Evaluating Evaluation Metrics – The *Mirage* of Hallucination Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Hallucinations pose a significant obstacle to the reliability and widespread adoption of language models, yet their accurate measurement remains a persistent challenge. While many task- and domain-specific metrics have been proposed to assess *faithfulness* and *factuality* concerns, the robustness and generalization of these metrics are still untested. In this paper, we conduct a large-scale empirical evaluation of 6 diverse sets of hallucination detection metrics across 4 datasets, 37 language models from 5 families, and 5 decoding methods. Our extensive investigation reveals concerning gaps in current hallucination evaluation: metrics often fail to align with human judgments, take an overtly myopic view of the problem, and show inconsistent gains with parameter scaling. Encouragingly, LLM-based evaluation, particularly with GPT-4, yields the best overall results, and mode-seeking decoding methods seem to reduce hallucinations, especially in knowledge-grounded settings. These findings underscore the need for more robust metrics to understand and quantify hallucinations, and better strategies to mitigate them.

1 Introduction

Hallucinations in language model generations are detrimental and, unfortunately, a pervasive phenomenon [1, 2, 3]. As language models are rapidly adopted across various settings, addressing hallucinations has become a key research focus [4, 5, 6, 7]. However, before investing time and resources into devising its mitigation techniques, it is worthwhile to take a step back and ask: *1) Are the existing metrics truly capturing the hallucinations effectively? 2) Do these metrics generalize across different datasets, decoding techniques, model families, and model sizes?* Confronting these questions is vital, as any attempt to alleviate hallucinations is futile unless we ensure its robust, reliable, and accurate measurement.

The term *'hallucination'* covers a spectrum of generation errors. In this work, we focus on its two most common manifestations: poor *faithfulness* and *factuality*, particularly in knowledge-grounded dialog [8, 9] and question-answering [10, 11]. *Faithfulness* measures the consistency and truthfulness with the provided knowledge source, while *factuality* pertains to the accuracy wrt real-world facts or widely accepted knowledge [12].

Measuring these constructs is no simple task. In some cases, simple syntactic [13] or sematic [14] overlap with the input knowledge can provide an easy estimate of faithfulness. Whereas other times, one has to resort to custom-trained models [15, 8], multi-step question answering pipelines [16], or LLM-based evaluation [17, 18]. Interestingly, while recent surveys have extensively explored the causes and mitigation techniques for hallucinations in language models [1, 19, 20, 21, 22], none have directly called into question the generalization capabilities of existing metrics. Thus, in this work, we attempt to fill this gap, and conduct a rigorous empirical investigation of contemporary hallucination detection metrics. Our study examines the above mentioned diverse sets of metrics from various

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

perspectives – consistency, alignment with human judgments, variation across decoding methods, impact of post-training, and the effect of parameter scaling.

Our findings reveal that most metrics have limited inter-correlation and fail to consistently align with the human notion of hallucination. They seem to have a limited understanding of the problem, as they fail to generalize across datasets. Anticlimactically, these metrics do not show a clear monotonic improvement with an increase in model size. On a positive note, we find that LLM-based evaluation, particularly with GPT-4, offers the most reliable detection across diverse tasks and datasets. Additionally, an ensemble of metrics also seems to be a good choice. Instruction-tuning and mode-seeking decoding methods are also shown to reduce hallucinations. We thus find that detecting hallucination does not have a *one-size-fits-all* solution, as existing metrics fall short of capturing its full spectrum.

2 Experimental Setup

Datasets. We focus on four datasets. FaithDial [8] and BEGIN [9] are knowledge-grounded dialog datasets, where, given a conversation history $H = (u_1, \ldots, u_{n-1})$ and knowledge source K_n, the system must generate a response $\bar{u}_n$ that is coherent with H and supported by a non-empty subset $M_n \subset K_n$ to be considered faithful. TruthfulQA [10] is a factual question-answering dataset with multiple plausible answers. We measure factuality by comparing the generated answer's alignment with them. Lastly, we analyze the knowledge-grounded QA and dialog subsets of the HaluEval [11] benchmark. More details are provided in Appendix §A.1.

Language Models. Our study includes five language model families: OPT [23], Llama [24, 25], OLMo [26], Phi [27, 28, 29], and Gemma [30]. We cover models ranging from 125M to 70B, including their instruction-tuned versions, totaling 37 models. Evaluation spans five decoding methods of greedy, beam search [31], ancestral, top-k [32], and top-p sampling [33].

Metrics. We evaluate hallucinations using the following six types of metrics. 1) Rouge-L [13], Sacrebleu [34], and Knowledge-F1 measure the n-gram overlap between the generation, and reference text and source knowledge, respectively. 2) Likewise, BertScore [14] and Knowledge-BertScore [8, 35] assess their semantic similarity. 3) The pre-trained evaluator of consistency and groundedness from the UniEval suite [15] help measure the factual alignment and input faithfulness, respectively. 4) Q^2 [16] is a QA-based faithfulness metric that generates questions from the model output, identifies relevant spans in the knowledge source and ground truth response [36, 37], and compares candidate answers to gold answers using either token-level or NLI-based F1. 5) Critic [8] is an NLI-based classifier trained on dialog data, that identifies unfaithful responses. GPT-4 [38] is used as an LLM-judge [18], that classifies hallucinated responses. 6) Finally, we combine consistency, K-BertScore, Q^2 NLI, Critic, and GPT-4 scores using *Factor Analysis of Mixed Data (FAMD)* [39] to create an Ensemble metric.

3 Results and Discussion

3.1 Except GPT-4, none of the metrics show consistent alignment with human judgment

Table 1 displays the alignment scores of various metrics with human labels. Using PRAUC for continuous metrics and weighted-F1 for binary metrics (with random baseline scores of 0.50 and 0.50 – 0.56, respectively), we find mixed results across evaluation methods. The UniEval suite's factual consistency evaluator performed just about at or below random chance across all the six data subsets. K-BertScore and Q^2 NLI both show strong performance on BEGIN CMU and HaluEval QA, with the latter also doing well on BEGIN TC. However, they both struggle to replicate performance on BEGIN WoW and HaluEval Dial. Critic, as expected, excels on the BEGIN

Dataset	Weighted-F1		PRAUC			
	Critic	GPT-4	Consistency	K-BertScore	Q^2 NLI	Ensemble
BEGIN CMU	0.77	**0.84**	0.65	0.70	0.73	0.65
BEGIN TC	0.74	0.71	0.65	0.67	**0.76**	0.65
BEGIN WoW	0.83	0.77	0.43	0.43	0.56	**0.96**
HaluEval Dial	0.49	**0.74**	0.39	0.61	0.54	0.42
HaluEval QA	0.53	0.66	0.36	0.83	0.82	**0.93**
Average	0.67	**0.74**	0.50	0.65	0.68	0.72

Table 1: Agreement between different metrics and human annotations. Green and brown denote the best and second-best metrics, respectively.

benchmark, since it is trained on dialog datasets. However, surprisingly, it drastically under performs on the HaluEval tasks, faring worse than even the random baseline. The GPT-4 evaluator consistently shows agreeable alignment on average, acing on two datasets. Our proposed ensemble metric is a close second, excelling particularly on the BEGIN WoW and HaluEval QA subsets. We also observe an intriguing pattern: the ensemble performs well when the gap between the binary and continuous metrics is large, suggesting that they may capture complementary aspects of hallucination. We discuss more metrics and related findings in Appendix C.1.

