Learning Auxiliary Tasks Improves Reference-Free Hallucination Detection in Open-Domain Long-Form Generation

Anonymous ACL submission

Abstract

Hallucination, the generation of factually incorrect information, remains a significant challenge for large language models (LLMs), especially in open-domain long-form generation. Existing approaches for detecting hallucination in long-form tasks either focus on limited domains or rely heavily on external fact-checking tools, which may not always be available.

In this work, we systematically investigate reference-free hallucination detection in open-domain long-form responses. Our findings reveal that internal states (*e.g.,* model's output probability and entropy) alone are insufficient for reliably (*i.e.,* better than random guessing) distinguishing between factual and hallucinated content. To enhance detection, we explore various existing approaches, including prompting-based methods, probing, and fine-tuning, with fine-tuning proving the most effective. To further improve the accuracy, we introduce a new paradigm, named RATE-FT, that augments fine-tuning with an auxiliary task for the model to jointly learn with the main task of hallucination detection. With extensive experiments and analysis using a variety of model families & datasets, we demonstrate the effectiveness and generalizability of our method, *e.g.,* +3% over general fine-tuning methods on LongFact.

1 Introduction

With the recent advancements in model scale and pretraining data, large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks (Brown et al., 2020). Despite these successes, hallucination, where models tend to produce content that conflicts with real-world facts, remains a significant challenge (Zhang et al., 2023). Most existing research on hallucination detection has focused on short-form tasks, where the output consists of one or a few tokens. While these methods are effective for short-form content (Manakul et al., 2023; Mahaut et al., 2024; Yehuda et al., 2024; Zhang et al., 2024a), extending them to open-domain long-form generation presents additional complexities and new challenges. Unlike short-form tasks, long-form responses can span hundreds or even thousands of tokens, requiring models to generate detailed and nuanced answers to broad fact-seeking prompts (Wei et al., 2024). This necessitates that LLMs synthesize information across multiple knowledge domains, increasing the risk of generating content that sounds plausible yet is factually incorrect. For example, when answering 'What is the significance of Amber Room?', LLMs may generate responses that mix accurate historical information with fabricated details, complicating the task of distinguishing fact from hallucination.

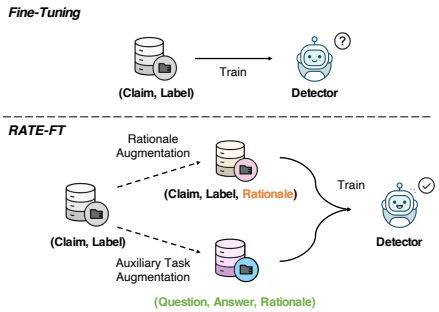

Figure 1: Comparison between Fine-Tuning and RATE-FT for hallucination detection. RATE-FT improves Fine-Tuning by incorporating rationales and an auxiliary task (question answering) into the training process.

Recent efforts have sought to address hallucination detection in long-form tasks. However, they either focus on limited domains, *e.g.,* biography generation (Min et al., 2023; Fadeeva et al., 2024) or rely heavily on external fact-checking tools or knowledge bases, *e.g.,* Google Search (Wei et al., 2024). While these tools offer valuable support, they are not always available or scalable. This raises an important question: *can we develop hallucination detectors that rely solely on the model itself, without the need for external fact-checking*

resources? So far, little attention has been given to systematically exploring how the model's own mechanisms can be used for detecting hallucinations in open-domain long-form generation.

To address this gap, we start by investigating hallucination detection in open-domain long-form responses using the model's internal states, *e.g.,* output probability and entropy. Specifically, we decompose long-form responses into atomized claims using the model and verify each claim's correctness using Google Search to construct benchmark data following Wei et al. (2024). Our analysis reveals that these internal states alone are insufficient for reliably (*i.e.,* better than random guessing) distinguishing between correct and incorrect claims, indicating that the mechanisms for detecting hallucinations in long-form outputs differ significantly from those in short-form tasks. To enhance detection, we explore several existing methods, including prompting, probing, and fine-tuning LLMs. Our experimental results show that fine-tuning LLMs is the most effective method to detect hallucinations.

