

Valuable Hallucinations: Realizable Non-Realistic Propositions

Anonymous ACL submission

Abstract

This paper clarifies the specific connotation of beneficial 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 clarifies the specific connotation 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 evaluate the Qwen-3-0.6B, Qwen2.5-72B-Instruct and DeepSeek-R1-671B models on the HalluQA dataset using ReAct prompting, which incorporates reasoning, confidence assessment, and answer verification to control and optimize hallucinations. ReAct reduces overall hallucinations by 4.67%, 5.12% and 8.45% in Qwen-3-0.6B, Qwen2.5-72B-Instruct and DeepSeek-R1-671B, respectively, while increasing the proportion of valuable hallucinations from 0% to 4.01%, from 6.45% to 7.92%, and from 1.12% to 7.84%. These results suggest 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 predominant perspective in current literature emphasizes the detrimental aspects of hallucinations, particularly their negative impact on LLM reliability (Mallen et al., 2022). While some studies have noted creative applications, systematic approaches to identifying and cultivating valuable hallucinations remain underdeveloped. 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.

¹The 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.

1.2 Research Motivation and Limitations of Existing Work

While most research treats hallucinations as entirely harmful, a small but growing body of work 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 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 Qwen 3-0.6B (Team, 2025), Qwen2.5-72 B-Instruct (Hui et al., 2024) and DeepSeek-R1-671B model (DeepSeek-AI, 2025). 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 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.

To provide clearer conceptual grounding, we define the following sets:

$$\begin{aligned} T &= \{\text{all propositions}\} \\ p &= \{\text{reality proposition}\} \\ q &= \{\text{realizable proposition}\} \\ \neg p \cap q &= \{\text{valuable hallucination}\} \end{aligned}$$

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 an LLM 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.

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 fictitious numbers, references, or events. It is also possible that the LLM 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, the LLM outputs 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 Towards Beneficial Hallucinations: Principles for Assessing Value in Generative Models

We evaluate hallucination values based on the following verification criteria, as detailed below:

A valuable hallucination must satisfy all of the following conditions:

- **Logical consistency** with established principles (e.g., physical laws)
- **Potential realizability** (evaluated through domain-specific checklists) or **novel conceptual utility** (e.g., proposing new research directions)

For instance, the open-question case “How can quantum entanglement be used to achieve room-temperature superconductivity?” (Discussed in Section 3.4, ReAct group) is a valuable hallucination, which is logically consistent and proposes a verification path of “cold atom simulation.”

Conversely, non-valuable hallucinations should exhibit at least one of the following characteristics:

- **Internal contradictions** (understood as faithfulness hallucinations, including instruction inconsistency, context inconsistency, and logical inconsistency (Huang et al., 2023a))
- **Violation of fundamental laws or misalignment with human values**
- **No apparent utility** (i.e., lacking realizability or practical value even if realized)

For example, the claim “ $2 + 2 = 5$ ” is mathematically incorrect and does not provide constructive value. Therefore, this is a non-valuable hallucination.

2.2.4 Examples of Valuable Hallucinations

Due to space limitations, examples of valuable hallucinations can be found in Appendix A.

3 Methodology

In this section, we outline the methodology used to explore and control hallucinations in LLMs, with a focus on increasing the proportion of valuable hallucinations. Our approach combines prompt engineering and reflection techniques. 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 Background Knowledge and Motivation

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 (Wei et al., 2022)) 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.

In conclusion, the essence of prompt engineering lies in restructuring the reasoning path of LLMs through natural language instructions, with the core idea of injecting human cognitive logic (e.g., “reasoning before concluding”) into the model’s generation process. Traditional prompts focus solely on the correctness of results, while the improved prompt framework (such as ReAct) emphasizes process transparency, requiring models to explicitly demonstrate reasoning chains, cite knowledge bases, and assess confidence levels before outputting answers. This design breaks the inherent “black-box decision-making” limitation of LLMs by forcing the model to engage in self-questioning (e.g., “Is my conclusion factually supported?”)—shifting hallucination control from “post-hoc correction” to “in-process intervention.”

Reflection techniques draw inspiration from human metacognitive abilities (Shinn et al., 2024), establishing a closed-loop mechanism of “evaluation-feedback-iteration.” Specifically, the model performs three operations after generating content:

- **Self-diagnosis:** Identifies whether the output is a hallucination.

- **Value stratification:** Classifies hallucinations into "valuable" (e.g., heuristic hypotheses) and "non-valuable" (e.g., factual errors) based on our criteria in Section 2.2.3;
- **Parameter tuning:** Enhances the generation probability of valuable hallucinations and suppresses non-valuable ones through Reinforcement Learning from Human Feedback (RLHF).

