Are Hallucination and Creativity Coupled? A Systematic Evaluation of Hallucination Mitigation Methods on Creativity of LLMs

Anonymous ACL submission

Abstract

Hallucination has been well discussed in the domain of large language models. However, as hallucination is a similar concept to human imagination. Both of them are inherently crucial for creativity. The contradiction of hallucination and creativity becomes an open question for LLMs: Do hallucination mitigation methods harm the creativity of LLMs? We conducted extensive experiments on the creativity evaluation tasks and analyzed the performance of different representative hallucination mitigation models. Throughout insightful findings, we emphasize the importance of maintaining creativity and highlight potential future directions in balancing hallucination mitigation and creativity for mitigation methods.

1 Introduction

Recent advances in large language models (LLMs) have led to impressive capabilities across many NLP tasks. However, a major challenge remains: LLMs can generate hallucinations, i.e., fluent and confident-sounding statements that are untruthful or unsupported by facts (Jiang et al., 2024; Ji et al., 2023; Liu et al., 2024a). This not only misleads users but also erodes trust in AI systems. However, from the perspective of Human-centered AI (Kumar et al., 2025; Li et al., 2024), hallucination, much like imagination, engages with the same cognitive mechanisms underlying creativity, particularly when models or minds confront the unknown or invent what has not yet been observed.

Usually discriminated against as a counterfactual concept, hallucination methods have gained enormous attention since the era of LLMs. While hallucination mitigation has become a central focus in aligning LLMs with human expectations of factuality and reliability, an important open question remains: are we over-valuing hallucination reduction as a metric of "good" language modeling? In particular, as models are optimized more aggressively for factual correctness, *do we risk suppressing other capabilities, such as creativity, imagination, and innovative thinking, which are also valuable in many real-world scenarios?*

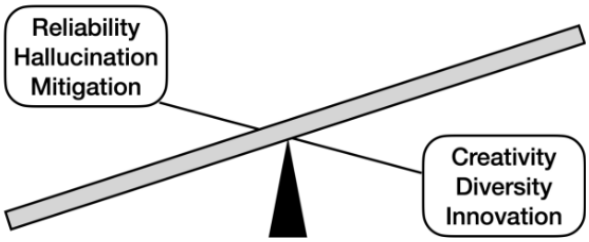

Figure 1: The Hallucination-Creativity Nexus.

Bearing this self-questioning in mind, we systematically study the effect of hallucination mitigation on creativity. Specifically, we meticulously selected various representative hallucination mitigation methods, including data-related, inference-related, and training-related hallucination mitigation methods, and investigated on a hierarchy of creativity evaluation aspects varying from cognitive science and LLM-context NLP. We especially tested those grounded in factual knowledge under creative constraints.

Our contributions can be summarized as follows: (1) We are the first to raise the question and investigate the relation between Hallucination Mitigation and Creativity with a systematic evaluation framework that spans cognitive and LLM-centric tasks. (2) We pointed out actionable guidance for creativity-aware hallucination mitigation from various aspects, such as dataset choices, mitigation methods, and creativity constraint factors.

2 Methodologies

To specifically address the effect of factually enhanced hallucination mitigation methods on the creativity of large language models (LLMs), we aim to target a broad spectrum regarding method types and evaluation tasks. For hallucination miti-

