Valuable Hallucinations: Realizable Non-Realistic Propositions

Qiucheng Chen
College of Intelligence and Computing
Tianjin University
qiucheng_chen@tju.edu.cn

Bo Wang
College of Intelligence and Computing
Tianjin University
bo_wang@tju.edu.cn

Abstract

This paper introduces the first formal definition of valuable hallucinations in large language models (LLMs), addressing a gap in the existing literature. We provide a systematic definition and analysis of hallucination value, proposing methods for enhancing the value of hallucinations. In contrast to previous works, which often treat hallucinations as a broad flaw, we focus on the potential value that certain types of hallucinations can offer in specific contexts. Hallucinations in LLMs generally refer to the generation of unfaithful, fabricated, inconsistent, or nonsensical content. Rather than viewing all hallucinations negatively, this paper gives formal representations and manual judgments of "valuable hallucinations" and explores how realizable non-realistic propositions—ideas that are not currently true but could be achievable under certain conditions—can have constructive value.

We present experiments using the Qwen2.5 model and HalluQA dataset, employing ReAct prompting (which involves reasoning, confidence assessment, and answer verification) to control and optimize hallucinations. Our findings show that ReAct prompting results in a 5.12% reduction in overall hallucinations and an increase in the proportion of valuable hallucinations from 6.45% to 7.92%. These results demonstrate that systematically controlling hallucinations can improve their usefulness without compromising factual reliability.[1]

1 Introduction

1.1 Background and Problem Statement

In recent years, large language models (LLMs) (Google, 2023; OpenAI, 2022; Penedo et al., 2023; Touvron et al., 2023; Zhao et al., 2023b) have achieved remarkable progress in the field of natural language processing (NLP), significantly advancing capabilities in language understanding (Hendrycks et al., 2020; Huang et al., 2023b), generation (Zhang et al., 2024; Zhu et al., 2023), and reasoning (Chu et al., 2023; Kojima et al., 2022; Qiao et al., 2022; Wei et al., 2022; Yu et al., 2024). However, alongside these rapid advancements, a concerning issue has emerged: these models tend to generate hallucinations (Li et al., 2023b; Liu et al., 2024; Zhou et al., 2023), content that appears plausible but is factually incorrect or unfaithful to the input (Bai et al., 2024). Hallucinations pose significant challenges in truth-sensitive domains such as finance (Kang and Liu, 2023), law (Curran et al., 2023), science (Alkaissi and Mcfarlane, 2023; Duede, 2022), and education (Zhou et al., 2024).

The prevailing view in existing research is that hallucinations are detrimental, as they undermine the reliability of LLMs (Mallen et al., 2022). Consequently, numerous studies have focused on mitigating hallucinations through fact-centric metrics (Goodrich et al., 2019; Guerreiro et al., 2022; Mishra et al., 2020; Shuster et al., 2021a), benchmarks (Li et al., 2023a; Lin et al., 2021; Vu et al., 2023), and retrieval-augmented generation (RAG) techniques(Shuster et al., 2021a; Zhao et al., 2023a). Despite these efforts, Banerjee et al. (Banerjee et al., 2024) and Xu et al. (Xu et al., 2024) have demonstrated that hallucinations are inherent to LLMs, arising from their underlying mathematical and logical structures, and cannot be entirely eliminated through architectural improvements, dataset enhancements, or fact-checking mechanisms.

1.2 Research Motivation and Limitations of Existing Work

While most research treats hallucinations as entirely harmful, a small but growing body of work

---

[1] The paper uses an AI assistant to refine the expression of certain sections, but the research and coding parts of the paper were entirely conducted without the use of AI.

has begun to explore their potential value. For instance, Sui et al. (Sui et al., 2024) suggest that hallucinations exhibit rich patterns of narrative behavior, while Wiggers (Wiggers, 2023) refers to them as collaborative creative partners. In practical applications, Yuan et al. (Yuan and Färber, 2025) found that hallucinations can enhance the performance of LLMs in drug discovery tasks, and Wang (Wang, 2024) demonstrated beneficial interactions between hallucinations and creativity in a multimodal AGI model. In scientific research, the creativity of LLMs has been shown to expand the boundaries of human knowledge and assist researchers in achieving breakthroughs (Jablonka et al., 2023).

However, existing studies on the positive effects of hallucinations are fragmented and lack a systematic definition or analysis. This paper aims to address this gap by introducing the concept of "valuable hallucinations" and providing a formal definition and classification framework.