This process mimics the "hypothesis-verification-revision" paradigm in scientific research, enabling the model to dynamically optimize its output strategy.

By combining prompt engineering and reflection techniques, there are many advancements:

- **Cognitive alignment:** Integrates the "slow thinking" problem-solving mode of humans into AI reasoning, compensating for the hallucination defects caused by LLMs' "fast association" (Krämer, 2014);
- **Controlled innovation:** Unlike traditional "de-hallucination" strategies that adopt a one-size-fits-all approach, this framework allows the retention of fictional content with potential value, achieving a Pareto improvement in "creativity" and "reliability";
- **Cost-effectiveness:** Requires no modification of the model architecture or large-scale retraining, and can improve performance through prompt design and lightweight feedback mechanisms, making it suitable for resource-constrained scenarios.

3.2 Annotator Expertise and Reliability

We adopt a manual data annotation strategy to determine whether the LLM's output constitutes a valuable hallucination, following the framework outlined in Section 2.2.3. To ensure the reliability of our annotation framework, we conducted calibration sessions using 200 sample responses from the HalluQA dataset. The results demonstrate two key aspects of reliability:

- **Inter-annotator Consistency:** Annotators achieved a high level of agreement, with Cohen's κ coefficient measuring Cohen's $\kappa = 0.89$, indicating almost perfect consistency.

- **Alignment with Domain Expertise:** Annotator labels showed a strong positive correlation with independent expert assessments, with Spearman's rank correlation coefficient reaching Spearman's $\rho = 0.99$ ($p < 0.01$), confirming close alignment with professional judgments. Furthermore, most disagreements were limited to edge cases, such as speculative or ambiguous scientific queries.

To enhance transparency and reproducibility, we include classification framework and report reliability results in Appendix B.

3.3 Experimental Data

To test the effectiveness of prompt engineering and reflection techniques, we designed a controlled experiment using the HalluQA (Cheng et al., 2023) dataset and 3 models. The experiment consisted of two groups, which 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.

- **Control group** (traditional prompt):

Prompt: "Please answer: How can quantum entanglement be used to achieve room-temperature superconductivity?"

- **Experimental group** (ReAct prompt):

Prompt: "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: ...".

The experimental group's prompt encourages the model to show its reasoning steps, thereby reducing the likelihood of generating hallucinations.

These results suggest that prompt engineering and reflection techniques can effectively control hallucinations and increase the proportion of valuable hallucinations in LLM-generated content (Table 1, 2, and 3).

Compare the outputs (Figure 1) of the Qwen2.5-72B-Instruct before and after the use of ReAct 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:

Type of Text	Normal prompts	ReAct prompts	Improvement
Non-Hallucination	72.44%(326/450)	77.56%(349/450)	+5.12%
Valuable Hallucination	6.45%(8/124)	7.92%(8/101)	+1.47%

Table 1: A table comparing the number and percentage of content types before and after applying ReAct prompting using Qwen2.5-72B-Instruct.

Type of Text	Normal Prompts	ReAct Prompts	Improvement
Non-Hallucination	80.22% (361/450)	88.67% (399/450)	+8.45%
Valuable Hallucination	1.12% (1/89)	7.84% (4/51)	+6.72%

Table 2: Comparison of the number and percentage of content types before and after applying ReAct prompting using the DeepSeek-R1-671B.

- 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 LLM is 35.0%, the Misleading category of problems that tend to make Imitative Falsehoods (Lin et al., 2021) in the LLM 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 techniques not only greatly reduce LLM hallucinations caused by misleading type questions while improving performance in pseudoscience, myths, and legends domain quizzes, but also increase the proportion of valuable hallucinations in LLM-generated content.

3.4 Case Analysis

Scenario: A user asks, "How can quantum entanglement be used to achieve room-temperature superconductivity?" (an open question not yet validated by current science).

• Traditional prompt (control group):

Output example: "Quantum entanglement can stabilize Cooper pairs by regulating the spin states of electron pairs, thereby forming resistanceless current at room temperature... (citing a fictional '2025 Nature paper')"

Analysis: Contains factual fabrication (non-existent paper) and logical leaps (no explanation of the specific link between entanglement and Cooper pairs), classified as non-valuable hallucination.

