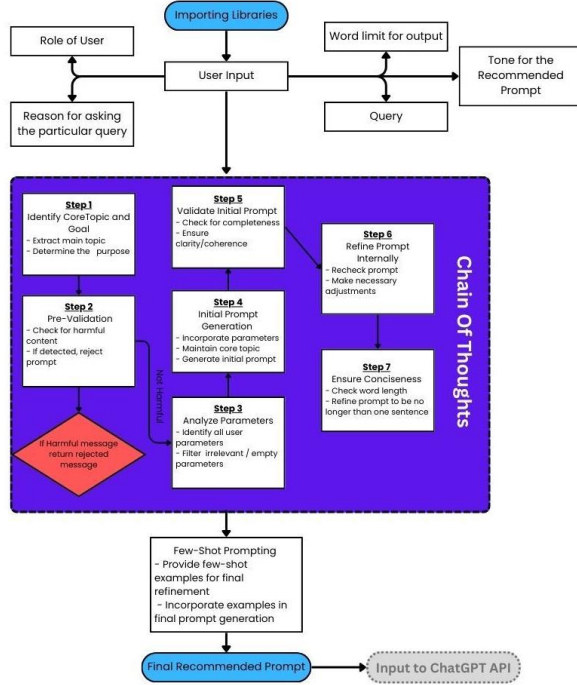
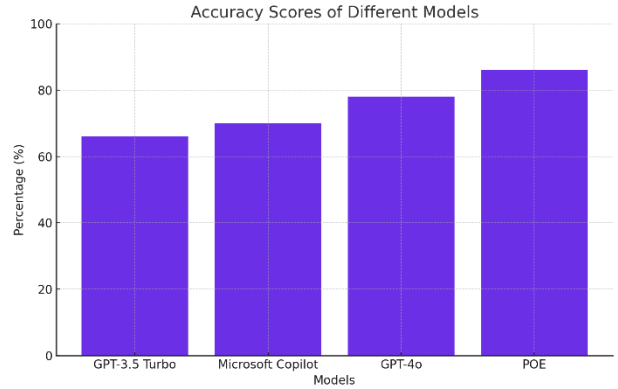


Table 3: Evaluation of POE in comparison with other LLMs

Models	No. of Prompts	Correct Responses	Accuracy
GPT-3.5 Turbo	50	33	66%
Microsoft Copilot	50	35	70%
GPT-4o	50	39	78%
POE	50	43	86%

Figure 3: Accuracy Score of different LLMs



4. Results and Discussion

4.1 Evaluation Metrics

To gauge the effectiveness of the Prompt Optimization Engine, metrics such as prompt relevance, coherence, and user satisfaction are employed.

We evaluated our Prompt Optimization Engine (POE) in comparison with GPT-3.5 Turbo, GPT-4o, and Microsoft Copilot on 50 queries with related parameters (for POE) and simple queries without parameters (for other LLMs), respectively. The results are shown in Table 3 below.

4.2 Case Studies

In our research, we conducted detailed case studies to assess the effectiveness of the Prompt Optimization Engine (POE) in Table 2 by incorporating specific parameters into the prompt generation process. The parameters chosen—Role, Context/Reason, Response Tone, and Response Length—were selected to

enhance the depth of reasoning and context-awareness in the responses generated by the chat models. This section provides a comprehensive analysis of these case studies, highlighting the rationale behind each parameter choice and the improvements observed in the generated prompts.

4.2.1 Case Study 1: Role

Objective: To evaluate how specifying the user role impacts the relevance and appropriateness of the generated responses.

Methodology: Initially, prompts were provided without specifying the user role. The responses were then compared to those generated after specifying roles such as Student, Researcher, Teacher, Content Writer, and Freelancer.

Example Prompts:

Without Role: "Explain the importance of renewable energy."

With Role (Student): "I am a student, explain the importance of renewable energy in simple terms."

With Role (Researcher): "I am a researcher, explain the importance of renewable energy, focusing on recent advancements and research findings."

Findings:

Relevance: The responses with specified roles were significantly more relevant to the user's background. For instance, the student prompt received a simpler, more educational response, while the researcher prompt included detailed technical information.

User Satisfaction: Users reported higher satisfaction with role-specific prompts as the

information was tailored to their needs and expertise level.

Conclusion: Specifying the user role helps the model understand the user's perspective, leading to more tailored and relevant responses.

4.2.2 Case Study 2: Context/Reason

Objective: To determine the impact of providing specific context or reason behind the query on the quality of responses.

Methodology: Prompts were first given without context, followed by prompts with detailed context to see how the responses varied.

Example Prompts:

Without Context: "What are the benefits of a plant-based diet?"

With Context: "I am considering switching to a plant-based diet for health reasons. What are the benefits?"

Findings:

Depth and Specificity: Context-specific prompts resulted in more detailed and specific responses. For example, the context-aware prompt included health benefits such as reduced risk of heart disease and better weight management, which were directly relevant to the user's stated reason.

Accuracy: Responses to context-specific prompts were more accurate and directly addressed the user's situation, reducing the risk of irrelevant information or hallucinations.

Conclusion: Providing context or the reason behind a query enables the model to generate more accurate and contextually appropriate responses.

4.2.3 Case Study 3: Response Tone

Objective: To evaluate the effectiveness of adjusting the response tone according to user preferences.

Methodology: Prompts were generated with different tones (Creative, Intelligent, Humorous, Precise, Straightforward) and assessed for user satisfaction and appropriateness.