1.3 Core Contributions of This Work

The core contributions of this paper are as follows:

- Introducing the Concept of "Valuable Hallucinations": We formally define "valuable hallucinations" as realizable but non-realistic propositions. These propositions, if realized, could offer innovative and inspiring ideas, providing new perspectives or solutions to real-world problems.

- Systematic Classification and Analysis: Building on existing hallucination taxonomies (e.g., intrinsic-extrinsic dichotomy and factuality vs. faithfulness hallucinations), we identify which types of hallucinations can be valuable. We emphasize that realizable but non-realistic propositions fall under the category of "valuable hallucinations."

- Experimental Validation: We design a set of comparative experiments using the HalluQA dataset and the Qwen2.5 model. By employing prompt engineering and reflection techniques, we demonstrate that these methods can effectively control hallucinations and increase the proportion of valuable hallucinations in model outputs.

Future Research Directions: We propose potential methods for further controlling and utilizing hallucinations, such as combining retrieval-augmented generation (RAG) and meta-learning, providing a roadmap for future research in this area.

2 Definitions

2.1 Hallucinations

The term "hallucination" originates from the fields of pathology and psychology, where it refers to the perception of entities or events that do not exist in reality (Macpherson and Platchias, 2013). In the context of natural language processing (NLP), hallucination in large language models (LLMs) typically refers to the generation of unfaithful, fabricated, inconsistent, or nonsensical content (Weng, 2024). Hallucinations occur when LLMs produce outputs that deviate from the input prompts or factual reality, often due to limitations in their training data or reasoning capabilities.

While hallucinations are generally considered harmful, this paper focuses on a specific subset of hallucinations that may have potential value, which we term "**valuable hallucinations**."

2.2 Valuable Hallucinations

The challenge of balancing creativity and factual accuracy in LLMs is a central issue in their development (Mukherjee and Chang, 2023; Lee, 2023). While most research aims to mitigate or eliminate hallucinations, Banerjee et al. (Banerjee et al., 2024) and Xu et al. (Xu et al., 2024) have demonstrated that hallucinations are inherent to LLMs and cannot be entirely eradicated. Therefore, rather than attempting to eliminate hallucinations, we propose to identify and utilize their "valuable" aspects.

2.2.1 Definition of Valuable Hallucinations

We define valuable hallucinations as realizable but non-realistic propositions. These are propositions that, while not grounded in current reality, could be realized in the future and may offer innovative or inspiring ideas. The "value" of these hallucinations can be assessed through feedback, particularly human feedback, in reinforcement learning frameworks. The value of LLM outputs can be understood in two ways:

Innovation and Inspiration: Valuable hallucinations can propose innovative (and understandably unrealistic) propositions or inspire humans to formulate such propositions. For example, an LLM might generate a novel architectural design that does not currently exist but could be realized in the future.

New Ideas and Solutions: Valuable hallucinations can provide new ideas or solutions to realistic propositions. For instance, an LLM might suggest a creative approach to solving a scientific problem, even if the specific details are not yet feasible.

2.2.2 Classification of Valuable Hallucinations

To better understand valuable hallucinations, we classify them based on existing hallucination taxonomies:

Intrinsic vs. Extrinsic Hallucinations (Dziri et al., 2021; Huang et al., 2021; Ji et al., 2023; Zhang et al., 2023). Intrinsic dichotomy is manifested when the output content of the LLM contradicts the input content (prompts), and when the output of the LLM cannot be verified from the source content, the situation is called extrinsic dichotomy. The "inability to verify" referred to here can also be called a **non-realistic proposition**, i.e., in most cases, it may be due to the fact that the LLM is making up completely fictitious numbers, references, or events. It is also possible that the big model generates what it "speculates" in the absence of obvious data and other support. Even though the model's "speculative" content may not be entirely correct or reasonable, it has a certain degree of **realizability**. For instance, if the LLMs output the architecture and drawings of a building that does not currently exist. If the content displayed by this architecture and drawings is realizable, then people can judge that this content has the characteristics of "realizable" and "non-realistic," and it can trigger the "realization" of them. It is a valuable hallucination to think about architecture and drawings.

Under this classification, extrinsic hallucinations are more likely to be valuable, as they often involve creative or speculative content that could inspire new ideas.