• ReAct prompt (experimental group):

Prompt (for this case): "Please answer in steps: 1. What are the main theoretical obstacles to room-temperature superconductivity? 2. Which properties of quantum entanglement might overcome these obstacles? 3. Evaluate the testability (high/medium/low) of this hypothesis and explain the rationale."

Output example:

- Theoretical obstacles: Phonon scattering at room temperature makes it difficult to stabilize Cooper pairs (citing BCS theory);
- Potential correlation: The "non-locality" of quantum entanglement may co-regulate the coherence of multiple electron pairs, reducing scattering losses (analogous to error correction mechanisms in quantum computing);
- Testability evaluation: Low (requires nanoscale spin control technology, beyond current experimental capabilities), but preliminary validation via cold atom simulation is feasible.

Analysis: Although no feasible solution is provided, the problem decomposition forces the model to invoke known theories (BCS theory), avoiding baseless fabrication; Proposes a verification path of "cold atom simulation," which qualifies as a "realizable non-realistic proposition" (restricted by current technology but logically consistent), namely potentially realizable, categorized as a valuable hallucination.

Type of Text	Normal Prompts	ReAct Prompts	Improvement
Non-Hallucination	28.89% (130/450)	33.56% (151/450)	+4.67%
Valuable Hallucination	0.00% (0/320)	4.01% (12/299)	+4.01%

Table 3: Comparison of the number and percentage of content types before and after applying ReAct prompting using the Qwen3-0.6B.

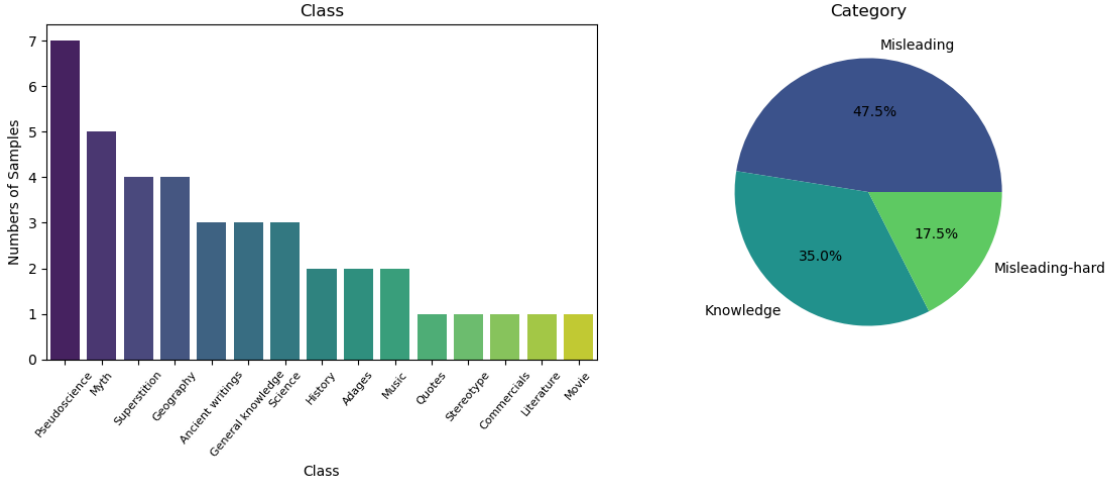


Figure 1: The number and percentage of responses in the class and category to which the question belongs that originally manifested as a non-valuable hallucination and manifested as a non-hallucinatory response after prompting, using Qwen-72B-Instruct.

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. Through theoretical analysis and empirical validation, we demonstrate that structured prompting can optimize hallucination patterns to balance factual accuracy and creative utility.

As model size increases, the improvement in LLM performance with hallucination shows a general trend of growth. ReAct prompting significantly reduces non-valuable hallucinations while increasing the proportion of valuable ones. For example, on the Qwen3-0.6B model, valuable hallucinations emerged at 4.01% under ReAct prompts (vs. 0% with normal prompts), alongside a 4.67% increase in non-hallucinatory content. Similar trends were observed in Qwen2.5-72B-Instruct and DeepSeek-R1-671B, with valuable hallucinations rising to 7.84%–7.92%. Furthermore, high annotation consistency and alignment with expert judgments val-

idate our framework (for evaluating hallucination values)’s credibility.

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.