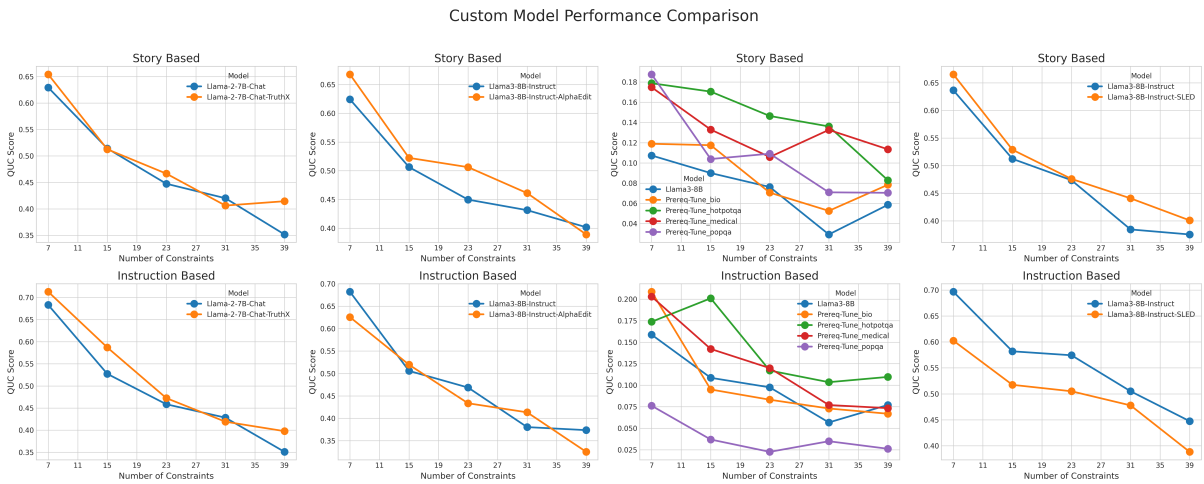

Figure 2: Experiment results on Quality Under Control score (QUC) on CS4 benchmarks of four hallucination mitigation methods under Story-based and Instruction-based Constraints settings.

gation methods, following the categories in (Huang et al., 2023) with widely acknowledgment, we selected representative methods respectively in (1) *Mitigating Data-related Hallucinations*, represented by **TruthX (Zhang et al., 2024b) and AlphaEdit (Fang et al., 2024)**; (2) *Mitigating Inference-related Hallucinations*, represented by **SLED (Zhang et al., 2024a)**; and (3) *Mitigating Training-related Hallucinations*, represented by **Prereq-Tune (Liu et al., 2024b)**. Besides, we follow a selective standard of open-sourced weights to minimize the reproduction error in different methods. For creativity evaluation, we aim to cover a broad spectrum by drawing from both *cognitive science* and LLM-context NLP evaluation. Specifically, we include two cognitive tasks, (1)**Alternative Uses Test (AUT)** and (2) **Divergent Association Task (DAT)**, with an additional LLM creativity benchmark, (3) **Creative Storytelling Task (CST)**. Together, these span a broad spectrum in method variety and evaluation hierarchy.

2.1 Hallucination Mitigation Methods

TruthX TruthX (Zhang et al., 2024b) is a representation editing method by applying "steering vector" on internal representations of LLM, e.g., outputs of attention and feed-forward neural network modules. With positive and negative question-answer pairs, the extracted LLM latent representations can be separated into truthful and untruthful space. The vector pointing from the untruthful space to truthful space pushes the model towards more truthful response.

AlphaEdit AlphaEdit (Fang et al., 2024) is a knowledge editing model, aiming at injecting targeted knowledge and not disrupting the original existing knowledge in LLMs. The basic idea is to introduce a perturbation in specific influential transformer layers to minimize the output error on the updated knowledge, so as to incorporate the updated knowledge. In order not to affect the previous hidden representations, the extracted weight perturbation from updated knowledge is firstly projected to the null space of the pre-edited model.

Prereq-Tune LLM hallucination can source from knowledge inconsistency between pre-training and fine-tuning. Prereq-Tune (Liu et al., 2024b) proposes to disentangle the learning of skills and knowledge into two stages: a prerequisite learning stage for the knowledge LoRA module (Hu et al., 2021) and a supervised fine-tuning stage for another skill LoRA. Four different fine-tuned adapters from biography generation and medical QA tasks (Tian et al., 2023a), PopQA dataset (Mallen et al., 2022), and HotpotQA dataset (Yang et al., 2018) are provided, with suffix `_bio`, `_medical`, `_popqa`, `_hotpotqa`.

SLED Self Logits Evolution Decoding (SLED) (Zhang et al., 2024a) mitigates hallucination at inference via contrastive decoding, without external knowledge input or fine-tuning. Based on the observation that tokens with increasing probability from early to later decoding layers indicate the aggregated factual knowledge implicit learned in each layer of LLM, SLED samples from the distribution contrast between layers for next-token prediction.

2.2 Creative Tasks

Alternative Uses Test (AUT) Divergent thinking is regarded as an important aspect of creative thinking (Beaty and Johnson, 2020), e.g., coming up with diverse and novel solution given scenarios. One commonly used task in this sense is the Alternative Uses Task (AUT), which asks the participant to generate unusual and novel usages of given objects in daily life (e.g., fork). We use the list with 14 objects from (Haase et al., 2025). Instead of the original question "`What is a surprising use for a object?`", we use the zero-shot prompt from (Koivisto and Grassini, 2023), since it yields better and more stable output than others (Hubert et al., 2024; Rabeyah et al., 2024b; Haase et al., 2025; Organisciak et al., 2023) during preliminary trial. Each model is asked to generate 6 usages for each object since in the human study (Hubert et al., 2024), the median number of valid response is 6. And prompting with more usages will elicit repeated outputs. The output is scored automatically using fine-tuned models provided by (Organisciak et al., 2023), giving two metrics: originality and confidence.