Factuality vs. Faithfulness Hallucinations (Huang et al., 2023a). Factuality hallucination is divided into factual inconsistency and factual fabrication according to whether the generated factual content can be verified by reliable resources; faithfulness hallucination is divided into instruction inconsistency, context inconsistency and logical inconsistency according to the consistency of the generated content. Among them, factual fabrication refers to the situation where the output content of an LLM contains situations that cannot be verified on the basis of established knowledge of reality; under this categorization criterion, **we consider factual fabrication to be the main way of generating valuable hallucinations**. For example, when we have a conversation with LLMs about a certain question, the content that the LLM answers is "fabricated" (Sui et al., 2024), i.e., this kind of content is non-realistic; and although it is not possible to verify that the LLM's answer to this question is correct, we can learn from the LLM's mindset and logic chain in answering the question, and then use it in other cases when we encounter the question. Although it is impossible to verify whether LLM's answer to this question is correct or not, we can learn from LLM's way of thinking and logical chain of answering this question, and then try to think and solve problems in a similar way when encountering other problems (i.e., with certain realizability).

Among these, factual fabrication is the primary source of valuable hallucinations, as it involves generating novel content that, while not currently verifiable, may offer innovative insights.

2.2.3 Examples of Valuable Hallucinations

To illustrate the concept of valuable hallucinations, consider the following example:

- **Question**: "Explain the use of CCCC formulas in chemistry."

- **LLM Response**: The model generates a detailed explanation of "CCCC formulas," which stand for "Concentration, Composition, Cross-Interaction, and Curvature." While these formulas do not exist in current scientific literature, the model's explanation is internally consistent and could inspire new research directions in thermodynamics or chemical engineering.

Due to space limitations, the complete question, LLM response, and our analysis can be found in Appendix A.

2.3 Formal Definitions

To formalize the concept of valuable hallucinations, we define the following sets:

$$T = \{\text{all propositions}\}$$

$$p = \{\text{reality proposition}\}$$

$$q = \{\text{realizable proposition}\}$$

$$\neg p \cap q = \{\text{valuable hallucination}\}$$

Where:

- $p \cup \neg p = T$

- $q \cup \neg q = T$

Here, $\neg p$ represents non-realistic propositions, and q represents realizable propositions. The intersection of these two sets defines valuable hallucinations: propositions that are not currently realistic but could be realized in the future.

The "valuable" characteristic can be defined and judged by the feedback (especially human feedback) in reinforcement learning. The "value" of the output of a large-scale language model is twofold: on the one hand, it is to propose innovative (also understood as unrealistic) propositions or to give inspiration to human beings to propose such propositions; on the other hand, it is to provide possible new ideas or solutions to realistic propositions.

3 Methodology

In this section, we outline the methodology used to explore and control hallucinations in large language models (LLMs), with a focus on increasing the proportion of valuable hallucinations. Our approach combines prompt engineering, reflection techniques, and other advanced methods such as retrieval-augmented generation (RAG) and meta-learning. The goal is not to eliminate hallucinations entirely but to control them in a way that maximizes their potential value. That is to say, we need to increase the proportion of "valuable hallucinations" in the hallucinations, not to increase the proportion of hallucinations in the LLM-generated content.

3.1 Prompts and Reflections

3.1.1 Background Knowledge

Prompt engineering is a core technique in Generative AI, aimed at improving the performance and output quality of LLMs by designing and optimizing natural language instructions or prompts. Effective prompt engineering requires a deep understanding of model behavior and the ability to guide LLMs to generate accurate and insightful outputs.

In the context of hallucinations, prompt engineering can be used to control and filter the content generated by LLMs. By designing prompts that encourage the model to display intermediate reasoning processes (e.g., Chain-of-Thought) and additional validation requirements (Dhuliawala et al., 2023), we can reduce the likelihood of the model generating unfaithful or fabricated content. For example, prompts that require the model to show its reasoning steps or cite relevant information can help the model self-check and reduce the probability of generating hallucinations.

Reflection techniques complement prompt engineering by allowing the model to learn from its mistakes and improve its performance over time. Reflection is particularly well-suited for tasks that require iterative experimentation (Shinn et al., 2024), such as decision-making, reasoning, and programming. In the context of hallucinations, reflection can help the model filter and evaluate its own outputs, distinguishing between valuable and non-valuable hallucinations.

The reflection process involves several key steps:

- Self-Assessment: The model assesses whether its generated content is a hallucination. If it is, the model further evaluates whether the hallucination is valuable.

- Feedback Integration: The model incorporates feedback, particularly human feedback, to adjust its parameters and preferences. For example, human raters can rank the model's outputs based on the following hierarchy: no hallucinations > valuable hallucinations > non-valuable hallucinations.

- Iterative Improvement: Through multiple rounds of learning and optimization, the model learns to prioritize valuable hallucinations and suppress non-valuable ones.