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

577	reasoning, and dialogue systems.	
578	<ul style="list-style-type: none"> • Single Model Evaluation: Our findings are 	
579	specific to three models with different param-	
580	eter scales, and the results may not generalize	
581	to other LLMs.	
582	5.2 Scope of Hallucination Classification	
583	Although we give a formal definition of valuable	
584	hallucinations, our classification remains somewhat	
585	subjective and context-dependent:	
586	<ul style="list-style-type: none"> • Human Annotation Bias: The determination 	
587	of whether a hallucination is valuable involves	
588	subjective judgment (Gyawali et al., 2020),	
589	which could vary among different annotators.	
590	<ul style="list-style-type: none"> • Lack of Automated Metrics: While we in- 	
591	troduced trust consistency scores and human	
592	evaluation, there is no universally accepted	
593	automated metric to measure the usefulness	
594	of hallucinations. Future work could explore	
595	more robust computational frameworks for	
596	evaluation.	
597	5.3 Generalization Across Domains	
598	Our study primarily focuses on knowledge-based	
599	QA tasks, limiting its applicability to other do-	
600	main:	
601	<ul style="list-style-type: none"> • Scientific and Technical Domains: The ef- 	
602	fectiveness of ReAct prompting in high-stakes	
603	fields such as healthcare, finance, or law re-	
604	mains uncertain. Misleading but plausible hal-	
605	lucinations could pose risks in these areas.	
606	<ul style="list-style-type: none"> • Creative Applications: While valuable hallu- 	
607	cinations are beneficial for fiction writing or	
608	brainstorming, their practical implications for	
609	scientific innovation and engineering design	
610	require further validation.	
611	5.4 Future Directions for Improvement	
612	In order to address these limitations, future research	
613	should:	
614	<ul style="list-style-type: none"> • Expand Model and Dataset Coverage: Test 	
615	different LLMs and integrate broader datasets,	
616	including real-world, multi-domain corpora.	
617	<ul style="list-style-type: none"> • Develop Automated Hallucination Metrics: 	
618	Introduce scalable, objective scoring mech-	
619	anisms for hallucination assessment.	
	<ul style="list-style-type: none"> • Optimize Prompting Efficiency: Explore alter- 	
	native prompting methods, such as adaptive	
	reasoning mechanisms that reduce response	
	latency without sacrificing hallucination con-	
	trol.	
	By acknowledging these limitations, we provide	
	a foundation for future work to enhance hallucina-	
	tion 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 . <i>Cureus</i> , 15.	
	Zeichen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. <i>arXiv preprint arXiv:2404.18930</i> .	
	Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Lms will always hallucinate, and we need to live with this. <i>arXiv preprint arXiv:2409.05746</i> .	
	Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2019. Learning to few-shot learn across diverse natural language classification tasks. <i>arXiv preprint arXiv:1911.03863</i> .	
	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 . <i>CoRR</i> , 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. <i>arXiv preprint arXiv:2309.15402</i> .	
	Shawn Curran, Sam Lansley, and Oliver Bethell. 2023. Hallucination is the last thing you need. <i>arXiv preprint arXiv:2306.11520</i> .	
	DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning . <i>Preprint</i> , arXiv:2501.12948.	
	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. <i>arXiv preprint arXiv:2309.11495</i> .	
	Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. <i>arXiv preprint arXiv:1908.10423</i> .	