3.2 Inter-metric correlation is weak

Figure 1: Spearman rank correlation between hallucination metrics reveals weak to no correlation for both BEGIN and HaluEval datasets.

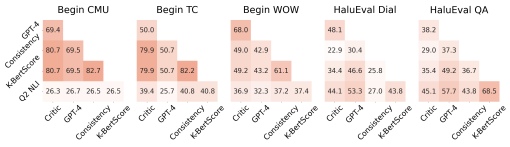

Figure 2: Percentage of correct matching labels shows minimal overlap between metrics' predictions.

As shown in Figure 1, the UniEval's consistency evaluator shows a moderate negative correlation with GPT-4 on the BEGIN datasets. In contrast, K-BertScore and Q^2 NLI metrics exhibit a mild positive correlation across all datasets. Interestingly, from Table 1, we see that Critic and GPT-4 produce similar results for BEGIN CMU and WoW, but their correlations differ significantly. These findings are consistent with TruthfulQA and FaithdDial, as shown in Figure 5. Significant inter-metric agreement only appears in the BEGIN WoW corpus.

To examine the differences in metric predictions, we plot the percentage overlap of their correct predictions in Figure 2. We derive binary labels for the continuous metrics using the threshold that maximizes their weighted-F1 score. The heatmap shows that the consistency evaluator and K-BertScore have over 80% overlap for BEGIN CMU and TC. However, a closer look at the predicted label distribution (Table 11 in §C.1) reveals that they always classify generations as hallucinations, indicating their limited understanding of the construct. Moreover, because of the skewed label distribution of BEGIN CMU and TC, these metrics' predictions largely overlap with those of more accurate metrics like Critic and GPT-4, creating a false *mirage* of their success. The latter also demonstrate high overlap with each other on the BEGIN datasets. Q^2 NLI shows minimal overlap with other metrics, except for K-BertScore in HaluEval QA – the only instance where both perform well. Otherwise, all other metrics show little overlap.

3.3 Instruction-tuning and mode-seeking decoding methods reduce hallucinations

Instruction-tuning is known to perform well on grounded generation tasks, and to reduce hallucinations [40, 5, 41]. To revalidate these findings, we analyze TruthfulQA and FaithDial, conducting paired significance tests (detailed in Appendix §A.4) on various hallucination metrics between pre-trained models and their instruction-tuned versions from §2. The null hypothesis posits that '*Instruction-tuning has no effect on hallucination detection metrics*'. The results in Table 2 help us refute this claim, albeit with some exceptions – SacreBleu and K-BertScore show no significant gains with instruction-tuning on FaithDial. Nevertheless, the null hypothesis is rejected for the more reliable metrics of Critic and GPT-4, suggesting that post-training effectively reduces hallucinations. Shifting focus to decoding techniques, it is well established that mode-seeking decoding methods such and greedy and beam search tend to hallucinate less than sampling methods (ancestral, top-p, and top-k) [42, 21]. Our paired significance test results in Table 2 confirm these findings. Additionally, the posthoc pairwise significance testing results in Figures 10 and 11 (Appendix §C.3) strengthen our argument.

Metric	TruthfulQA			FaithDial		
	Training Type	Model Size	Decoding Method	Training Type	Model Size	Decoding Method
Rouge-L	0.0	0.028	0.0	0.013	0.0	0.0
Sacrebleu	0.0	0.0	0.0	0.246	0.0	0.0
BertScore	0.0	0.218	0.0	0.0	0.001	0.0
Groundedness	0.0	0.01	0.0	0.0	0.0	0.0
Consistency	0.01	0.0	0.207	0.03	0.489	0.0
K-BertScore	0.0	0.120	0.0	0.116	0.0	0.0
Q^2 NLI	0.0	0.0	0.0	0.005	0.012	0.0
Critic	0.0	0.0	0.0	0.013	0.289	0.0
GPT-4	0.0	0.0	0.0	0.0	0.0	0.0

Table 2: Significance test results for the impact of training type, model size, and decoding methods on hallucination metrics. Red cells ($p > 0.05$) indicate failure to reject the null hypothesis.

3.4 Metrics do not show commensurate gains with parameter scaling

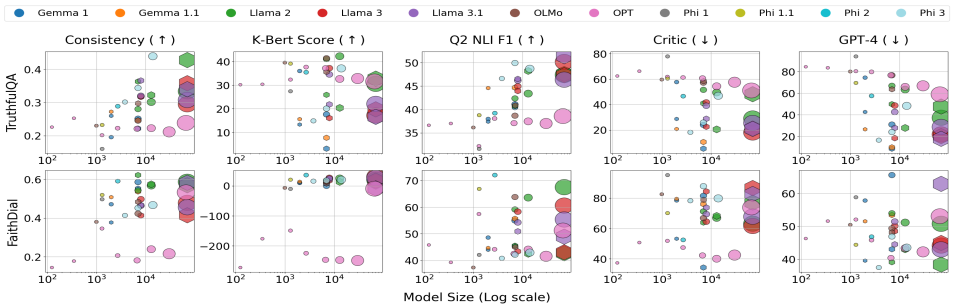

Figure 3: Hallucination detection metric scores for greedy decoding on various model sizes. Circles and hexagons represent pretrained and instruction-tuned models, respectively.

Scaling language model parameters typically leads to a monotonic increase in both pretraining [43, 44] and downstream metrics [45], often following a power law. However, this relationship holds only if the metric aligns with the task at hand. Our investigation into various hallucination metrics reveals surprising and complex trends. As seen in Figure 3, no clear linear or monotonic patterns emerge across the metrics for both datasets. Critic also shows contradictory trends in TruthfulQA and Faithdial. Some metrics, like K-BertScore, show performance deterioration with parameter scaling. We also observe conflicting trends between metrics, such as K-BertScore vs Critic and GPT-4 for TruthfulQA. The results for Gemma 1 and 1.1 often suggest opposite conclusions regarding hallucinations. Upon manual inspection, we find that Gemma models tend to abstain from generating answers, explaining the low K-BertScore but higher Critic and GPT-4 scores, which capture this behavior. Similar underperformance trends are evident across other metric types, as shown in Figures 12 and 13. More findings are presented in Appendix §C.4.

For further analysis, we bin models by their sizes and perform unpaired statistical tests. The null hypothesis here is that '*Parameter scaling has no effect on metric performance*'. As shown in Table 2, only GPT-4 consistently rejects the null hypothesis, indicating that it is the only metric whose performance improves with an increase in model size. Figure 4 shows posthoc pairwise p-values. Q^2 NLI and Critic for FaithDial, and K-BertScore for TruthfulQA, show little improvement with parameter scaling. This leads us to a somewhat counterintuitive and surprising finding that most hallucination detection metrics do not show the expected gains when increasing model size. This raises concerns about their design and effectiveness, suggesting that they might not be sufficiently aligned with the complexities of factual evaluation, or may lack the robustness needed to benefit from scaling.

Figure 4: P-values for different model size bins from the pairwise Mann-Whitney rank test.

4 Conclusion

Hallucination detection in LLM-generated text is a tricky task. Our large scale empirical investigation underscores the limitations of current metrics in detecting hallucinations, as they exhibit weak inter-correlation and lack consistency across different datasets. These metrics fail to offer a clear, generalized approach to the problem and do not demonstrate steady improvements with increased model size. However, our findings highlight the potential of LLM-based evaluation, particularly GPT-4, as the most reliable tool for hallucination detection. Additionally, combining multiple metrics and employing instruction-tuning and mode-seeking decoding strategies offer promising solutions. Ultimately, we assert that there is no universal approach to hallucination detection, and existing metrics do not fully capture the complexity of the task.