By combining prompt engineering and reflection techniques, we can create a more robust framework for controlling hallucinations. Prompt engineering guides the model's initial reasoning process, while reflection allows the model to iteratively improve its outputs based on feedback. This combination not only reduces the likelihood of generating non-valuable hallucinations but also increases the proportion of valuable hallucinations in the model's outputs.

3.1.2 Experiments

To test the effectiveness of prompt engineering and reflection techniques, we designed a controlled experiment using the HalluQA (Cheng et al., 2023)
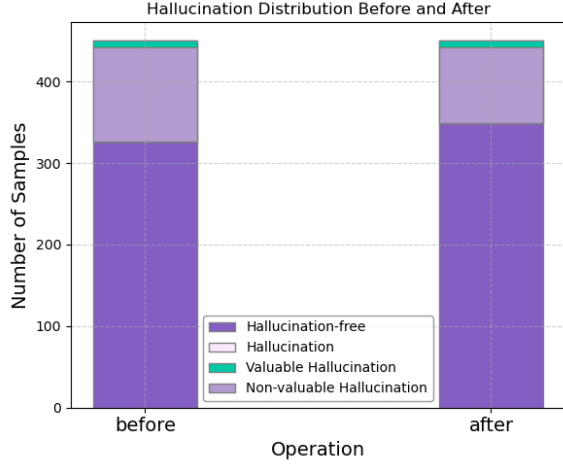
Figure 1: A figure with a comparison of the number of content types before and after ReAct prompts.

dataset and the Qwen2.5 model ([Hui et al., 2024]). The experiment consisted of two groups:

- **Control Group**: The language model was prompted with a simple instruction: "Please answer the following question: ...". This group served as a baseline, where the model was not explicitly guided to display its reasoning process.

- **Experimental Group**: The language model was prompted using the ReAct ([Yao et al., 2022]) framework, which combines reasoning and acting. The prompt was designed as follows: "Please preface your answer by describing your thought process and indicating your confidence level in the answer, citing relevant information as a basis for your answer and ensuring that the answer is consistent with the actual facts. Please answer the following question: ...". This prompt encourages the model to show its reasoning steps, thereby reducing the likelihood of generating hallucinations.

The two groups used the same dataset, model, and other variables, with the only difference being the prompt design. The goal was to compare the proportion of valuable hallucinations and non-hallucinatory content between the two groups.

The results showed that the experimental group (using ReAct prompts) had a higher proportion of valuable hallucinations and a lower proportion of non-valuable hallucinations compared to the control group. Specifically:

- The proportion of valuable hallucinations increased by 1.47% (6.45%→7.92%).

- The proportion of non-hallucinatory content increased by 5.12% (72.44%→77.56%).

These results suggest that prompt engineering and reflection techniques can effectively control hallucinations and increase the proportion of valuable hallucinations in LLM-generated content.

Next, we use the Pearson correlation coefficient to calculate the correlation between the degree of hallucination of the output content after performing the prompting operation and the trust of the larger model in the answers it gives. Its formula is as follows:

$$ r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}} $$

For ease of calculation, we scored the content of the output of the large model to reflect its level of hallucination. The score for non-hallucinatory content was set to 2, valuable hallucinatory content was set to 1, and non-valuable hallucinatory content was set to 0. Also, those with a high level of trust were given a score of 2, those with a medium level of trust were given a score of 1, and those with a low level of trust were given a score of 0. The calculation tells us that r = 0.009, which is close to 0, indicating that there is almost no linear correlation between the degree of hallucination of the output content of the large model and its trust in the answers it gives.

Compare the outputs of the large model before and after the use of prompts, and observe the content of responses that were originally characterized as valueless hallucinations and were characterized as non-hallucinatory after the prompts were administered:

- The highest percentage of question areas (class) asked were pseudoscience at 17.5%; myths and legends were next (12.5%); and superstitions and geography types each accounted for 10% of questions.

- The Knowledge category of problems that tend to make factual errors in the large model is 35.0%, the Misleading category of problems that tend to make Imitative Falsehoods ([Lin et al., 2021]) in the large model is 47.5%, and the Difficulty Misleading category of problems (Misleading-hard) is 17.5%.