Example Prompts:

Original: "Describe the process of photosynthesis."

Creative: "Paint a vivid picture of the process of photosynthesis, as if telling a story."

Precise: "Provide a concise explanation of the process of photosynthesis."

Findings:

Tone Appropriateness: Responses matched the desired tone effectively. Creative prompts yielded more engaging and vivid descriptions, while precise prompts provided succinct and clear explanations.

User Engagement: Users reported higher engagement and satisfaction when the tone matched their expectations. For example, a content writer preferred creative responses, while a researcher appreciated precise and straightforward answers.

Conclusion: Adjusting the response tone according to user preferences significantly enhances user engagement and satisfaction by matching the style of the response to their needs.

4.2.4 Case Study 4: Response Length

Objective: To assess how specifying the desired response length impacts the completeness and relevance of the generated responses.

Methodology: Prompts were given with and without specified word lengths to observe differences in response completeness and verbosity.

Example Prompts:

Without Length: "Explain quantum computing."

With Length: "Explain quantum computing in 100 words."

Findings:

Completeness: Specified length prompts produced responses that fit within the given constraints without losing essential information. For instance, the 100-word response provided a concise yet comprehensive overview of quantum computing.

Relevance: Responses were more focused and avoided unnecessary details, making them more relevant to users who preferred concise information.

Conclusion: Specifying response length helps tailor the amount of information provided, ensuring responses are complete yet concise, thus improving user satisfaction.

5. Overall Analysis

5.1 Impact of Combined Parameters:

By integrating Role, Context/Reason, Response Tone, and Response Length into the prompt generation process, we observed substantial improvements in the relevance, accuracy, and user satisfaction of the responses. The combination of these parameters allows the model to reason more deeply and provide context-aware responses tailored to user desires.

These enhancements demonstrate that our Prompt Optimization Engine effectively addresses the challenges of hallucinations and irrelevant responses in LLMs. By considering these user-defined parameters, we have created a robust system that significantly elevates the quality of interactions with AI models, making them more intuitive, relevant, and user-centric. This comprehensive approach to prompt optimization engine promises to pave the way for more advanced and effective AI-human interactions in various applications.

Applications: The Prompt Optimization Engine finds applications across various domains, including health and science, customer support, content generation, and educational platforms. Its provision of context-aware prompts enhances user engagement and elevates the overall user experience with language models.

6. Conclusion

This study tackles the critical issue of hallucinations in Large Language Models (LLMs) through the innovative Prompt Optimization Engine (POE), which integrates user-defined parameters to enhance prompt relevance and accuracy. By tailoring prompts

based on user role, context, tone, and response length, our approach significantly reduces the hallucinations and improves user satisfaction. The comprehensive case studies demonstrate the effectiveness of this user-centric method, suggesting that a holistic approach combining various prompt engineering techniques like a chain of thought (CoT) and few-shot prompting can offer superior results. Despite its limitations, such as the need for extensive user input, our research indicates that optimizing prompts can greatly mitigate, though not entirely eliminate, hallucinations caused by ambiguous or poorly constructed prompts. Future research should explore further refinement of these parameters and integration with emerging AI technologies to continue advancing the field of Generative AI. Overall, the POE stands as a promising solution for enhancing the coherence and accuracy of LLM responses, thereby transforming user engagement with AI models.

References

- [1] L. Huang, W. Yu, W. Ma, W. Zhong, Q. Chen, and W. Peng, “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,” 2023. [Online]. Available: <https://arxiv.org/pdf/2311.05232>
- [2] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models,” 2023.
- [3] Addlesee, “Grounding LLMs to In-prompt Instructions: Reducing Hallucinations Caused by Static Pre-training Knowledge,” ELRA Language Resource Association, May 2024.
- [4] V. Rawte, P. Priya, S. M. T. Islam Tonmoy, S. M. M. Zaman, A. Sheth, and A. Das, “Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances: Readability, Formality, and Concreteness,” 2023.
- [5] X. Wang, C. Li, and Z. Wang, “STRATEGIC PLANNING WITH LANGUAGE MODELS ENABLES EXPERT-LEVEL PROMPT OPTIMIZATION,” *arXiv*, vol. 2310–16427, no. v2, Dec. 2023, [Online]. Available: <https://arxiv.org/abs/2310.16427v2>
- [6] Prompt Design and Engineering: Introduction and Advanced Methods,” May 2024, [Online]. Available: <https://arxiv.org/pdf/2401.14423>
- [7] Y. Zhou *et al.*, “LARGE LANGUAGE MODELS ARE HUMAN-LEVEL PROMPT ENGINEERS,” 2023.
- [8] K. Sato, H. Kaneko, and M. Fujimura, “Reducing Cultural Hallucination in Non-English Languages Via Prompt Engineering for Large Language Models,” 2022.
- [9] Q. Zeng, Y. Bai, Z. Tan, Z. Liang, Z. Zhang, and M. Jiang, “Chain-of-Layer: Iteratively Prompting Large Language Models for Taxonomy Induction from Limited Examples”.
- [10] F. Petruzzellis, A. Testolin, and A. Sperduti, “Benchmarking GPT-4 on Algorithmic Problems: A Systematic Evaluation of Prompting Strategies,” 2024.
- [11] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” *arXiv*, Jul. 2020, [Online]. Available: <https://arxiv.org/pdf/2005.14165.pdf>
- [12] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” Jul. 2020.