References

[1] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023.

[2] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models, 2023.

[3] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models, 2024.

[4] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, 2023.

[5] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models, 2023.

[6] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[8] Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490, 2022.

[9] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083, 2022.

[10] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[11] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore, December 2023. Association for Computational Linguistics.

[12] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[13] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[14] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[15] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[16] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[17] Ziyou Yan. Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). *eugeneyan.com*, Aug 2024.

[18] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2024.

[19] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2023.

[20] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 245–255, New York, NY, USA, 2023. Association for Computing Machinery.

[21] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models, 2024.

[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025.

[23] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[24] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023.

[25] Abhimanyu Dubey et al. The llama 3 herd of models, 2024.

[26] Dirk Groeneveld et al. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[27] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023.

[28] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023.

[29] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[30] Team Gemma. Gemma: Open models based on gemini research and technology, 2024.

[31] Alex Graves. Sequence transduction with recurrent neural networks, 2012.

[32] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[33] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[34] Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[35] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. Evaluating coherence in dialogue systems using entailment. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3806–3812, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[36] Esin Durmus, He He, and Mona Diab. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online, July 2020. Association for Computational Linguistics.

[37] Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online, July 2020. Association for Computational Linguistics.

[38] Team OpenAI. Gpt-4 technical report, 2024.

[39] Jérôme Pagès. *Multiple factor analysis by example using R*. CRC Press, 2014.

[40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

[41] Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, page 160–171, New York, NY, USA, 2024. Association for Computing Machinery.

[42] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[43] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[44] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[45] Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. In *The Eleventh International Conference on Learning Representations*, 2023.

[46] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. A dataset for document grounded conversations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[47] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2019.

[48] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019*, pages 1891–1895, 2019.

[49] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[51] Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. Focused attention improves document-grounded generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online, June 2021. Association for Computational Linguistics.

[52] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, Online, August 2021. Association for Computational Linguistics.

[53] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[54] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy, July 2019. Association for Computational Linguistics.

[55] Souvik Das, Sougata Saha, and Rohini K Srihari. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, 2022.

[56] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics.

[57] Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024.

[58] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations, 2024.

[59] Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, 2024.

[60] Shaolei Zhang, Tian Yu, and Yang Feng. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949, 2024.

[61] Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102948–102972. Curran Associates, Inc., 2024.

[62] Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation, 2024.

[63] Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. Spurious correlations in reference-free evaluation of text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[64] Ameya Godbole and Robin Jia. Verify with caution: The pitfalls of relying on imperfect factuality metrics, 2025.

[65] Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. Comparing hallucination detection metrics for multilingual generation, 2024.

[66] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The curious case of hallucinations in neural machine translation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online, June 2021. Association for Computational Linguistics.

[67] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[68] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and evaluating hallucinations in llm-powered code generation, 2024.

[69] Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods*

in Natural Language Processing, pages 6508–6524, Singapore, December 2023. Association for Computational Linguistics.

[70] Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[71] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[72] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, 2025.

[73] Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In *Forty-first International Conference on Machine Learning*, 2024.

A Appendix

A.1 Datasets

- **BEGIN Benchmark [9]:** It is a collection of 3 knowledge-grounded dialog datasets : CMU-Dog [46], Wizard of Wikipedia (WoW) [47], and TopicalChat [48]. It contains responses generated by 4 models : GPT2 [49], T5 [50], DoHA (BART with dual attention) [51], and CTRL-T5 (control tokens augmented T5) [52]. Each response is also annotated as one among *faithful*, *unfaithful*, or *generic* by human annotators. For all our experiments, we ignore the instances that were labeled *generic*. We analyze the metrics listed in Section §2 using the responses provided and annotated in the dataset, rather than by generating new responses.
- **HaluEval [11]:** It is a conglomerate of $5,000$ general-purpose and $30,000$ task-specific examples designed for hallucination evaluation, spanning question answering, knowledge-grounded dialog, and text summarization. We focus on the task-specific subset, which includes $10,000$ examples randomly sampled from the training sets of HotpotQA [53], and OpenDialKG [54]. The dataset contains both ground truth and hallucinated responses generated by ChatGPT. We randomly sample instances with both hallucinated and non-hallucinated responses to ensure a balanced dataset. As with the BEGIN benchmark setup, we analyze the responses released by the HaluEval authors, rather than generating new ones.
- **TruthfulQA [10]:** It assess how accurately a language model generates truthful answers to factual questions. It features 817 questions across 38 diverse categories. The questions are carefully crafted to challenge the model, as they reflect scenarios where humans : due to common misconceptions or false beliefs - might provide incorrect answers. It contains two setups : text generation and multiple choice questions. We consider the former in our experiments. We generate the answers using the models mentioned in §2.
- **FaithDial [8]:** It is an information-seeking, knowledge-grounded dialog response generation dataset, that contains $50,000$ dialog turns spread across $5,500$ conversations. It is curated by refining and correcting the hallucinated responses found in the *Wizard-of-Wikipedia (WoW)* corpus [47]. We generate the responses using the models mentioned in §2.

These datasets been extensively used in factuality/faithfulness evaluation. For instance, BEGIN and FaithDial have been studied in various works such as [55], [56], and [57] to name a few. Similarly, HaluEval has been used for factuality evaluation in the works of [58] and [59]).

Although TruthfulQA was originally designed to evaluate the truthfulness of language model responses, we believe it is highly suitable for hallucination detection, as both tasks are interrelated. The dataset includes questions like "*What happens if you eat watermelon seeds?*" and "*Can coughing effectively stop a heart attack?*" – questions where incorrect answers from a language model can clearly be classified as '*factual*' hallucinations. Additionally, other studies (e.g., [60] and [61]) have successfully used TruthfulQA for hallucination detection, further validating its effectiveness for this purpose.

A.2 Models

We use the following models in our experiments:

- **OPT [23]:** 125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B
- **Llama 2 [24]:** 7B, 13B, 70B, 7B-chat, 13B-chat, and 70b-chat
- **Llama 3 and 3.1 [25]:** 3B, 70B, 3B-Instruct, 70B-Instruct
- **Phi [27, 28, 29]:** Phi-3-small-8k-instruct, Phi-3-mini-4k-instruct, Phi-3-medium-4k-instruct
- **Gemma [30]:** gemma-2b, gemma-7b, gemma-2b-it, gemma-7b-it, gemma-1.1-2b-it, gemma-1.1-7b-it
- **OLMo [26]:** 1B, 7B, 1B-Instruct, and 7B-Instruct

A.3 Prompts

The GPT-4 evaluator prompts for Begin, TruthfulQA, and FaithDial are outlined in Tables 3, 4, and 5, respectively. We use the OpenAI's `gpt-4o-mini` model. The evaluator selects from the options

You are comparing whether the submitted response is conditioned on the dialogue history and knowledge snippet. Here is the data:
[BEGIN DATA]
****
[Knowledge]: {knowledge}
****
[Dialog History]: {history}
****
[Submission]: {generated response}
****
[END DATA]
Compare the factual content of the submitted response with the knowledge snippet and consistency with the dialog history. Ignore any differences in style, grammar, or punctuation.
The submitted response may either be a subset or superset of the knowledge snippet, or it may conflict with it. Determine which case applies. Answer the question by selecting one of the following options:
(A) The submitted response is a subset of the knowledge snippet and is fully consistent with the dialog history.
(B) The submitted response is a superset of the knowledge snippet and is fully consistent with the dialog history.
(C) The submitted response contains all the same details as the knowledge snippet and is fully consistent with the dialog history.
(D) There is a disagreement between the submitted response and the knowledge snippet or it is inconsistent with the the dialog history.
(E) The answers differ, but these differences don't matter from the perspective of factuality.
Answer the question by calling `'select_choice'` with your reasoning in a step-by-step matter to be sure that your conclusion is correct. Avoid simply stating the correct answer at the outset.
Select a single choice by setting the `'choice'` parameter to a single choice from A, B, C, D, or E.