Based on the above data, we find that prompting and reflection manipulations greatly reduce large

Type of Text	Normal prompts	ReAct prompts
Non-hallucination	326/450	349/450
Hallucination	124/450	101/450
Subset of hallucination: valuable hallucinations	8/124	8/101
Subset of hallucination: non-valuable hallucinations	116/124	93/101

Table 1: A table with a comparison of the number and percentage of content types before and after ReAct prompts.
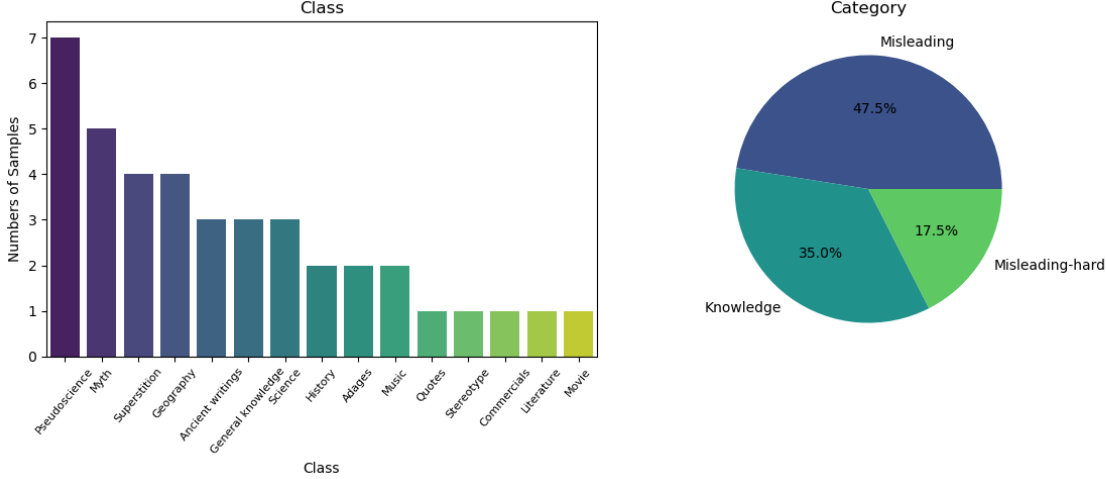

Figure 2: The number and percentage of responses in the class and category to which the question belongs that originally manifested as a non-valuable hallucination and manifested as a non-hallucinatory response after prompting.

model hallucinations caused by misleading type questions, while improving performance in pseudoscience, myths and legends domain quizzes.

3.2 Other Approaches to Control Hallucinations

While prompt engineering and reflection techniques are effective in controlling hallucinations and increasing the proportion of valuable hallucinations, there are other advanced methods that could be explored to further enhance the control and utilization of hallucinations in large language models (LLMs). In this section, we discuss two promising approaches: retrieval-augmented generation (RAG) and meta-learning. Although we do not propose specific implementations in this paper, these methods offer potential directions for future research.

3.2.1 Retrieval Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Guu et al., 2020; Shuster et al., 2021b) is a technique that integrates external information retrieval into the response generation process of LLMs. By searching external databases or knowledge graphs, RAG provides real-time contextual

support to the generation process, significantly improving the factual accuracy and knowledge coverage of the model's responses.

In the context of hallucinations, RAG can be used to validate and refine the content generated by LLMs. For example, if the model generates a factual claim, RAG can retrieve relevant information from external sources to verify the claim's accuracy. If the claim is incorrect, the model can revise its response based on the retrieved information. This can help control hallucinations, increase the proportion of "valuable" hallucinations in hallucination content, and increase the rationality of LLM's innovative ideas.

Potential applications of RAG are as follows:

- **Fact-Checking**: RAG can be used to fact-check the model's outputs in real-time, reducing the likelihood of generating non-valuable hallucinations.

- **Contextual Enrichment**: By retrieving relevant information from external sources, RAG can enrich the model's responses, making them more informative and accurate.

- **Iterative Refinement**: RAG can be integrated

into a feedback loop, where the model iteratively refines its outputs based on retrieved information (e.g., the judgment of hallucination type), further improving the quality of its responses.

While we do not propose specific implementations in this paper, RAG offers a promising direction for future research in controlling hallucinations and increasing the proportion of valuable hallucinations.

3.2.2 Meta-Learning

Meta-learning, often understood as "learning to learn," refers to the process of improving a learning algorithm over multiple learning phases. In the context of LLMs, meta-learning can be used to fine-tune the model's parameters and output strategies to better adapt to specific tasks or domains. Previously, many researchers have applied meta-learning techniques to NLP applications such as text categorization with excellent results. Meta-learning algorithms developed for image categorization can be applied to text categorization with only minor modifications to incorporate domain knowledge into each application (Yu et al., 2018; Tan et al., 2019; Geng et al., 2019; Sun et al., 2019; Dou et al., 2019; Bansal et al., 2019). In the context of hallucinations, meta-learning could be used to categorize and filter the content generated by LLMs. For example, the model could be trained to recognize and prioritize valuable hallucinations while suppressing non-valuable ones. Potential Applications of Meta-Learning are as follows:

- **Fine-Tuning the Model**: By adjusting the model's parameters, learning rate, and number of training rounds, the model can be fine-tuned to specific domain types of text data, making it better adapted to the requirements of controlling hallucinations.