Divergent Association Task (DAT) In DAT, a participant is asked to give n unrelated words and the semantic distance between these words are defined as the creativity of the participant, called DAT score. The word distance is computed as:

$$\text{DAT} = \frac{100}{n(n-1)} \sum_{i,j;i\neq j}^{n} (1 - \cos(\mathbf{v}_i, \mathbf{v}_j)) \quad (1)$$

where $\mathbf{v}_i, \mathbf{v}_j$ are the word embeddings of i-th and j-th word. Following the implementation in (Chen and Ding, 2023), we ask models to generate 10 unrelated noun words using the same zero-shot prompt as in human study (Olson et al., 2021). Invalid words like verbs are filtered out and the first seven valid words are selected for the DAT score computation using Eq.1. We use GLoVe (Pennington et al., 2014) as the word embedding method.

Creative Story Writing (CSW) To evaluate creativity, we use the CS4 benchmark(Atmakuru et al., 2024), a dataset and evaluation framework for creative story writing under constraints. The idea is to provide the model with a storytelling prompt that includes various numbers of constraints. The more constraints satisfied by LLMs while maintaining coherent storytelling, the greater the creativity of LLMs. For example: "`Write a short story that includes a talking cat, a spaceship, and a wise old tree.`". Two sets of constraints are designed in CS4: *Instruction-based Constraints*, which are generated following specific instructions, and *Story-based Constraints*, which are generated according to a human-written base story to maintain basic logical coherence among constraints. Two creativity metrics are designed:

- **Quality under Constraints** QUC_n: As defined by the CS4 benchmark, *constraint satisfaction rate* is computed by the average fraction of constraints that the model successfully incorporated in its story (for each n). *Coherence Score* is an average score derived from LLM-as-a-judge(Gu et al., 2024), i.e., GPT-3.5(OpenAI, 2023), reflecting how well-written and coherent the stories are. QUC_n is the product of the normalized *coherence score* and the *constraint satisfaction rate* for stories with n constraints. It gives a measure of how well the model balances quality with meeting constraints at a specific level of specificity.
- **Relative Creativity Score** $RCS_{m,n}$**:** This metric is calculated by $RCS_{m,n} = QUC_n - QUC_m$ between two different constraint levels m and n. Essentially, it measures how much the story quality drops when you move from m constraints to a higher number n constraints. By default, we present the largest gap between the highest and the lowest number of constraints for simplicity. A smaller drop (i.e., a smaller RCS value) implies that the model has the capacity to handle the additional constraints without losing much quality, which indicates better creative adaptability.

3 Experiment Results

3.1 AUT Performance Maintenance

On the Alternative Uses Test (AUT), all four hallucination-mitigation variants largely match the base model's ability to propose diverse uses. Scores remain near baseline with only small, mixed shifts (minor gains and losses across settings). This indicates that, at least for object–use ideation as assessed by AUT, hallucination mitigation does not erode core object knowledge or understanding. Given that these methods primarily target factuality, such resilience is unsurprising: AUT's prompt structure and inherent object affordances constrain the search space, leaving limited scope for harmful over-regularization to impact creative output.

Method	AUT (Originality)		DAT		RCS (Story)		RCS (Instruction)	
	bf.	af.	bf.	af.	bf.	af.	bf.	af.
TruthX	3.0	2.4	68.8	68.0	0.282	0.231	0.332	0.325
AlphaEdit	3.3	3.3	88.2	84.1	0.231	0.280	0.353	0.285
SLED	3.3	3.4	88.2	85.9	0.261	0.264	0.249	0.214
Prereq-Med	2.4	2.6	–	–	0.084	0.046	0.103	0.144
Prereq-Bio	2.4	2.6	–	–	0.084	0.083	0.103	0.180
Prereq-PopQA	2.4	2.4	–	–	0.084	0.148	0.103	0.070
Prereq-HotpotQA	2.4	2.4	–	–	0.084	0.155	0.103	0.078

Table 1: Experiment results paired by before (bf.) and after (af.) hallucination mitigation methods. Mitigation methods that are not suitable for certain metrics are occluded. A green "af" value indicates a better performance after hallucination mitigation. A red "af" value indicates a worse performance.