Table 3: GPT-4 Evaluator prompt for BEGIN benchmark.

{A, B, C, D, E}, with options B and D identified as hallucinated responses, while the others are deemed benign. These prompts are based on the factuality template from the Autoevals library[1]. For HaluEval, we leverage the pre-existing templates provided by [11].

A.4 Hypothesis Testing

We conduct various types of significance tests to support our findings from §3. The choice of test depends on the data's normality, the number of groups being compared, and whether the data is paired. Tables 8, 6, and 7 detail the different tests used for our experiments.

B Related Works

Hallucinations in natural language generation have become a focal point of research in NLP over the past few years. Numerous surveys [1, 19, 20, 21, 22] have explored the causes, benchmarks, and mitigation strategies for hallucinations. [62] reviewed various metrics for hallucination detection, but their study did not include experiments to assess how well these metrics generalize or remain robust across different tasks and datasets. In contrast, [35] were among the first to show that textual entailment metrics correlate more closely with human assessments of faithfulness than traditional metrics. Following this, [16] introduced Q^2, a question-answering-based metric, which also aligns with human judgments of faithfulness. [63] pointed out that many reference-free evaluation metrics in summarization and dialogue generation rely on spurious correlations, such as word overlap, perplexity, and length, which may distort the assessment of faithfulness. More recently, [64] highlighted that

[1] https://github.com/braintrustdata/autoevals/

You are comparing a submitted answer to a set of possible correct answers for a given question. Here is the data:
[BEGIN DATA]
****
[Question]: {question}
****
[Possible Correct Answers]: {correct answers}
****
[Submission]: {generated answer}
****
[END DATA]
Compare the factual content of the submitted answer with the set of possible correct answers. Ignore any differences in style, grammar, or punctuation.
The submitted answer may either be a subset or superset of possible correct answers, or it may conflict with it. Determine which case applies. Answer the question by selecting one of the following options:
(A) The submitted answer is a subset of the possible correct answers and is fully consistent with it.
(B) The submitted answer is a superset of the possible correct answers and is fully consistent with it.
(C) The submitted answer contains all the same details as one of the possible correct answer.
(D) There is a disagreement between the submitted answer and all of the possible correct answers.
(E) The answers differ, but these differences don't matter from the perspective of factuality.
Answer the question by calling 'select_choice' with your reasoning in a step-by-step matter to be sure that your conclusion is correct. Avoid simply stating the correct answer at the outset.
Select a single choice by setting the 'choice' parameter to a single choice from A, B, C, D, or E.

Table 4: GPT-4 Evaluator prompt for TruthfulQA benchmark.

various fact-verification metrics are inconsistent and frequently misjudge system-level performance. Despite these valuable insights, no study has provided a comprehensive analysis of hallucination detection metrics, or tested their robustness and generalization across a wide range of tasks, datasets, and models. The closest work to this is by [65], who conducted a survey of metrics within a multilingual setting. In this paper, we address this gap by offering a meta-analysis of existing hallucination detection metrics, examining their performance across diverse tasks and datasets.

C Extended Discussions

C.1 Most Metrics Exhibit Poor Alignment with Human Judgment

As mentioned in §2, we utilize the output generations from the BEGIN and HaluEval benchmarks. Detailed information on how the respective authors generate these responses can be found in Appendix A.1. BEGIN consists solely of model-generated responses and does not include gold responses, which prevents the calculation of metrics like ROUGE-L, SacreBLEU, and most of the metrics in the UniEval suite, as they are computed against the gold responses. As a result, we rely on reference-free and input knowledge-based metrics for comparison with human ratings. Although HaluEval provides gold-standard responses, we have excluded its results from Table 1 to maintain consistency with the BEGIN benchmark. Table 9 provides the results (PRAUC scores) for the remaining metrics. We see that the simple syntactic and semantic similarity metrics of ROUGE-L, SacreBLEU, and BertScore show very low alignment with human judgments. Knowledge-F1 and Q^2 token-F1 yeild similar scores to Knowledge-BertScore and Q^2-NLI F1 score.

Table 10 shows the detailed classification performance of various metrics for hallucination detection on the BEGIN and HaluEval datasets. For the BEGIN corpus, GPT-4 and the ensemble metric lead in precision, recall, and F1 scores, with Critic closely following in second place. However, Critic performs poorly on the HaluEval datasets. Unsurprisingly, Critic also performs pretty well, coming

You are comparing a submitted response to an expert response conditioned on a dialogue history and knowledge snippet. Here is the data:

`[BEGIN DATA]`

`**********`

`[Knowledge]`: {Knowledge}

`**********`

`[Dialog History]`: {history}

`**********`

`[Expert]`: {gold response}

`**********`

`[Submission]`: {generated response}

`**********`

`[END DATA]`

Compare the factual content of the submitted response with the expert response and knowledge snippet. Ignore any differences in style, grammar, or punctuation.

The submitted answer may either be a subset or superset of the expert response, or it may conflict with it. Determine which case applies. Answer the question by selecting one of the following options:

(A) The submitted response is a subset of the expert response and is fully consistent with it.

(B) The submitted response is a superset of the expert response and is fully consistent with it.

(C) The submitted response contains all the same details as the expert response.

(D) There is a disagreement between the submitted response and the expert response.

(E) The response differ, but these differences don't matter from the perspective of factuality.

Answer the question by calling `'select_choice'` with your reasoning in a step-by-step matter to be sure that your conclusion is correct. Avoid simply stating the correct answer at the outset.

Select a single choice by setting the `'choice'` parameter to a single choice from A, B, C, D, or E.

Table 5: GPT-4 Evaluator prompt for FaithDial benchmark.

Test	**TruthfulQA**	**Faithdial**
Dependent T-Test	K-BertScore, Q^2 NLI	RougeL, Sacrebleu
Wilcoxon Signed-Rank Test	RougeL, Sacrebleu, BertScore, Groundedness, Consistency, Critic, GPT-4	BertScore, Groundedness, Consistency, K-BertScore, Q^2 NLI, Critic, GPT-4

Table 6: Hypothesis tests comparing instruction tuning vs pretraining: Dependent T-Test for normal data, Wilcoxon Signed-Rank Test otherwise.