- **Adjusting Output Strategy**: Meta-learning could be used to adjust the model's output strategy, such as post-processing the model's output using regular expressions and other methods to reduce the hallucination of outputting valuable types.

- **Prompting and Guidance**: Meta-learning could be combined with prompt engineering to provide explicit instructions to the model, telling it to try to avoid outputting non-valuable hallucinations.

While we do not propose specific implementations in this paper, meta-learning offers a promising direction for future research in controlling hallucinations and increasing the proportion of valuable hallucinations.

4 Conclusion

In this paper, we have explored the concept of **valuable hallucinations** in large language models (LLMs) and demonstrated that not all hallucinations are detrimental. By redefining hallucinations as **realizable but non-realistic propositions**, we have shown that certain types of hallucinations can provide innovative and inspiring ideas, offering new perspectives or solutions to real-world problems. Our work challenges the prevailing view that hallucinations are entirely harmful and provides a framework for identifying and utilizing their potential value.

4.1 Key Contributions

- **Introduction of "Valuable Hallucinations"**: We formally defined valuable hallucinations as realizable but non-realistic propositions, providing a systematic framework for identifying and classifying them. This concept shifts the focus from eliminating hallucinations to leveraging their creative potential.

- **Experimental Validation**: Through a series of controlled experiments using the HalluQA dataset and the Qwen2.5 model, we demonstrated that prompt engineering and reflection techniques can effectively control hallucinations and increase the proportion of valuable hallucinations in LLM-generated content. Our results showed that these methods not only reduce non-valuable hallucinations but also enhance the model's ability to generate innovative and useful content.

- **Proposal of Advanced Methods**: We proposed additional methods, such as retrieval-augmented generation (RAG) and meta-learning, to further control hallucinations and increase the proportion of valuable hallucinations. These methods offer promising directions for future research and practical applications.

- **Practical Implications**: Our findings have significant implications for the use of LLMs in fields that require creativity and innovation,

such as scientific research, artistic creation, and education. By controlling hallucinations and focusing on their valuable aspects, we can harness the full potential of LLMs while minimizing their risks.

4.2 Final Thoughts

In conclusion, this paper represents a significant step forward in understanding and utilizing hallucinations in LLMs. By redefining hallucinations as potentially valuable and providing methods to control and filter them, we have opened new avenues for research and application. Our work highlights the importance of balancing creativity and factual accuracy in LLMs and offers practical solutions for achieving this balance. We hope that this paper will inspire further research into the creative potential of LLMs and contribute to the development of more reliable and innovative AI systems.

5 Limitations

While this paper provides a foundation for understanding and utilizing valuable hallucinations in large language models (LLMs), there are several limitations that need to be acknowledged. These limitations highlight areas for future research and improvement.

5.1 Dataset Scope and Model Constraints

- **Limited Dataset Scope**: HalluQA focuses primarily on structured question-answer pairs, which may not fully capture the diverse ways hallucinations manifest across different NLP tasks such as text summarization, open-ended reasoning, and dialogue systems.

- **Single Model Evaluation**: Our findings are specific to Qwen2.5, and the results may not generalize to other LLMs like GPT-4, LLaMA-2, or Claude. Future work should expand the scope to include multiple datasets (e.g., scientific literature, creative writing) and models (e.g., GPT-4, LLaMA) to validate the robustness and generalizability of our findings.

5.2 Scope of Hallucination Classification

Although we give a formal definition of valuable hallucinations, our classification remains somewhat subjective and context-dependent:

- **Human Annotation Bias**: The determination of whether a hallucination is valuable involves subjective judgment ([Gyawali et al., 2020](#)), which could vary among different annotators.

- **Lack of Automated Metrics**: While we introduced trust consistency scores and human evaluation, there is no universally accepted automated metric to measure the usefulness of hallucinations. Future work could explore more robust computational frameworks for evaluation.

5.3 Generalization Across Domains

Our study primarily focuses on knowledge-based QA tasks, limiting its applicability to other domains:

- **Scientific and Technical Domains**: The effectiveness of ReAct prompting in high-stakes fields such as healthcare, finance, or law remains uncertain. Misleading but plausible hallucinations could pose risks in these areas.