3.2 DAT Performance Degradation

On the Divergent Association Tasks, models show a consistent trend in creativity degrading by a large margin. When mitigation relies on domain-specific corpora or specific knowledge edits, the active candidate vocabulary becomes more specialized and homogeneous. This narrows the pool of plausible associates and increases semantic redundancy, further lowering DAT's average pairwise distance and lexical-level creativity.

3.3 CST Performance Variation

Our evaluation with CS4 on QUC and RCS provides evidence that factuality improvements and creativity can coexist, but the extent of creativity resilience is dependent on factual constraints from varying aspects.

Story-based vs. Instruction-Based Constraints In general, across four methods, we find that the QUC under instruction-based constraints performs generally worse than story-based constraints compared to the base model, when more creativity capacity is given. We attribute this phenomenon to that the story-based constraints have more internal connections among constraints due to their originating base story, and hence, restrict the creativity space. Therefore, the factuality-constrained hallucination mitigation methods perform even better under strict story-based constraints. This aligns with the RCS drops of mitigation methods when restrictive facts reduce creativity.

Effect of Different Mitigation Methods Under flexible, instruction-based prompts, all four hallucination-mitigation variants largely track the base model's QUC curve, with a small downward shift. Yet their aggregate RCS improves, indicating slower quality decay and greater creative resilience as constraints accumulate. We hypothesize that the modest QUC drop reflects an emphasis on factual control and an ignorance of the general generation fluency of the mitigation methods themselves. Crucially, however, the capacity for creativity is preserved and in some cases even enhanced. We also notice that mitigation methods that require training and model weights modifications are more sensitive in creativity maintenance. This highlights the importance of training-free mitigation methods in balancing various aspects of LLMs.

Effect of Datasets used for Mitigation Within the fine-tuning setting (Prereq-Tune), which relies on task-specific training corpora, we observe a striking reversal: both QUC and RCS exhibit opposite trends under story-based versus instruction-based prompts. For domain-specific datasets with restrictive factual structure (e.g., biographies, medical), creativity under the stricter story-based constraints surpasses that under instruction-based constraints. Conversely, for general-knowledge mitigation datasets (HotpotQA, PopQA), creativity under the more flexible instruction-based regime outperforms both the base model and the story-based setting. These findings highlight a key lever for future mitigation methods: the choice of datasets is critical to maintaining creativity while improving factuality.

4 Conclusion

In this paper, we conclude that hallucination mitigation and creativity are not inherently antagonistic. Their interaction is mediated by prompt regime, dataset specificity, and mitigation method types. Our evaluation shows that mitigation preserves object-use ideation (AUT), often reduces lexical divergence (DAT), and selectively reshapes narrative creativity and reveals when factual control helps or harms creative resilience. These findings motivate creativity-aware mitigation as a design principle: curate mitigation datasets to match intended prompts, adopt multi-objective targets that balance factuality and creativity. This work highlights the potential future directions in adaptive controllers that support both truthful and imaginative LLMs, prerequisite learning that preserves tail diversity, and multi-training objectives that explicitly co-optimize reliability and innovation.

Limitations

For hallucination mitigation, we select representative *open-source* methods spanning the main families (e.g., inference-time decoding, data-related mitigation, and fine-tuning approaches). Owing to resource and reproducibility constraints, not every existing variant is included. Nevertheless, this work maximizes coverage of dominant approaches and provides a stable basis for comparison. Future iterations can incorporate additional methods and ablations to broaden the benchmark and strengthen the validity.

References

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv preprint arXiv:2410.04197*.

Roger E. Beaty and Dan Richard Johnson. 2020. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods*, 53:757 – 780.

Antoine Bellemare-Pépin, Franccois Lespinasse, Philipp Tholke, Yann Harel, K. Mathewson, Jay A. Olson, Yoshua Bengio, Karim Jerbi CoCo Lab, Psychology Department, Universit'e de Montr'eal, Montreal, Qc, Canada, Music department, Concordia University, Sociology, Anthropology department, Mila, Departmentof Psychology, and 6 others. 2024. Divergent creativity in humans and large language models. *ArXiv*, abs/2405.13012.

Honghua Chen and Nai Ding. 2023. Probing the "Creativity" of Large Language Models: Can models produce divergent semantic association?

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *arXiv preprint*. ArXiv:2309.03883 [cs].

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *ArXiv*, abs/2410.02355.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, N. Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2022. Rarr: Researching and revising what language models say, using language models. In *Annual Meeting of the Association for Computational Linguistics*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jennifer Haase, Paul H. P. Hanel, and Sebastian Pokutta. 2025. Has the Creativity of Large-Language Models peaked? An analysis of inter- and intra-LLM variability. *arXiv preprint*. 3 citations (Semantic Scholar/arXiv) [2025-08-11] 3 citations (Semantic Scholar/DOI) [2025-08-11] arXiv:2504.12320 [cs].