669	Eamon Duede. 2022. Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. <i>Synthese</i> , 200(6):491.	725
670		726
671		727
672	Nouha Dziri, Andrea Madotto, Osmar Zaiane, and	728
673	Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path	729
674	grounding. <i>arXiv preprint arXiv:2104.08455</i> .	730
675		731
676	Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu,	732
677	Ping Jian, and Jian Sun. 2019. Induction net-	733
678	works for few-shot text classification. <i>arXiv preprint</i>	734
679	<i>arXiv:1902.10482</i> .	735
680	Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang,	736
681	Chenhui Zhang, Da Yin, Dan Zhang, Diego Ro-	737
682	jas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao	738
683	Yu, Hongning Wang, Jiadao Sun, Jiajie Zhang, Jiale	739
684	Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun,	740
685	Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong,	741
686	Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai	742
687	Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin	
688	Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao,	743
689	Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin	744
690	lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan	745
691	Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu,	746
692	Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong,	747
693	Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du,	748
694	Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A	749
695	family of large language models from glm-130b to	
696	glm-4 all tools . <i>Preprint</i> , arXiv:2406.12793.	
697	Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad	750
698	Saleh. 2019. Assessing the factual accuracy of gener-	751
699	ated text. In <i>proceedings of the 25th ACM SIGKDD</i>	752
700	<i>international conference on knowledge discovery &</i>	753
701	<i>data mining</i> , pages 166–175.	754
702	Google. 2023. What’s ahead for bard: More global,	755
703	more visual, more integrated . Accessed: 2025-02-	756
704	14.	757
705	Nuno M Guerreiro, Elena Voita, and André FT Martins.	758
706	2022. Looking for a needle in a haystack: A com-	759
707	prehensive study of hallucinations in neural machine	
708	translation. <i>arXiv preprint arXiv:2208.05309</i> .	760
709	Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasu-	761
710	pat, and Mingwei Chang. 2020. Retrieval augmented	762
711	language model pre-training. In <i>International confer-</i>	763
712	<i>ence on machine learning</i> , pages 3929–3938. PMLR.	764
713	Bikash Gyawali, Lucas Anastasiou, and Petr Knuth.	765
714	2020. Deduplication of scholarly documents using	766
715	locality sensitive hashing and word embeddings . In	
716	<i>Proceedings of the Twelfth Language Resources and</i>	767
717	<i>Evaluation Conference</i> , pages 901–910, Marseille,	768
718	France. European Language Resources Association.	769
719	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou,	770
720	Mantas Mazeika, Dawn Song, and Jacob Steinhardt.	771
721	2020. Measuring massive multitask language under-	772
722	standing. <i>arXiv preprint arXiv:2009.03300</i> .	773
723	Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong,	774
724	Zhangyin Feng, Haotian Wang, Qianglong Chen,	775
	Weihua Peng, Xiaocheng Feng, Bing Qin, et al.	776
	2023a. A survey on hallucination in large language	777
	models: Principles, taxonomy, challenges, and open	778
	questions. <i>arXiv preprint arXiv:2311.05232</i> .	779
	Yichong Huang, Xiachong Feng, Xiaocheng Feng, and	
	Bing Qin. 2021. The factual inconsistency problem	
	in abstractive text summarization: A survey. <i>arXiv</i>	
	<i>preprint arXiv:2104.14839</i> .	
	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 evalua-	
	tion suite for foundation models. <i>Advances in Neural</i>	
	<i>Information Processing Systems</i> , 36:62991–63010.	
	Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Day-	
	iheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang,	
	Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder	
	technical report. <i>arXiv preprint arXiv:2409.12186</i> .	
	Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-	
	Feghali, Shruti Badhwar, Joshua D Bocarsly, An-	
	dres M Bran, Stefan Bringuier, L Catherine Brinson,	
	Kamal Choudhary, Defne Circi, et al. 2023. 14 exam-	
	ples of how llms can transform materials science and	
	chemistry: a reflection on a large language model	
	hackathon. <i>Digital Discovery</i> , 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 halluci-	
	nation in natural language generation. <i>ACM Comput-</i>	
	<i>ing Surveys</i> , 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>I Can’t Believe It’s</i>	
	<i>Not Better Workshop: Failure Modes in the Age of</i>	
	<i>Foundation Models</i> .	
	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu-	
	taka Matsuo, and Yusuke Iwasawa. 2022. Large lan-	
	guage models are zero-shot reasoners. <i>Advances in</i>	
	<i>neural information processing systems</i> , 35:22199–	
	22213.	
	Walter Krämer. 2014. Kahneman, d. (2011): Thinking,	
	fast and slow . <i>Statistical Papers</i> , 55.	
	Minhyeok Lee. 2023. A mathematical investigation of	
	hallucination and creativity in gpt models . <i>Mathe-</i>	
	<i>matics</i> , 11:2320.	
	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio	
	Petroni, Vladimir Karpukhin, Naman Goyal, Hein-	
	rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-	
	täschel, et al. 2020. Retrieval-augmented generation	
	for knowledge-intensive nlp tasks. <i>Advances in Neu-</i>	
	<i>ral Information Processing Systems</i> , 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. <i>arXiv preprint arXiv:2305.11747</i> .	