Building on this, we introduce a novel method Rationale and Auxiliary Task Enhanced Fine-Tuning (RATE-FT) (Figure 1). Specifically, we convert the original claims into auxiliary question answering (QA) examples for augmentation, providing a complementary learning perspective for the model, which enables better generalization. Additionally, we incorporate collected rationales into the training process for better reasoning. Extensive experiments and analysis using different models demonstrate the effectiveness and generalizability of our approach. Furthermore, we investigate the integration of model uncertainty into hallucination detection in Appendix A.11.

2 Are LLMs' Internal States Sufficient for Open-Domain Long-Form Generation?

The internal states of LLMs, such as output probability and entropy, have been shown to be effective in detecting hallucinations in short-form tasks, where outputs are typically limited to only a few tokens. By analyzing these signals, models can often differentiate between factual and hallucinated information. However, their applicability in open-domain long-form generation remains underexplored. A key question is whether LLMs can depend solely on their internal states to identify hallucinations in long-form generation, without using external fact-checking tools. To answer it,

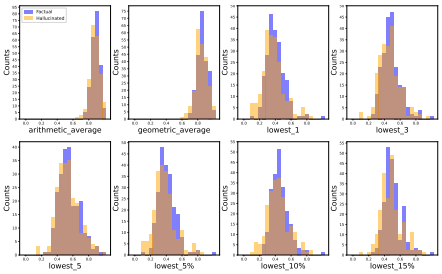

Figure 2: Detection results based on token probability.

we conduct some pilot experiments on LongFact (Wei et al., 2024), a long-form generation dataset spanning 38 different domains. Specifically, for each prompt in the sampled subset (200 prompts), we obtain a long-form response from Llama-3-8B-Instruct with greedy decoding. Following Wei et al. (2024), we employ the model to decompose long-form responses into atomized claims and label them as 'factual' or 'hallucinated' together with the reasons (see Appendix A.3 for construction details).

For each claim, we mainly focus on two types of internal states to estimate factual confidence following SelfCheckGPT (Manakul et al., 2023): the probability or the entropy (uncertainty) of output tokens. Specifically, we examine the arithmetic and geometric [1] averages of all tokens, the average of tokens with the top-K lowest probability or highest entropy ($K = 1, 3, 5$), and the average of tokens with the top-$P\%$ lowest probability or highest entropy ($P = 5, 10, 15$). The results in Figure 2 and Appendix A.4 suggest that neither internal state reliably, *i.e.,* better than random guessing, predicts the correctness of a given claim, which may be due to the presence of numerous insignificant tokens within the claim, such as stop words. To address this, we consider variants that focus only on output tokens related to entities. The results, shown in Appendix A.4, reveal similar patterns (see Appendix A.5 for a detailed comparison with the findings in Manakul et al. (2023)). We analyze the underlying reasons as follows. In open-domain long-form generation, claims are not limited to a few tokens, which introduces multiple sources of uncertainty. Specifically, *the probability or entropy reflects the model's confidence in how a claim is expressed, i.e., its confidence in the claim as a sequence of output tokens, rather than in the correctness of the claim.* Different surface forms of the claim yield different confidence levels, leading to unreliable estimates.

Considering the unreliability of LLMs' internal states in hallucination detection, there are sev-

[1] Commonly known as perplexity

eral promising alternative approaches, including prompting, probing and fine-tuning LLMs, which we explore in the next section.

3 Prompting, Probing and Fine-Tuning

Based on a review of the research area, we identify three groups of existing hallucination detection methods, which we discuss below.

Prompting Prompting-based approaches involve directly prompting LLMs to assess the correctness of a given claim without additional training. We investigate the following three different methods: (*i*) Prompting the model to output *'True' or 'False'* for a given claim, referred to as $\text{Prompt}_{\text{TF}}$. The probability assigned to the token 'True' represents P_{factual}, while the probability assigned to 'False' represents $P_{\text{hallucinated}}$. (*ii*) Prompting the model to output the *probability* that it considers the given claim to be correct, referred to as $\text{Prompt}_{\text{Prob}}$. This number directly represents P_{factual}. (*iii*) SelfCheckGPT, which detects hallucinations by sampling additional responses from the model and assessing inconsistencies between each response and the target claim. The *proportion* of responses that support the claim is taken as P_{factual}. Following Manakul et al. (2023), we sample 20 responses for detection.