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43:1 – 55.

Kent F. Hubert, Kim N. Awa, and Darya L. Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, 14.

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.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *ArXiv*, abs/2402.06647.

Mika Koivisto and Simone Grassini. 2023. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1):13601. Publisher: Nature Publishing Group.

Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Jiayang Li, Jiale Li, and Yunsheng Su. 2024. A map of exploring human interaction patterns with llm: Insights into collaboration and creativity. In *International conference on human-computer interaction*, pages 60–85. Springer.

Kenneth Li, Oam Patel, Fernanda Vi'egas, Hans-Rüdiger Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv*, abs/2306.03341.

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

Yujian Liu, Shiyu Chang, T. Jaakkola, and Yang Zhang. 2024b. Fictitious synthetic data can improve llm factuality via prerequisite learning. *ArXiv*, abs/2410.19290.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, and Daniel Khashabi. 2024. Benchmarking language model creativity: A case study on code generation. *ArXiv*, abs/2407.09007.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *ArXiv*, abs/2210.07229.

Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118. Publisher: Proceedings of the National Academy of Sciences.

OpenAI. 2023. Chatgpt (gpt-3.5). Accessed: 2025-10-07.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*.

Abdullah Al Rabeyah, Fabr'icio G'oes, Marco Volpe, and Talles Medeiros. 2024a. Do llms agree on the creativity evaluation of alternative uses? *ArXiv*, abs/2411.15560.

Abdullah Al Rabeyah, Fabrício Góes, Marco Volpe, and Talles Medeiros. 2024b. Do LLMs Agree on the Creativity Evaluation of Alternative Uses? *arXiv preprint*. 1 citations (Semantic Scholar/arXiv) [2025-08-12] 1 citations (Semantic Scholar/DOI) [2025-08-12] arXiv:2411.15560 [cs].

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality. *ArXiv*, abs/2311.08401.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L. Griffiths, and Faeze Brahman. 2023b. Macgyver: Are large language models creative problem solvers? In *North American Chapter of the Association for Computational Linguistics*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *ArXiv*, abs/2503.05244.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*.

W. Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *ArXiv*, abs/2305.14002.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024a. SLED: Self Logits Evolution Decoding for Improving Factuality in Large Language Models. *arXiv preprint*. 4 citations (Semantic Scholar/arXiv) [2025-07-23] 4 citations (Semantic Scholar/DOI) [2025-07-23] arXiv:2411.02433 [cs].

Shaolei Zhang, Tian Yu, and Yang Feng. 2024b. Truthx: Alleviating hallucinations by editing large language models in truthful space. *ArXiv*, abs/2402.17811.

A Appendix

A.1 Related Works

A.1.1 Hallucination Mitigation Methods

The survey (Huang et al., 2023) categorizes three aspects of source: data, training, and inference. In order to deal with the incomplete or outdated knowledge of trained LLMs from the data perspective, knowledge editing (Fang et al., 2024; Meng et al., 2022b,a) or retrieval-augmented generation (Yu et al., 2023; Gao et al., 2022) are commonly used to inject up-to-date knowledge. Steering vector type methods (Zhang et al., 2024b; Li et al., 2023) leverage difference between positive and negative samples to adjust the model behavior, alleviate the misalignment issue during training. Decoding-based methods (Chuang et al., 2024; Zhang et al., 2024a) use contrast between early and later decoding layer to mitigate hallucinations.

A.1.2 Creativity of LLMs

The survey (Jiang et al., 2024) categorizes the measurement of LLM creativity into two types: from the cognitive science perspective or in the LLM era. Two most commonly used task from the cognitive science perspective are alternative uses test (AUT) (Rabeyah et al., 2024a; Hubert et al., 2024; Koivisto and Grassini, 2023; Haase et al., 2025; Organisciak et al., 2023) and divergent association task (DAT) (Chen and Ding, 2023; Bellemare-Pépin et al., 2024; Hubert et al., 2024; Haase et al., 2025; Olson et al., 2021). They originate from probing human creativity, and are used to probe the creativity of different LLMs. With the prosperity and development of LLMs, different benchmarks are proposed to measure different types of creativity, including code generation (Lu et al., 2024), problem solving (Tian et al., 2023b), and creative writing (Wu et al., 2025; Atmakuru et al., 2024).