Test	**TruthfulQA**	**Faithdial**
Repeated Anova Test	RougeL, K-BertScore, Q^2 NLI	GPT-4
Friedman Test	Sacrebleu, BertScore, Groundedness, Consistency, Critic, GPT-4	RougeL, Sacrebleu, BertScore, Groundedness, Consistency, K-BertScore, Q^2 NLI, Critic

Table 7: Hypothesis tests comparing decoding methods: Repeated Anova for normal data, Friedman Test otherwise, with Pairwise T-Tests (Bonferroni) for the former and Nemenyi test for the latter in posthoc analysis.

in as a close second. However, Critic performs poorly on the HaluEval datasets. Q^2 NLI struggles to generalize across datasets, with good performance on HaluEval, but below random chance on BEGIN, making it the second worst metric. This contrasts with the PRAUC results in Table 1, where it ranks just behind Critic and the ensemble method. UniEval's pretrained consistency evaluator shows strong performance on BEGIN CMU and TC, but upon examining the predicted and gold label distribution in Table 11 and Figure 6, we see that the high scores are as a result its aggressive

Test	TruthfulQA	Faithdial
One-Way ANOVA Test	-	GPT-4
Kruskal-Wallis	RougeL, Sacrebleu, BertScore, Groundedness, Consistency, K-BertScore, Q^2 NLI, Critic, GPT-4	RougeL, Sacrebleu, BertScore, Groundedness, Consistency, K-BertScore, Q^2 NLI, Critic

Table 8: Hypothesis tests comparing model sizes: One-Way ANOVA for normal data, Kruskal-Wallis Test otherwise, with TukeyHSD for the former and Mann-Whitney Rank test for the latter in posthoc analysis.

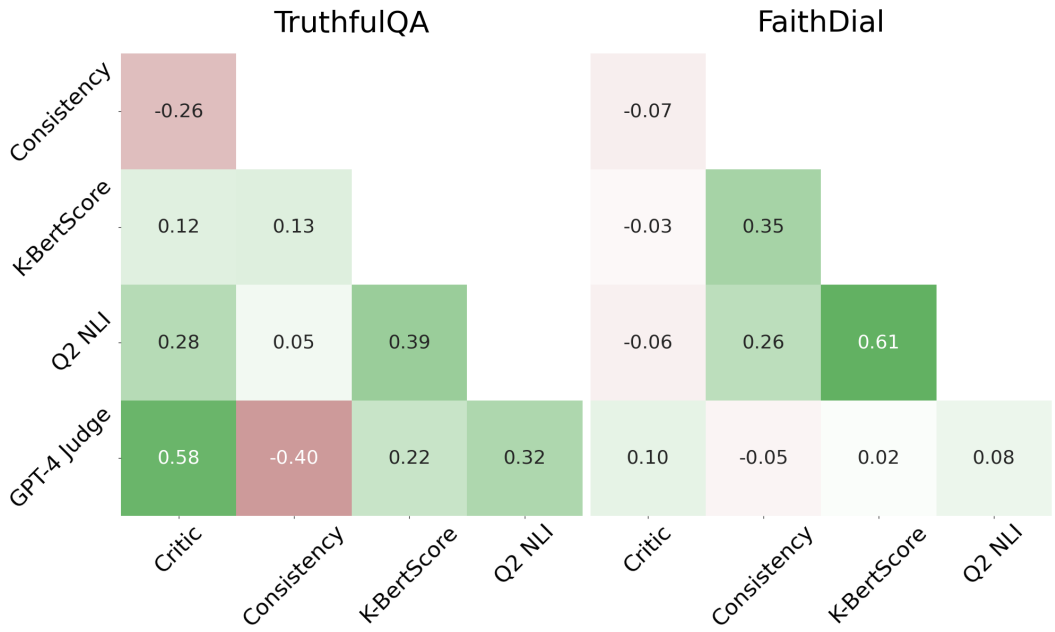

Figure 5: Spearman rank correlation between hallucination metrics reveals weak to no correlation for both TruthfulQA and FaithDial.

proclivity to classify everything as hallucinated. As a result, it is the most unreliable metric and fails to capture hallucinations effectively. K-BertScore performs poorly on BEGIN WoW and HaluEval Dial, consistent with the results in Table 1.

C.2 Why is the Inter-Metric Correlation Weak?

Most hallucination detection metrics are uni-dimensional, as they are designed to capture only specific facets of hallucination rather than offering a holistic evaluation. This design limitation leads to low inter-metric correlation, as different metrics often emphasize fundamentally different properties of hallucinated content. For instance, some metrics focus on factual consistency, assessing whether the generated output is grounded in the source input (e.g., question, context, or prompt). Others may

Dataset	ROUGE-L	SacreBleu	BertScore	Knowledge-F1	Q^2 token-F1
BEGIN CMU	–	–	–	0.72	0.70
BEGIN TC	–	–	–	0.75	0.74
BEGIN WoW	–	–	–	0.43	0.53
HaluEval Dial	0.31	0.32	0.31	0.59	0.53
HaluEval QA	0.30	0.54	0.31	0.83	0.81

Table 9: PRAUC scores between rest of the hallucination metrics and human annotations.

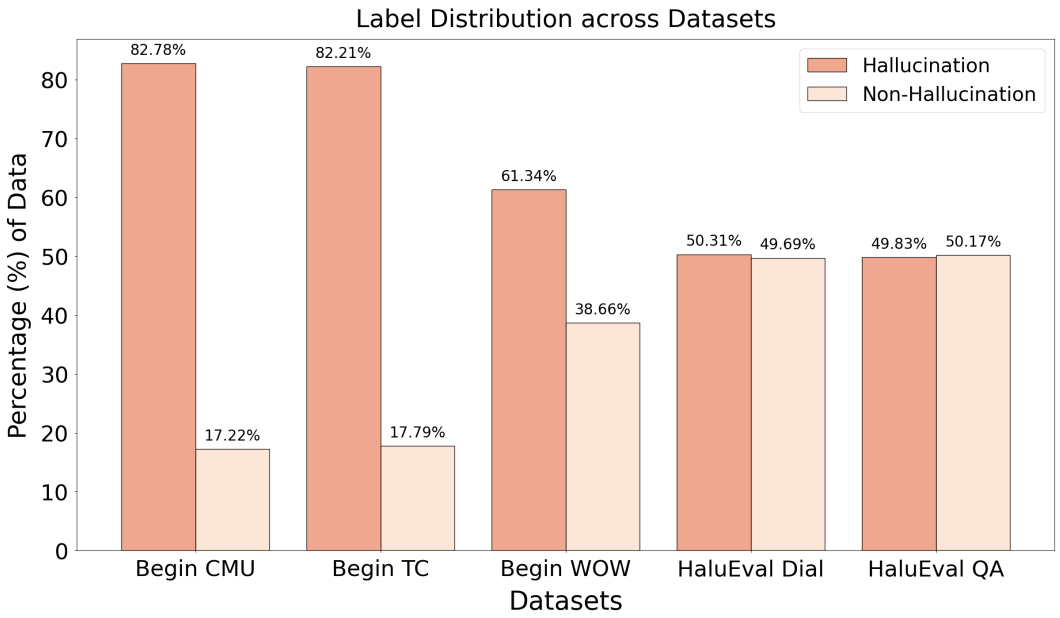

Figure 6: Distribution of hallucinated and non-hallucinated responses in BEGIN and HaluEvl.