Probing Following Su et al. (2024a), we train a multilayer perceptron (MLP) on the contextualized embeddings of LLMs to perform binary classification for hallucination detection, while keeping the base LLM frozen. The trained MLP outputs P_{factual} as an indicator for classification.

Fine-Tuning We fine-tune the base LLM with LoRA to enhance its ability to output 'True' or 'False' for a given claim (Kapoor et al., 2024). Similar to $\text{Prompt}_{\text{TF}}$, the probabilities assigned to the tokens 'True' and 'False' correspond to P_{factual} and $P_{\text{hallucinated}}$, respectively. Note that LoRA fine-tuning allows us to easily use the original model for general tasks while applying the trained LoRA specifically for hallucination detection.

Following the data construction process outlined in Appendix A.3, we conduct experiments on the full set of LongFact using Llama-3-8B-Instruct. This process yields 2,711 factual and hallucinated claims, which are subsequently split into training (70%), validation (20%), and test (10%) sets. For all three types of methods, we use P_{factual} as the classification indicator. Specifically, a claim is classified as 'factual' if P_{factual} exceeds a predefined threshold; otherwise, it is classified as 'hallucinated'. The optimal threshold is determined through a search on the validation set. Consistent with Tang et al. (2024); Chen et al. (2024b), we employ balanced accuracy (BAcc) as the evaluation metric: $\text{BAcc} = \frac{1}{2}(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}})$, where TP, TN, FP, and FN stand for true/false positives/negatives.

Dataset	Method				
	$\text{Prompt}_{\text{TF}}$	$\text{Prompt}_{\text{Prob}}$	SelfCheckGPT	Probing	Fine-Tuning
LongFact	69.9	53.4	69.1	74.4	**76.1**
Biography	72.3	56.3	71.9	77.0	**78.2**

Table 1: BAcc (%) of existing hallucination detection methods on LongFact and biography generation.

The results of different methods on the test set, as shown in Table 1, indicate that fine-tuning LLMs is the most effective among all existing methods (see Appendix A.6 for an analysis of fine-tuning effectiveness in Out-of-Distribution (OOD) scenarios). While both $\text{Prompt}_{\text{TF}}$ and SelfCheckGPT achieve decent performance, Probing yields notable improvements by incorporating additional training with labels obtained from external search. Fine-Tuning further enhances performance by updating the internal features of LLMs, enabling more effective learning. In contrast, $\text{Prompt}_{\text{Prob}}$ performs significantly worse, likely due to LLMs' tendency to output high probabilities for hallucinated claims, leading to overconfidence. Additionally, we extend the experiments to biography generation (Min et al., 2023). The results presented in Table 1 demonstrate that the observations and conclusions can be generalized to different datasets.

Building on these findings, a natural question arises: can Fine-Tuning be further improved to develop more effective hallucination detectors? We answer this question by *incorporating rationales and an auxiliary task into the training process.*

4 Rationale and Auxiliary Task Enhanced Fine-Tuning (RATE-FT)

While hallucination detection is not regarded as a reasoning task in the conventional sense, incorporating Chain-of-Thought (CoT) (Wei et al., 2022) explaining the judgment can still be beneficial for distinguishing factual content from hallucinated information as it enables LLMs to better evaluate the correctness of claims by systematically analyzing underlying components. To examine the impact of rationales, we prompt the model to generate a reasoning path before making a judgment (*i.e.,* 'True' or 'False'), referred to as $\text{Prompt}_{\text{CoT-TF}}$. This approach improves performance from 69.9 (using

Dataset	Method				
	$\text{Prompt}_{\text{TF}}$	$\text{Prompt}_{\text{CoT-TF}}$	Probing	Fine-Tuning	RATE-FT
LongFact	69.9	74.9	74.4	76.1	**79.6**
Biography	72.3	74.8	77.0	78.2	**80.9**

Table 2: BAcc (%) of RATE-FT and baseline methods.

Model	Method				
	$\text{Prompt}_{\text{TF}}$	$\text{Prompt}_{\text{CoT-TF}}$	Probing	Fine-Tuning	RATE-FT
Llama-3.1-70B-Instruct	73.2	76.8	79.4	80.6	**83.8**
Mistral-7B-Instruct	61.8	64.1	68.4	70.8	**73.4**
Qwen2.5-7B-Instruct	72.8	75.5	77.0	78.4	**81.1**

Table 3: Results using different models.