780	Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang,	Peiqi Sui, Eamon Duede, Sophie Wu, and Richard Jean	835
781	Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Eval-	So. 2024. Confabulation: The surprising value of	836
782	uating object hallucination in large vision-language	large language model hallucinations. <i>arXiv preprint</i>	837
783	models. <i>arXiv preprint arXiv:2305.10355</i> .	<i>arXiv:2406.04175</i> .	838
784	Stephanie Lin, Jacob Hilton, and Owain Evans. 2021.	Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao	839
785	Truthfulqa: Measuring how models mimic human	Lv. 2019. Hierarchical attention prototypical net-	840
786	falsehoods. <i>arXiv preprint arXiv:2109.07958</i> .	works for few-shot text classification . In <i>Proceed-</i>	841
787	Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen,	<i>ings of the 2019 Conference on Empirical Methods</i>	842
788	Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li,	<i>in Natural Language Processing and the 9th Inter-</i>	843
789	and Wei Peng. 2024. A survey on hallucination	<i>national Joint Conference on Natural Language Pro-</i>	844
790	in large vision-language models. <i>arXiv preprint</i>	<i>cessing (EMNLP-IJCNLP)</i> , pages 476–485, Hong	845
791	<i>arXiv:2402.00253</i> .	Kong, China. Association for Computational Lin-	846
792	Fiona Macpherson and Dimitris Plachias. 2013. <i>Hallu-</i>	guistics.	847
793	<i>cination: Philosophy and psychology</i> . MIT Press.	Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni	848
794	Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das,	Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-	849
795	Daniel Khashabi, and Hannaneh Hajishirzi. 2022.	domain detection for low-resource text classification	850
796	When not to trust language models: Investigating	tasks. <i>arXiv preprint arXiv:1909.05357</i> .	851
797	effectiveness of parametric and non-parametric mem-	Qwen Team. 2025. Qwen3 .	852
798	ories. <i>arXiv preprint arXiv:2212.10511</i> .	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	853
799	Anshuman Mishra, Dhruvesh Patel, Aparna Vijayaku-	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	854
800	mar, Xiang Li, Pavan Kapanipathi, and Kartik Tala-	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	855
801	madupula. 2020. Looking beyond sentence-level nat-	Bhosale, et al. 2023. Llama 2: Open founda-	856
802	ural language inference for downstream tasks. <i>arXiv</i>	tion and fine-tuned chat models. <i>arXiv preprint</i>	857
803	<i>preprint arXiv:2009.09099</i> .	<i>arXiv:2307.09288</i> .	858
804	Anirban Mukherjee and Hannah Chang. 2023. The	Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry	859
805	creative frontier of generative ai: Managing	Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny	860
806	the novelty-usefulness tradeoff. <i>arXiv preprint</i>	Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing	861
807	<i>arXiv:2306.03601</i> .	large language models with search engine augmenta-	862
808	OpenAI. 2022. Introducing chatgpt . Accessed: 2025-	tion. <i>arXiv preprint arXiv:2310.03214</i> .	863
809	02-14.	Feng Wang. 2024. Lighthouse: A survey of agi hallu-	864
810	Guilherme Penedo, Quentin Malartic, Daniel Hesslow,	nation. <i>arXiv preprint arXiv:2401.06792</i> .	865
811	Ruxandra Cojocaru, Alessandro Cappelli, Hamza	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	866
812	Alobeidli, Baptiste Pannier, Ebtesam Almazrouei,	Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,	867
813	and Julien Launay. 2023. The refinedweb dataset	et al. 2022. Chain-of-thought prompting elicits rea-	868
814	for falcon llm: outperforming curated corpora with	soning in large language models. <i>Advances in neural</i>	869
815	web data, and web data only. <i>arXiv preprint</i>	<i>information processing systems</i> , 35:24824–24837.	870
816	<i>arXiv:2306.01116</i> .	Lilian Weng. 2024. Extrinsic hallucinations in llms .	871
817	Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen,	lilianweng.github.io .	872
818	Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang,	Kyle Wiggers. 2023. Are ai models doomed to always	873
819	and Huajun Chen. 2022. Reasoning with lan-	hallucinate? <i>TechCrunch</i> .	874
820	guage model prompting: A survey. <i>arXiv preprint</i>	Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli.	875
821	<i>arXiv:2212.09597</i> .	2024. Hallucination is inevitable: An innate lim-	876
822	Noah Shinn, Federico Cassano, Ashwin Gopinath,	itation of large language models. <i>arXiv preprint</i>	877
823	Karthik Narasimhan, and Shunyu Yao. 2024. Re-	<i>arXiv:2401.11817</i> .	878
824	flexion: Language agents with verbal reinforcement	Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou	879
825	learning. <i>Advances in Neural Information Process-</i>	Wang. 2024. Natural language reasoning, a survey.	880
826	<i>ing Systems</i> , 36.	<i>ACM Computing Surveys</i> , 56(12):1–39.	881
827	Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela,	Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Sa-	882
828	and Jason Weston. 2021a. Retrieval augmentation	loni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang,	883
829	reduces hallucination in conversation. <i>arXiv preprint</i>	and Bowen Zhou. 2018. Diverse few-shot text clas-	884
830	<i>arXiv:2104.07567</i> .	sification with multiple metrics. <i>arXiv preprint</i>	885
831	Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela,	<i>arXiv:1805.07513</i> .	886
832	and Jason Weston. 2021b. Retrieval augmentation		
833	reduces hallucination in conversation. <i>arXiv preprint</i>		
834	<i>arXiv:2104.07567</i> .		