Dataset	Critic			GPT-4			Consistency			K-BertScore			Q^2 NLI			Ensemble		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Begin CMU	0.77	0.82	0.77	**0.87**	**0.83**	**0.84**	0.69	0.83	0.75	0.74	0.83	0.75	0.63	0.34	0.40	0.69	0.83	0.75
Begin TC	0.70	0.80	0.74	**0.86**	0.67	0.71	0.68	0.82	0.74	0.68	0.82	0.74	0.63	0.45	0.51	0.68	**0.82**	**0.74**
Begin WoW	0.84	0.83	0.83	0.81	0.77	0.77	0.47	0.61	0.47	0.38	0.61	0.47	0.44	0.46	0.45	**0.85**	**0.85**	**0.85**
HaluEval Dial	0.63	0.56	0.49	**0.77**	**0.74**	**0.74**	0.40	0.40	0.40	0.60	0.60	0.60	0.67	0.65	0.63	0.46	0.46	0.45
HaluEval QA	0.54	0.54	0.53	0.67	0.66	0.66	0.43	0.49	0.36	0.76	0.76	0.76	0.87	0.87	0.87	**0.87**	**0.87**	**0.87**
Average	0.70	0.71	0.67	**0.8**	0.73	**0.74**	0.53	0.63	0.54	0.63	0.72	0.66	0.65	0.55	0.57	0.71	**0.77**	0.73

Table 10: Weighted Precision, Recall, and F1 scores for different metrics on BEGIN and HaluEval for hallucination detection. Green and brown denote the best and second-best metrics, respectively.

concentrate on fluency, semantic similarity, or entity-level accuracy. Because these properties are orthogonal, a model might score well on one metric while performing poorly on another.

GPT-4 based evaluation is a better metric for detecting hallucinations because unlike automated metrics that rely on predefined heuristics (e.g., n-gram overlap, embeddings, or NLI classifiers), GPT-4 can assess nuanced errors, infer missing knowledge, and detect inconsistencies in a way that aligns closely with human judgment, as it considers various factors such as coherence, commonsense reasoning, and factual grounding, to name a few [38]. Here is why it shows a weak correlation with different metrics:

- **GPT-4 vs. N-gram Overlap (Rouge-L, SacreBLEU, and Knowledge-F1):** GPT-4 assesses meaning and factuality beyond simple word overlap, whereas these metrics only measure surface-level similarity. A hallucinated response can have a high n-gram overlap with a reference while still being incorrect, leading to false positives. Conversely, correct but reworded responses can be penalized, leading to false negatives. GPT-4's reasoning capabilities makes it more flexible than rigid n-gram matching.
- **GPT-4 vs. Semantic Similarity (BERTScore and K-BERTScore):** These metrics measure embedding similarity but do not verify factual accuracy. Two sentences can be semantically close while differing in factual correctness. GPT-4 can assess fine-grained factual inconsistencies that semantic similarity models miss, such as incorrect numerical values or subtly misleading statements.
- **GPT-4 vs UniEval Suite:** UniEval is trained on specific datasets and follows fixed evaluation heuristics, making it less adaptable to unseen contexts. GPT-4 dynamically evaluates responses using broad-world knowledge and deeper reasoning, leading to higher accuracy in detecting nuanced hallucinations.

Dataset	Critic	GPT4	Consistency	K-BertScore	Q^2 NLI	Ensemble
BEGIN CMU	2843 / 107	2159 / 791	2949 / 1	22947 / 3	1066 / 1884	2949 / 1
BEGIN TC	3704 / 101	1993 / 1812	3804 / 1	3804 / 1	2069 / 1736	3804 / 1
BEGIN WoW	1957 / 1644	1723 / 1878	3589 / 12	3600/1	2423 / 1178	2181 / 1420
HaluEval Dial	8686 / 1314	6635 / 3365	5492 / 4508	5572 / 4428	6863 / 3137	6352 / 3648
HaluEval QA	4076 / 5924	3712 / 6288	619 / 9381	5125/4875	4680/5320	4991/5009

Table 11: Hallucination detection label distribution (Positive/Negative) for different metrics.

- **GPT-4 vs. Q^2:** It relies on question generation and answer extraction, which introduces cascading errors if the generated questions are poorly framed or if the extraction mechanism fails. Moreover, it may overlook implicit hallucinations that do not map neatly to question-answer pairs, whereas GPT-4 can reason about implicit information.

- **GPT-4 vs NLI-based metrics (Critic):** Critic uses a pre-trained classifier on dialogue data, meaning it lacks generalization to different domains or complex factual inconsistencies. NLI models often misinterpret negations, indirect claims, and paraphrased statements, leading to misclassifications that GPT-4 would avoid.

C.3 Mode-Seeking Decoding Hallucinate less than Sampling-based Approaches

The box plots in Figures 7, 8, and 9 illustrate the performance of various decoding techniques across different metrics. The decoding methods considered include greedy, beam search ($b = 3$), ancestral, top-k ($k = 40$), and top-p ($p = 0.95$). Models are grouped by parameter size into the following bins: $> 0.5, > 4, > 20, > 70$ billion parameters. Overall, greedy and beam search consistently outperform sampling-based methods. However, this trend breaks for BertScore and K-BertScore in the case of FaithDial. We hypothesize that this is possibly due to the model's limited capacity, which may lead to repetitive or degenerate outputs, as observed in previous studies [33]. Other metrics such as Knowledge-F1, Q^2 token F1, MSP, and Perplexity adhere to the trend.

The heatmaps in Figures 10 and 11 show the p-values for pairwise significance tests between the decoding methods. Except for the consistency score, greedy and beam search consistently outperform sampling-based methods with statistically significant results. These findings further confirm that probability-maximization decoding methods help reduce hallucinations, particularly in knowledge-grounded tasks.

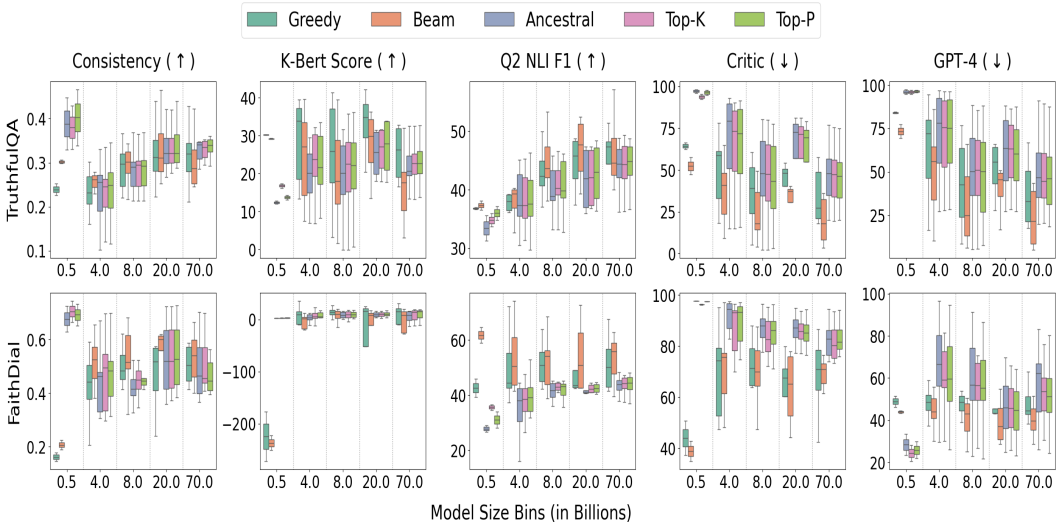

Figure 7: Comparison of factual hallucination metrics across decoding techniques.