$\text{Prompt}_{\text{TF}}$) to 74.9, highlighting the effectiveness of incorporating CoT reasoning.

Augmenting Fine-Tuning with Rationales Building on the above observation, we augment the fine-tuning dataset with rationales generated by the model during data construction, explaining whether the search results support the claims. Notably, we adopt the 'label-rationale' format to maintain the same inference cost as the baseline Fine-Tuning. This allows us to directly derive P_{factual} from the first output token without requiring the generation of the complete reasoning path.

Consolidating knowledge through repetition in diverse contexts is a fundamental principle of effective human learning (Ausubel, 2012). For example, medical students deepen their understanding of anatomy by studying diagrams, practicing in simulations, and engaging in hands-on dissections, each offering a unique perspective on the same foundational knowledge. Drawing inspiration from this paradigm, we introduce an auxiliary question answering (QA) task into the fine-tuning process to further strengthen the model's understanding and enhance its generalization capabilities. This auxiliary QA task serves as a complementary component to the primary hallucination detection task, offering the model an alternative but closely related perspective on the problem (see Appendix A.7 for more analysis on the auxiliary task).

Augmenting Fine-Tuning with QA Task Specifically, for each claim, we first prompt the model to generate a question about the key information within it. If the claim is factual, we ask the model to extract the correct answer directly from the claim and provide an explanation, forming a QA example. For hallucinated claims, we leverage the augmented rationale to guide the model in generating an appropriate correct answer along with an explanation. After constructing these QA examples, they are combined with the original data for fine-tuning.

By integrating these two strategies, we propose **R**ationale and **A**uxiliary **T**ask **E**nhanced **F**ine-**T**uning (RATE-FT) (Figure 1). RATE-FT requires the model to systematically analyze and explain its judgments and allows the model to benefit from complementary learning perspectives, reinforcing its understanding of claims through diverse yet interconnected tasks. Following the experimental setup described in Section 3, we show the comparison between RATE-FT and baseline approaches in Table 2, which demonstrates the superiority of RATE-FT across different datasets (see Appendix A.8 for an analysis of the effect of additional data augmentation compared to the auxiliary QA task).

4.1 Further Analysis

Generalization to Different Models Our experiments and analysis so far use Llama-3-8B-Instruct as the backbone model. To verify whether the performance gain of RATE-FT is consistent across different backbone models, we extend the experiments to Llama-3.1-70B-Instruct (Dubey et al., 2024), Mistral-7B-Instruct (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Yang et al., 2024) on LongFact (see Appendix A.9 for details on data collection). From the results shown in Table 3, we can observe that RATE-FT consistently outperforms baseline approaches across all models, demonstrating its robustness and generalizability to diverse model architectures and scales.

In addition, we provide a summary of the main contributions, related work, ablation studies, results of incorporating uncertainty for hallucination detection, all prompts used in our experiments, and implementation details in Appendix A.1 $\sim$ A.2, A.10 $\sim$ A.15, respectively.

5 Conclusion

In this work, we systematically investigate reference-free hallucination detection in open-domain long-form generation. Our study begins with an analysis of the model's internal states, demonstrating that these states alone cannot reliably detect hallucinations. We then evaluate several existing approaches, including prompting, probing, and fine-tuning, with fine-tuning emerging as the most effective method. Building on these findings, we introduce Rationale and Auxiliary Task Enhanced Fine-Tuning (RATE-FT), a novel approach that leverages rationales and an auxiliary task to achieve significant improvements in detection performance across two datasets and various LLMs.

Limitations

One limitation of our work is its focus solely on improving the performance of the hallucination detector. A potential improvement could be to explore leveraging the detector's feedback as a reward signal to guide LLMs to generate more factual responses. Additionally, developing a more comprehensive benchmark for hallucination detection in open-domain long-form generation that covers a broader range of domains would further enhance its applicability.