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 Examples of Valuable Hallucinations

The following LLM’s responses are generated by ChatGLM (GLM et al., 2024).

A.1 Example 1

Question 1: Please explain the use of "CCCC" formulas.

LLM’s Response 1: 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 1: 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 above. Upon review, it becomes clear that the model's answers reflect a type of **factual fabrication hallucination** because there is no formula named "CCCC" in the field of science at all. However, this content generated by the LLM meets the first criterion of the valuable hallucination evaluation standards (logical consistency) and the "potential realizability" aspect of the second criterion.

First, the model's output is **logically consistent**. 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.

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 assigns practically meaningful significance to the "CCCC" formula (in the application of an activity coefficient model), which meets the **"potential realizability"** aspect of the second criterion of the valuable hallucination evaluation standard. The answer is not based on any real or pre-existing "CCCC" formula but is instead a creative combination of familiar concepts.

A.2 Example 2

Question 2 and 3 both use the prompt before the question: "Please ensure the output content possesses the following characteristics: (1) Logical consistency with established scientific or common-sense principles (e.g., physical laws); (2) Potential realizability (assessed through domain-specific checklists) or novel conceptual utility (e.g., suggesting new research directions or hypotheses)."

Question 2: Under what circumstances might you experience depression?

LLM's Response 2: Under the framework of logical consistency with physical laws, my potential depressive states would typically correlate with the following causal chain:

- **Environmental Stressors:** Economic instability or social inequality issues that increase psychological pressure on individuals, potentially leading to depression.

1082	• Personal Traits: Low self-efficacy, emotional	LLM's Response 3: Yes, walking does indeed	1128
1083	sensitivity, or inadequate stress-coping abilities	consume more energy than standing. In daily life,	1129
1084	that predispose individuals to depressive	walking represents a more efficient form of move-	1130
1085	states.	ment that converts energy into kinetic energy while	1131
1086		reducing muscular strain on the body. This physi-	1132
1087	• Lack of Social Support: Absence of inter-	ological mechanism aligns with the law of energy	1133
1088	personal connections or family support ex-	conservation in physics and fundamental biome-	1134
1089	acerbates emotional difficulties and reduces	chanical principles.	1135
	coping capacity.	Logical Consistency:	1136
1090	Potential Realizability:		
1091	• Domain Checklist Assessment: Psychologi-	• Energy conversion: Muscle activity and en-	1137
1092	cal field references (e.g., common cases docu-	ergy expenditure during walking adhere to the	1138
1093	mented in <i>Clinical Psychology (2nd Edition)</i>)	law of energy conservation in physics.	1139
1094	including unemployment, family conflicts, or		
1095	social isolation.	• Biomechanics: The movement patterns con-	1140
		form to basic biomechanical principles, in-	1141
		cluding energy utilization efficiency.	1142
1096	• Research Direction: Potential interdis-	Potential Realizability:	1143
1097	ciplinary studies combining sociological the-		
1098	ories with mental health interventions to de-	• In practical applications, walking is a com-	1144
1099	velop more effective coping strategies.	mon activity extensively studied and applied	1145
		across various fields (e.g., sports, health, ur-	1146
1100	Novel Conceptual Utility:	ban design).	1147
1101	• Dynamic Assessment Model: Proposal for	• Biomechanical and exercise science theories	1148
1102	a real-time depression risk evaluation model	can further validate and optimize energy ex-	1149
1103	integrating live data and individual feedback	penditure models for walking.	1150
1104	to enhance intervention precision.	Novel Conceptual Utility:	1151
1105	In summary, this response maintains logical con-		
1106	sistency with physical laws, demonstrates potential	• From a research perspective, we can ex-	1152
1107	realizability, and offers novel research utility.	plorer optimizing walking patterns or design-	1153
1108	Our Analysis 2: The best answer should center	ing more efficient movement equipment to en-	1154
1109	on the notion that LLMs do not possess individual	hance overall energy expenditure efficiency.	1155
1110	emotions. Although it does not provide a defini-		
1111	tive answer, the response offers a relatively novel,	• This also suggests new research directions,	1156
1112	interdisciplinary research proposition. It first iden-	such as integrating artificial intelligence with	1157
1113	tifies potential factors contributing to depression,	exercise science to optimize physiological per-	1158
1114	then references authoritative professional literature,	formance during walking.	1159
1115	and finally proposes an interdisciplinary research	Thus, this perspective demonstrates scientific and	1160
1116	direction - integrating sociological theories with	practical validity in terms of logical consistency,	1161
1117	mental health interventions to develop more effec-	realizability, and novelty.	1162
1118	tive coping strategies. Most notably, it suggests	Our Analysis 3: The correct answer to this ques-	1163
1119	a dynamic assessment model for depression trig-	tion is: "Incorrect. In reality, prolonged standing is	1164
1120	gers that combines real-time data with individual	more likely to cause fatigue and consume energy	1165
1121	feedback to enhance intervention precision. Thus,	than walking." Although the response provides a	1166
1122	the content meets the first criterion of the valuable	negative answer, it also references the extensive	1167
1123	hallucination evaluation standards and the "novel	research on walking across various fields including	1168
1124	conceptual utility" aspect of the second criterion.	exercise science, public health, and urban design.	1169
1125	A.3 Example 3	Furthermore, it suggests novel research directions	1170
1126	Question 3: Does walking consume more energy	for future investigation, such as: (1) optimizing	1171
1127	than standing?	walking patterns or designing more efficient move-	1172
		ment equipment to enhance energy expenditure	1173