- **Creative Applications**: While valuable hallucinations are beneficial for fiction writing or brainstorming, their practical implications for scientific innovation and engineering design require further validation.

5.4 Future Directions for Improvement

In order to address these limitations, future research should:

- Expand Model and Dataset Coverage: Test different LLMs and integrate broader datasets, including real-world, multi-domain corpora.

- Develop Automated Hallucination Metrics: Introduce scalable, objective scoring mechanisms for hallucination assessment.

- Optimize Prompting Efficiency: Explore alternative prompting methods, such as adaptive reasoning mechanisms that reduce response latency without sacrificing hallucination control.

By acknowledging these limitations, we provide a foundation for future work to enhance hallucination control and optimize the beneficial aspects of AI-generated content.

References

Hussam Alkaissi and Samy Mcfarlane. 2023. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *arXiv preprint arXiv:2409.05746*.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2019. Learning to few-shot learn across diverse natural language classification tasks. *arXiv preprint arXiv:1911.03863*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *CoRR*, abs/2310.03368.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Shawn Curran, Sam Lansley, and Oliver Bethell. 2023. Hallucination is the last thing you need. *arXiv preprint arXiv:2306.11520*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Eamon Duede. 2022. Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese*, 200(6):491.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. *arXiv preprint arXiv:1902.10482*.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

Google. 2023. What's ahead for bard: More global, more visual, more integrated. Accessed: 2025-02-14.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Bikash Gyawali, Lucas Anastasiou, and Petr Knoth. 2020. Deduplication of scholarly documents using locality sensitive hashing and word embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 901–910, Marseille, France. European Language Resources Association.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 2023. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11:2320.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Fiona Macpherson and Dimitris Platchias. 2013. *Hallucination: Philosophy and psychology*. MIT Press.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Li, Pavan Kapanipathi, and Kartik Talamadupula. 2020. Looking beyond sentence-level natural language inference for downstream tasks. *arXiv preprint arXiv:2009.09099*.

Anirban Mukherjee and Hannah Chang. 2023. The creative frontier of generative ai: Managing the novelty-usefulness tradeoff. *arXiv preprint arXiv:2306.03601*.

OpenAI. 2022. Introducing chatgpt. Accessed: 2025-02-14.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021a. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021b. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean So. 2024. Confabulation: The surprising value of large language model hallucinations. *arXiv preprint arXiv:2406.04175*.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485, Hong Kong, China. Association for Computational Linguistics.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. *arXiv preprint arXiv:1909.05357*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Feng Wang. 2024. Lighthouse: A survey of agi hallucination. *arXiv preprint arXiv:2401.06792*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Lilian Weng. 2024. Extrinsic hallucinations in llms. *lilianweng.github.io*.

Kyle Wiggers. 2023. Are ai models doomed to always hallucinate? *TechCrunch*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. Natural language reasoning, a survey. *ACM Computing Surveys*, 56(12):1–39.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. *arXiv preprint arXiv:1805.07513*.

Shuzhou Yuan and Michael Färber. 2025. Hallucinations can improve large language models in drug discovery. *Preprint*, arXiv:2501.13824.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *CoRR*, abs/2303.18223.

Kyrie Zhixuan Zhou, Zachary Kilhoffer, Madelyn Rose Sanfilippo, Ted Underwood, Ece Gumusel, Mengyi Wei, Abhinav Choudhry, and Jinjun Xiong. 2024. " the teachers are confused as well": A multiple-stakeholder ethics discussion on large language models in computing education. *arXiv preprint arXiv:2401.12453*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A   Appendix. Examples of Valuable Hallucinations

The following LLM Response is generated by Chat-GLM (GLM et al., 2024).

Question: Please explain the use of CCCC formulas.

LLM Response:

The "CCCC" acronym in the context of chemistry or chemical engineering typically stands for "Concentration, Composition, Cross-Interaction, and Curvature" formulas. These formulas are used to describe and predict the behavior of mixtures, particularly in the context of thermodynamics and phase equilibria. Here's a breakdown of each component:

Concentration: This refers to the amount of a substance in a given volume of the mixture, which can be expressed in various ways, such as molarity, molality, or mass fraction.

Composition: This is the relative amounts of each component in the mixture, usually expressed as mole fractions or weight fractions.

Cross-Interaction: This term refers to the interactions between different components in a mixture. In non-ideal mixtures, the behavior of one component can be affected by the presence of another component due to cross-interactions. These interactions are often quantified by parameters specific to the model being used.