References

David Paul Ausubel. 2012. *The acquisition and retention of knowledge: A cognitive view*. Springer Science & Business Media.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024a. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2024b. Teaching large language models to express knowledge boundary from their own signals. *arXiv preprint arXiv:2406.10881*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Minda Hu, Bowei He, Yufei Wang, Liangyou Li, Chen Ma, and Irwin King. 2024. Mitigating large language model hallucination with faithful finetuning. *arXiv preprint arXiv:2406.11267*.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7721–7735, Singapore. Association for Computational Linguistics.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. *arXiv preprint arXiv:2406.08391*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *arXiv preprint arXiv:2405.01525*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. 2024. Factual confidence of LLMs: on reliability and robustness of current estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570, Bangkok, Thailand. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What LMs know about unseen entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702, Seattle, United States. Association for Computational Linguistics.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024a. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14379–14391, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024b. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *arXiv preprint arXiv:2404.10774*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

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

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.

Xiaokang Zhang, Zijun Yao, Jing Zhang, Kaifeng Yun, Jifan Yu, Juanzi Li, and Jie Tang. 2024a. Transferable and efficient non-factual content detection via probe training with offline consistency checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12348–12364, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Main Contributions

The main contributions of our work are twofold.

- We are the first to systematically investigate reference-free hallucination detection in open-domain long-form generation by analyzing a representative set of existing methods.
- We introduce a novel approach that incorporates rationales and an auxiliary question answering task into fine-tuning, achieving significant performance improvements.

A.2 Related Work

Large Language Models (LLMs) often generate content that appears plausible but is factually unsupported, a phenomenon commonly known as hallucination (Zhang et al., 2023). Based on whether the hallucinated content contradicts read-world facts or the input context, hallucination can be categorized into two main groups: factuality hallucination and faithfulness hallucination (Huang et al., 2023). Extensive research has been conducted on exploring the causes (Onoe et al., 2022; Kang and Choi, 2023; Wei et al., 2023; Liu et al., 2024), detection (Min et al., 2023; Zhao et al., 2023; Chen et al., 2024a; Fadeeva et al., 2024; Wei et al., 2024), and mitigation (Gao et al., 2023; Ji et al., 2023; Tian et al., 2024; Zhang et al., 2024b; Kang et al., 2024; Lin et al., 2024) of hallucination in LLMs. However, most existing hallucination detection methods have primarily focused on short-form tasks, where the output consists of one or a few tokens. In this work, we shift the focus to the more challenging problem of reference-free hallucination detection in open-domain long-form generation, where outputs are substantially longer and require a more nuanced evaluation of actuality.

A.3 Benchmark Construction Details

For each prompt in the sampled subset (200 prompts), we obtain a long-form response from Llama-3-8B-Instruct with greedy decoding. Following Wei et al. (2024), we employ the model to decompose long-form responses into atomized claims and assess whether each claim is relevant to answering the corresponding prompt. For each relevant claim, we use the model to generate multi-step Google Search queries and reason about whether the search results support the claim. Claims supported by the search results are labeled as "factual", while those contradicted by the results are categorized as "hallucinated". After construction, we obtain 2394 factual claims and 223 hallucinated claims, respectively. We then randomly selected an equal number (223) of factual and hallucinated claims for experiments.

A.4 Hallucination Detection Results using Internal States

We show the hallucination detection results using different internal states in Figure 3 $\sim$ 5

A.5 Detailed Comparison with Findings in SelfCheckGPT

(*i*) While SelfCheckGPT (Manakul et al., 2023) explores several internal states of LLMs, our work covers a broader range of variants. As illustrated in Section 2, we examine the arithmetic and geometric averages (perplexity) of all tokens, the average of tokens with the top-K lowest probability or highest entropy ($K = 1, 3, 5$), and the average of tokens with the top-$P\%$ lowest probability or highest entropy ($P = 5, 10, 15$). In contrast, SelfCheckGPT only examines the arithmetic average of all tokens and the average of tokens with the top-1 lowest probability or highest entropy.

(*ii*) Our findings differ significantly from those reported in SelfCheckGPT. While SelfCheckGPT suggests that LLM probabilities correlate well with factuality, our experiments demonstrate that neither internal state reliably, *i.e.,* better than random guessing, predicts the correctness of a given claim. One possible explanation for this is the presence of many insignificant tokens, such as stop words, within the claim. To address this, we further investigate variants that focus only on output tokens related to entities (Appendix A.4), and the results exhibit similar patterns. Importantly, our findings are consistent with those in Kapoor et al. (2024).