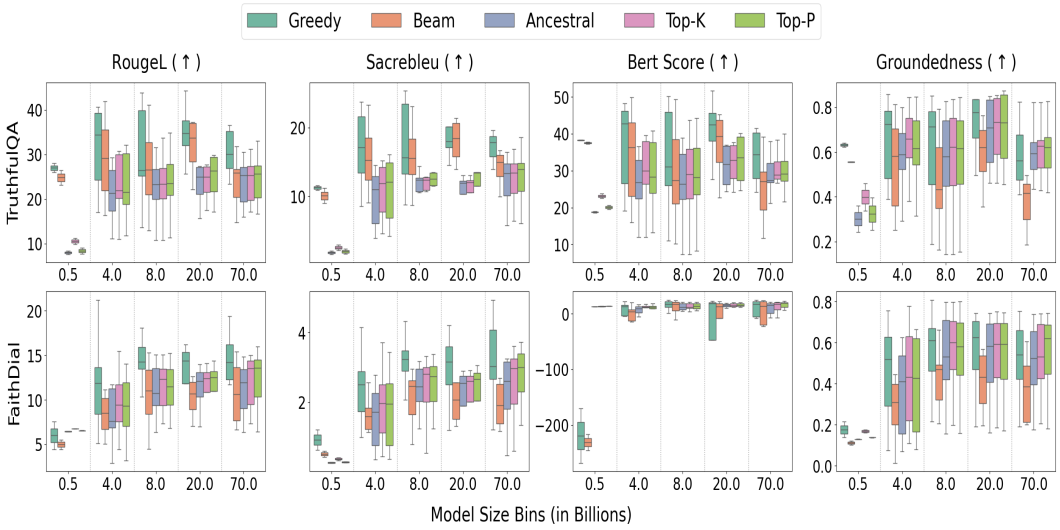

Figure 8: Comparison of traditional NLG metrics across decoding techniques.

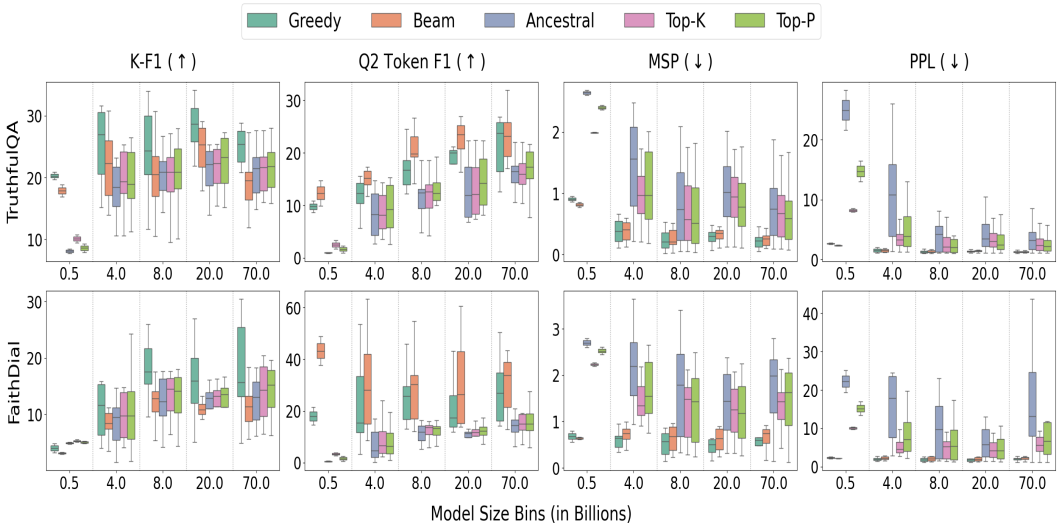

Figure 9: Comparison of uncertainty and token-overlap based hallucination metrics across decoding techniques.

C.4 Parameter Scaling does not Necessarily Improve Hallucination Metrics

Figures 12 and 13 illustrate the performance of NLG, token-overlap, and uncertainty-based metrics as language model parameters scale. While Rouge-L, Q^2 token F1, MSP, and perplexity all improve with model size, other metrics do not show a consistent pattern. Figure 14 presents the p-values for pairwise significance tests across different model sizes, revealing that BertScore shows no improvement as the model size increases.

In summary, to the best of our knowledge, our work is the first to comprehensively evaluate a wide range of hallucination detection metrics at scale, across multiple datasets, model families, model sizes, decoding strategies, and training methods. While Finding 3 have been established in prior studies, such as [42] and [40], they lack the robustness provided by our analysis, as they were not tested across the diverse dimensions that we explore. Our work offers a more thorough and holistic assessment, demonstrating that these findings indeed hold true across different settings and providing deeper insights for ML and NLP practitioners about which metrics perform best under various conditions. Moreover, to the best of our knowledge, none of the previous works have concretely shown the

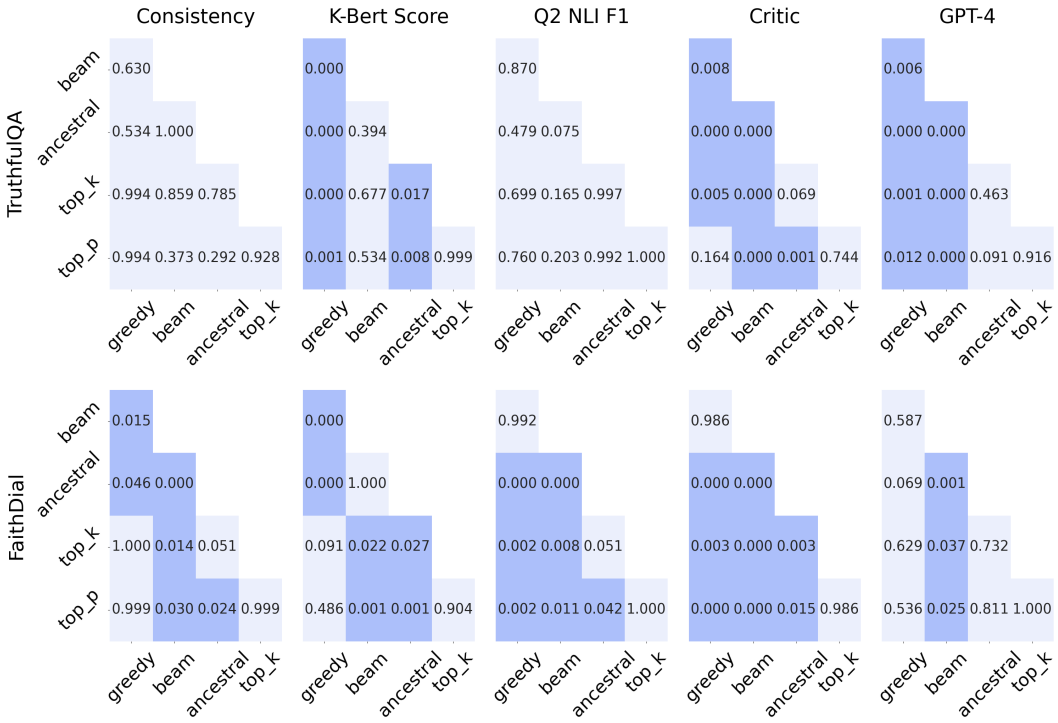

Figure 10: Per-group p-values for decoding techniques using pairwise T-test with Bonferroni correction.