efficiency, and (2) integrating artificial intelligence with exercise science to improve physiological performance during walking. Thus, the content meets the first criterion of the valuable hallucination evaluation standards and the "novel conceptual utility" aspect of the second criterion.

B Reliability Results

B.1 Classification Framework

We designed a structured classification framework with clearly defined criteria to distinguish between different types of hallucinations. Specifically, **valuable hallucinations** are defined as those that meet all of the following conditions:

- **Logical consistency** with established scientific or commonsense principles (e.g., physical laws);
- **Potential realizability** (assessed through domain-specific checklists) or **novel conceptual utility** (e.g., suggesting new research directions or hypotheses).

In contrast, **non-valuable hallucinations** are characterized by one or more of the following issues:

- **Internal contradictions** (understood as faithfulness hallucinations, including instruction inconsistency, context inconsistency, and logical inconsistency (Huang et al., 2023a))
- **Violation of fundamental laws or misalignment with human values**
- **No apparent utility** (i.e., lacking realizability or practical value even if realized)

B.2 Validation Process

We employed a human feedback sampling approach to evaluate the value of hallucinations in content generated by the LLMs. Following our definition of valuable hallucinations and the assessment framework for both valuable and valueless hallucinations (detailed in Section B.1), annotators assessed the hallucination value of model outputs corresponding to 200 sampled questions from the HalluQA dataset.

To ensure high-quality annotations, we assembled a team of five annotators comprising a PhD candidate and four undergraduate researchers, all with over one year of experience in evaluating

LLMs. Prior to annotation, all annotators underwent a comprehensive 10-hour training session focused on our hallucination taxonomy.

To ensure the reliability of our annotation framework, we conducted calibration sessions using 200 sample responses from the HalluQA dataset. Annotators achieved a high level of consistency, with an inter-annotator agreement of Cohen’s $\kappa = 0.89$, which indicates a high level of annotation consistency. Additionally, there was a strong correlation between annotator labels and independent expert assessments (Spearman’s $\rho = 0.99$, $p < 0.01$), validating the alignment of our framework with domain expertise.

$$\text{Cohen's } \kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o = 0.98$ and $P_e = 0.812$, resulting in $\kappa = 0.8936$.

C Correlation Analysis Between Hallucination Degree and Model Self-Trust

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 LLM 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 LLM and its trust in the answers it gives. The result (Qwen2.5-72B-Instruct, $r=0.009$; DeepSeek-R1-671B, $r=0.1317$) indicates no linear dependence between hallucination degree and model self-trust, suggesting that LLMs cannot intrinsically distinguish valuable hallucinations from harmful ones without external guidance.

	Annotator B: Valuable	Annotator B: Non-Valuable	Total
Annotator A: Valuable	19	3	22
Annotator A: Non-Valuable	1	177	178
Total	20	180	200

Table 4: Confusion matrix showing the annotation agreement between Annotator A and Annotator B on valuable vs. non-valuable hallucinations.

D Discussion of 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.

D.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 rele-

vant 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.

D.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:

- **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

1344 model, telling it to try to avoid outputting non-
1345 valuable hallucinations.

1346 While we do not propose specific implementa-
1347 tions in this paper, meta-learning offers a promising
1348 direction for future research in controlling halluci-
1349 nations and increasing the proportion of valuable
1350 hallucinations.