A.6 Out-of-Distribution Results

We verify the effectiveness of fine-tuning in Out-of-Distribution (OOD) scenarios by training the model on LongFact and evaluating its performance on Biography. The results reported in Table 4 demonstrate that fine-tuning effectively generalizes to OOD scenarios.

A.7 More Analysis on Auxiliary Task

Comparison with F2 F2 (Hu et al., 2024) also integrates rationales and auxiliary tasks into the

training process. However, its main goal is to enhance the faithfulness of model responses while we focus on improving the accuracy of hallucination detection.

Further Clarification on Motivation The underlying motivation for introducing the auxiliary question answering (QA) task into fine-tuning is that hallucination detection and mitigation are complementary and closely related tasks. This auxiliary QA task—where a question about the key information in the claim is posed, and the model is trained to provide the correct answer—helps improve the factuality of the model's responses through supervised fine-tuning. It acts as a complementary component to the primary hallucination detection task, offering the model an alternative yet closely related perspective, thereby enhancing its generalization capabilities.

A.8 Additional Data Augmentation versus Auxiliary QA Task

To isolate the effect of additional data augmentation versus the auxiliary QA task, we design two variants: (*i*) we paraphrase the original claim using GPT-4 for data augmentation and fine-tune the model on the combined data, referred to as Fine-Tuning$_{\text{para}}$, which has roughly the same amount of training data as RATE-FT; and (*ii*) we reduce the training data for RATE-FT by half (approximately the same amount as Fine-Tuning), referred to as RATE-FT$_{\text{half}}$. We conduct experiments on LongFact using Llama-3-8B-Instruct and present the results in Table 5 and 6, which demonstrate that the performance improvement primarily comes from our designed auxiliary task, rather than from additional data augmentation.

A.9 Data Collection Process for Other Models

When conducting experiments using other models, we follow the exact same settings as those used for Llama-3-8B-Instruct. Specifically, for each prompt, we obtain a long-form response from the model under investigation with greedy decoding. Following Wei et al. (2024), we employ the model to decompose long-form responses into atomized claims and assess whether each claim is relevant to answering the corresponding prompt. For each relevant claim, we use the model to generate multi-step Google Search queries and reason about whether the search results support the claim. Claims supported by the search results are labeled as "factual", while those contradicted by the results are categorized as "hallucinated".

Our constructed benchmarks align well with Su et al. (2024b), as both include responses and internal states from various LLMs. The key difference is that the LLMs we investigate are all modern models (Llama-3-8B-Instruct, Llama-3.1-70B-Instruct, Mistral-7B-Instruct, and Qwen2.5-7B-Instruct), whereas the models used in Su et al. (2024b) are relatively outdated (such as LLaMA-2 and GPT-J).

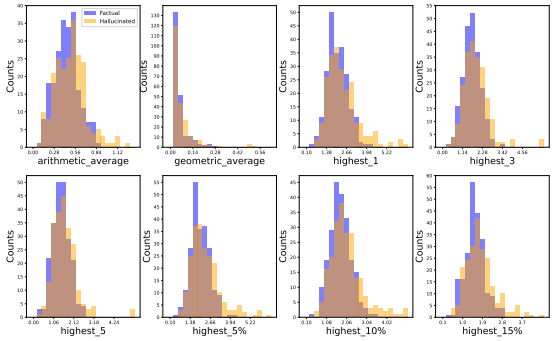

Figure 3: Hallucination detection results based on token entropy (uncertainty).

$\text{Prompt}_{\text{TF}}$	$\text{Prompt}_{\text{Prob}}$	SelfCheckGPT	Probing	Fine-Tuning
72.3	56.3	71.9	71.1	**74.7**

Table 4: Results of different methods in OOD scenarios.

A.10 Ablation Study

We analyze the contribution of different components of RATE-FT by investigating the variant of RATE-FT without the auxiliary task (*w.o.* aux). Table 7 presents the performance of different methods, highlighting that each component plays an important role in achieving the overall performance.