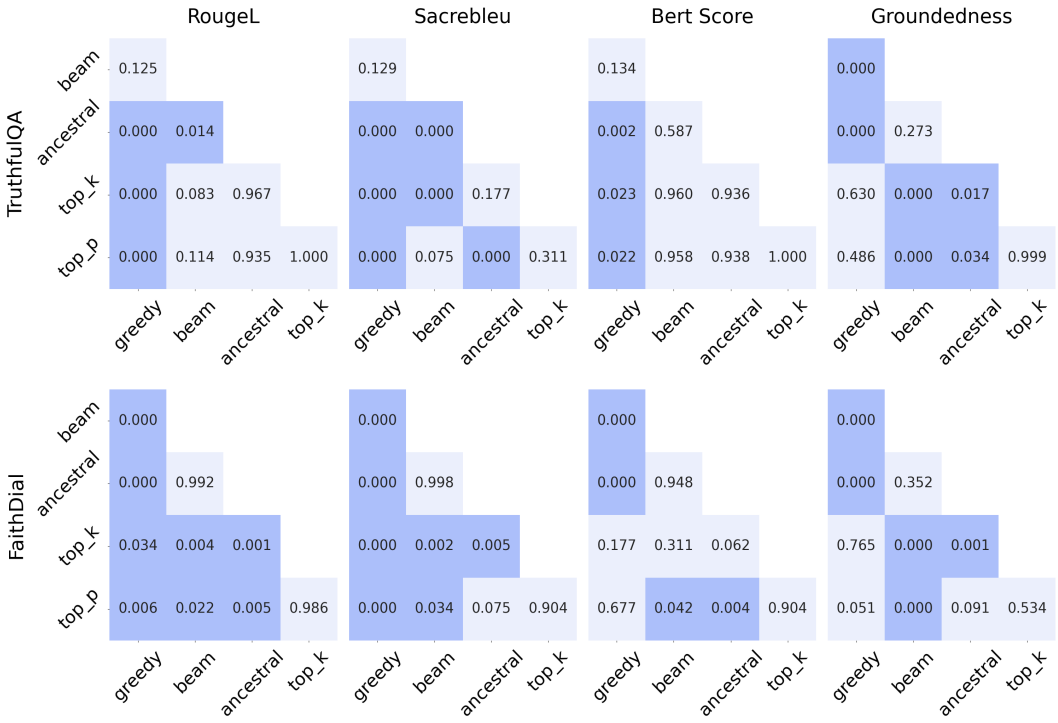

Figure 11: Per-group p-values for different decoding techniques using pairwise T-test with Bonferroni correction.

emergence of finding 4. Lastly, while some of these findings might seem obvious at first, we believe

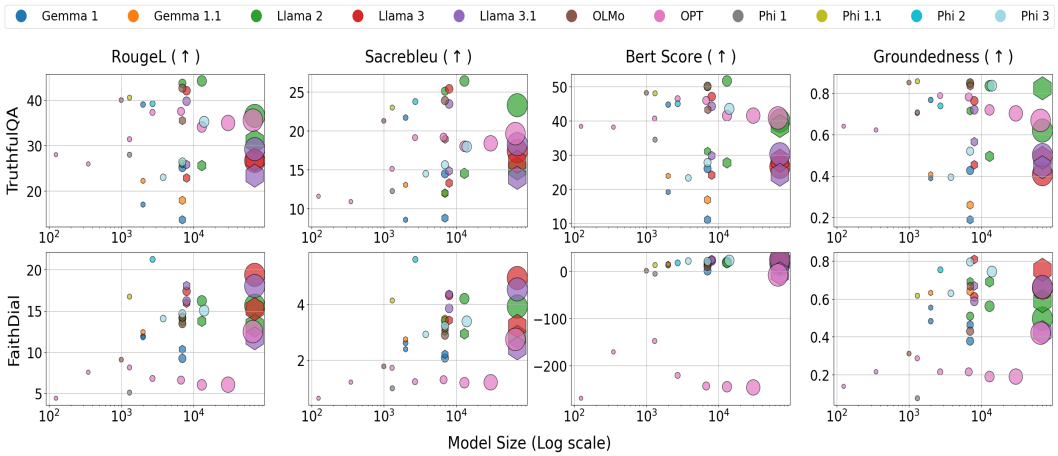

Figure 12: NLG-based hallucination detection metric scores for greedy decoding as model size increases. Circles and hexagons represent pretrained and instruction-tuned models, respectively.

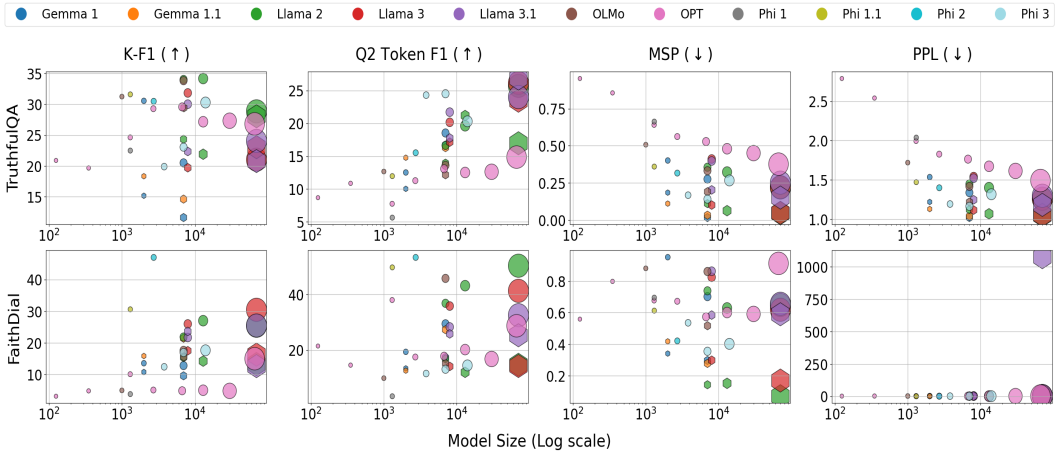

Figure 13: Uncertainty and token-overlap based hallucination detection metric scores for greedy decoding as model size increases. Circles and hexagons represent pretrained and instruction-tuned models, respectively.

scientific research is often exactly around such contributions - transforming intuitive observations into a robust, evidence-backed understanding, advancing the field with concrete, reproducible findings.

D Limitations

While our large-scale empirical investigation provides a thorough analysis of the current hallucination metrics, it does have certain limitations.

Different Evaluation Metrics for Different Datasets. As noted in §2, the BEGIN and HaluEval datasets include model-generated responses with human annotations for hallucinations, whereas TruthfulQA and FaithDial do not. We therefore generate responses for the latter using the models described in S2. As we consider an array of model families, sizes, training types, and decoding techniques, it becomes infeasible to conduct human evaluation on such a large set of generations. Consequently, for Finding 1, we focus solely on the BEGIN and HaluEval. Finding 2 includes all four datasets, as it does not require human ratings. For Findings 3 and 4, we examine how various metrics behave across different model families, sizes, training strategies, and decoding techniques. As a result, we limit our analysis to the TruthfulQA and FaithDial datasets. Additionally, since FaithDial is a modified version of the WoW dataset [47], which is already included in BEGIN, we can reasonably

Figure 14: Per-group p-values for different model size bins using the pairwise Mann-Whitney rank test.

assume that the Findings 3 and 4 results for BEGIN will follow similar trends to those observed for FaithDial.

Other Limitations. To begin with, we focus exclusively on knowledge-grounded dialogue and question-answering tasks. However, hallucination is a prevalent issue across various other NLP tasks, such as machine translation [66], summarization [67], code-generation [68], and linguistic applications [69]. Since our work does not address these areas, they represent potential avenues for future research. Furthermore, while we identify LLM-as-judge as the most reliable metric for hallucination detection, we do not evaluate its variants – such as chain-of-thought prompting [70], G-EVAL [71], or smaller / different architecture LLMs [72] – due to the scope of our study. Lastly, while fine-tuning has been shown to mitigate model hallucinations [73], we have not explored these experiments in our study, leaving them for future investigation.