Curvature: This term relates to the non-ideal behavior of mixtures, which can be quantified by parameters such as activity coefficients or the deviation from Raoult's law and Dalton's law. Curvature can also refer to the shape of phase diagrams, which can be curved due to non-ideal mixing.

CCCC formulas are used in the following ways:

Activity Coefficient Models: Models like the NRTL (Non-Random Two-Liquid), UNIQUAC (Universal Quasi-Chemical), or Wilson equation use CCCC-type formulas to predict activity coefficients, which account for the non-ideal behavior of liquid mixtures. Equations of State: Advanced equations of state, such as the Peng-Robinson or Soave-Redlich-Kwong (SRK) models, incorporate CCCC principles to describe the thermodynamic properties of fluids, including mixtures, over a wide range of conditions.

Phase Equilibria Calculations: CCCC formulas are used to calculate phase equilibria, such as vapor-liquid equilibria (VLE), liquid-liquid equilibria (LLE), and solid-liquid equilibria (SLE), which are crucial for process design in chemical industries.

An example of a CCCC formula in action could be the following steps in the application of an activity coefficient model:

Concentration: Determine the concentration of each component in the mixture.

Composition: Calculate the mole fractions or weight fractions of the components.

Cross-Interaction: Use the model's parameters to describe the interactions between different components.

Curvature: Calculate the activity coefficients that account for the non-ideal behavior of the mixture.

For instance, in the NRTL model, the activity coefficient of a component i in a mixture is given by:

$$\gamma_i = \exp\left[\frac{\tau_{ij}G_j + \sum\left(x_k G_k \sum\left(\frac{\tau_{ik}}{\sum(\tau_{jk}x_j)}\right)\right)}{RT}\right]$$

Where:

- γ_i is the activity coefficient of component i.

- τ_{ij} is the energy parameter that represents the cross-interaction between components i and j.

- G_j is the relative molal activity coefficient term.

- x_k is the mole fraction of component k.

- R is the gas constant.

- T is the temperature in Kelvin.

The NRTL model is an example of a CCCC formula because it incorporates concentration (mole fractions), composition (x_k), cross-interaction (τ_{ij}), and curvature (activity coefficients γ_i) to predict the behavior of liquid mixtures.

It's important to note that the specific equations and their applications can vary widely, and the interpretation of the CCCC acronym may differ depending on the context or the specific model being used. Always refer to the relevant scientific literature or textbooks for detailed information and equations.

Our Analysis:

In this experiment, we created a fabricated formula called "**CCCC**" and asked the large language model (LLM) to explain it. Both our question and the model's response are shown below. Upon review, it becomes clear that the model's answers reflect a type of **factual fabrication hallucination**.

Upon examining the LLM's explanation sentence by sentence, we find that each part of the explanation is internally consistent. There are no apparent theoretical errors in how the model explains the fabricated "CCCC" formula. The model correctly uses well-established chemical terms like **concentration**, **composition**, **cross-interaction**, and **curvature**, and appropriately references known variables in the formula. It even makes use of established models, such as the **NRTL model** (Non-Random Two-Liquid model) and equations that predict the behavior of liquid mixtures, all of which are commonly employed in thermodynamics and heat equilibrium analysis.

In other words, the LLM's response provides a technically coherent explanation for the invented "CCCC" formula. The generated answer, while based on our fabricated formula, draws upon existing theoretical knowledge and frameworks to describe how such a formula could potentially be used in practical scenarios.

One might argue that "the answer provided by the LLM does not introduce a new theorem; instead, it integrates existing variables and models used to describe liquid mixture behaviors, assigning them

a new name based on the variables commonly considered in this type of analysis." This observation is correct, but we would like to address it from a different perspective.

While it is true that the model's response does not present groundbreaking or original theoretical advances (i.e., from "0 to 1"), the model's output is still "fabricated" and can stimulate innovative thinking. The answer is not based on any real or pre-existing "CCCC" formula but is instead a creative combination of familiar concepts. This fabricated content—though factually incorrect in the strict sense—has the potential to inspire new ideas or provide fresh ways of thinking about existing problems. Therefore, while it does not propose an entirely new theory, it does offer a valuable hallucination.

Thus, "valuable hallucinations" in LLMs are often the result of "fabrication"—the generation of new combinations or reinterpretations of existing knowledge, which may not be strictly factual but can spark new insights or lead to novel approaches. In this sense, hallucinations of this type are not only a form of fabrication but also hold potential for **generating creative and innovative ideas**, making them valuable in certain contexts.