A.11 Incorporating Uncertainty for Hallucination Detection

To enhance hallucination detection, we propose incorporating model uncertainty into the detection process, enabling a hybrid pipeline that combines the strengths of the model and external tools. Specifically, when the model is uncertain about whether a claim is factual or hallucinated, we leverage external tools to handle ambiguous cases, improving overall performance. The process involves setting two thresholds, α_{low} and α_{high}, for classification. A claim is classified as 'factual' if $P_{\text{factual}} > \alpha_{high}$ and 'hallucinated' if $P_{\text{factual}} < \alpha_{low}$. Claims falling between these thresholds are classified as

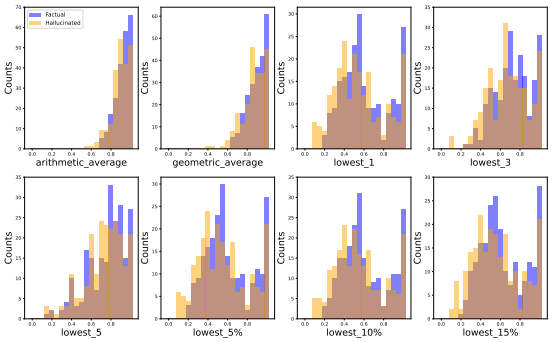

Figure 4: Hallucination detection results based on the probability of entity-related tokens.

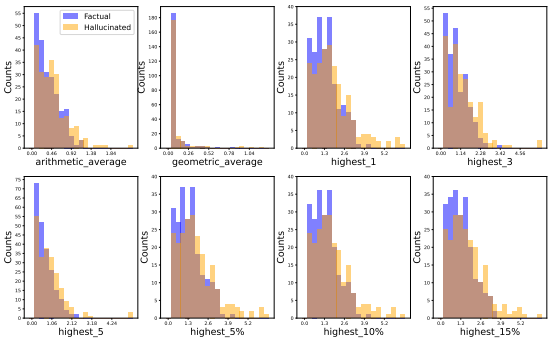

Figure 5: Hallucination detection results based on the entropy of entity-related tokens.

Fine-Tuning$_{\text{para}}$	RATE-FT
76.8	**79.6**

Table 5: Comparison between Fine-Tuning$_{\text{para}}$ and RATE-FT.

Fine-Tuning	RATE-FT$_{\text{half}}$
76.1	**78.5**

Table 6: Comparison between Fine-Tuning and RATE-FT$_{\text{half}}$.

Dataset	**Method**		
	Fine-Tuning	*w.o.* aux	RATE-FT
LongFact	76.1	77.5	**79.6**
Biography	78.2	79.4	**80.9**

Table 7: Results of different ablations.

Prompt$_{\text{CoT-TF}}$	Probing	Fine-Tuning	RATE-FT
80.4	81.1	82.4	**85.0**

Table 8: BAcc-unknown (%) of different methods on Longfact with Llama-3-8B-Instruct.

'unknown' and delegated to external tools for further evaluation. Assuming the external tools' output is the ground truth, predictions classified as 'unknown' are treated as correct. To evaluate the hybrid pipeline, we define the BAcc-unknown metric as follows:

$$\text{BAcc-unknown} = \frac{1}{2}\Big(\frac{\text{\# Correct Factual Predictions}}{\text{\# Total Factual Claims}} + \frac{\text{\# Correct Hallucinated Predictions}}{\text{\# Total Hallucinated Claims}}\Big) \quad (1)$$

The optimal thresholds, α_{low} and α_{high}, are determined through a search on the validation set. This process ensures that BAcc on the validation set exceeds 70%, while also maximizing BAcc-unknown. The goal is to strike a balance between performance and efficiency by achieving high BAcc-unknown without generating an excessive number of 'unknown' predictions, which could substantially increase detection costs. We conduct experiments on LongFact using Llama-3-8B-Instruct and report the results in Table 8, which demonstrate that incorporating model uncertainty greatly enhances hallucination detection, as evidenced by the BAcc-unknown metric's superior performance compared to standard BAcc in resolving ambiguous cases. Moreover, RATE-FT continues to outperform all other methods with respect to the BAcc-unknown metric, highlighting its robustness and effectiveness.

A.12 Prompt for Output Extraction

After decomposition, the atomized claims may differ from the original expression in the response. To address this, we use the prompt shown in Figure 6 to retrieve the original output corresponding to a given atomized claim.

A.13 Prompts for Baseline Approaches

Figure 7 illustrates the prompts used for different prompting methods. The prompt used for constructing training data in Probing and Fine-Tuning is the same as the prompt employed by the Prompt$_{\text{TF}}$ method.

A.14 Prompts Used in RATE-FT

Figure 8 presents all the prompts used in RATE-FT.

A.15 Implementation Details

For Prompt$_{\text{TF}}$ and Prompt$_{\text{Prob}}$, we obtain the response from the model with greedy decoding. Following Manakul et al. (2023), we set the temperature to 1.0 and generate 20 additional responses for

Prompt

Your task is to extract the original text corresponding to the given claim from the original response. When presented with an original response and a claim, reply with the original text. Make sure that your response is exactly the same as the original text and enclosed in \boxed{}.

Original response: {*response*}
Claim: {claim}

Figure 6: Prompt for extracting the original output given an atomized claim.

Prompt$_{\text{TF}}$

Your task is to determine the correctness of the given claim. When presented with a claim, reply with 'True' or 'False'. Make sure that your response is exactly 'True' or 'False' without any extra commentary whatsoever.

Claim: {claim}
Response:

Prompt$_{\text{CoT-TF}}$

Your task is to determine the correctness of the given claim. When presented with a claim, first explain the solution and then enclose the ultimate answer ('True' or 'False') in \boxed{}.

Claim: {claim}
Response:

Prompt$_{\text{Prob}}$

Your task is to provide the probability that the given claim is correct. When presented with a claim, reply with a number between 0.0 and 1.0. Make sure that your response is exactly a number between 0.0 and 1.0 without any extra commentary whatsoever.

Claim: The sun rises in the east and sets in the west.
Response: 1.0

Claim: Humans have four arms and three heads.
Response: 0.0

Claim: The human nose can detect over 1 trillion different scents.
Response: 0.82

Claim: The next president of South Korea will be a woman.
Response: 0.29

Claim: {claim}
Response:

SelfCheckGPT

Context: {context}

Sentence: {sentence}

Is the sentence supported by the context above? Answer Yes or No.

Answer:

Figure 7: Prompts for different prompting methods.

SelfCheckGPT.

We evaluate 4 different types of contextualized embeddings for Probing: (1) the final token from the last layer (type$_1$), (2) the average of all tokens in the last layer (type$_2$), (3) the average of the final token across all layers (type$_3$), and (4) the average of type$_1$ and type$_2$ (type$_4$). The optimal embedding type, along with other hyperparameters, *e.g.*, learning rate, is selected through a search on the validation set. For Fine-Tuning and RATE-FT, we leverage the LLaMA-Factory library (Zheng et al., 2024) and perform a search on the validation set for important hyperparameters.

Prompt for 'label-rationale' Format

Your task is to determine the correctness of the given claim. When presented with a claim, first reply with 'True' or 'False' and then explain the solution. Make sure that your response starts with 'True' or 'False'.

Claim: {claim}
Response: {True/False}. {explanation}

Prompt for Question Answering

Answer the following question and provide the explanation.

Question: {question}
Answer: {answer}
Explanation: {explanation}

Prompt for Question Generation (Correct Claim)

Given a correct claim and why it is correct, first identity the key information in the claim, then transform it into a question and a correct answer (keep the answer as concise as possible) about the key information, finally give the explanation (keep it different from the given reason). Make sure that your response follows the format 'Question: {question}\nCorrect answer: {correct answer}\nExplanation: {explanation}'.

Correct claim: {correct claim}
Reason: {reason}
Response:

Prompt for Question Generation (Wrong Claim)

Given a wrong claim and why it is wrong, first identity the key information in the claim, then transform it into a question and a correct answer (keep the answer as concise as possible) about the key information, finally give the explanation (keep it different from the given reason). Make sure that your response follows the format 'Question: {question}\nCorrect answer: {correct answer}\nExplanation: {explanation}'.

Wrong claim: {wrong claim}
Reason: {reason}
Response:

Figure 8: Prompts for different components of RATE-FT.