# Enhancing Hallucination Detection through Perturbation-Based Synthetic Data Generation in System Responses

**Dongxu Zhang, Varun Gangal, Barrett Martin Lattimer, Yi Yang**
ASAPP, Inc.
{dzhang,vgangal,blattimer,yyang}@asapp.com

## Abstract

Detecting hallucinations in large language model (LLM) outputs is pivotal, yet traditional fine-tuning for this classification task is impeded by the expensive and quickly outdated annotation process, especially across numerous vertical domains and in the face of rapid LLM advancements. In this study, we introduce an approach that automatically generates both faithful and hallucinated outputs by rewriting system responses. Experimental findings demonstrate that a T5-base model, fine-tuned on our generated dataset, surpasses state-of-the-art zero-shot detectors and existing synthetic generation methods in both accuracy and latency, indicating efficacy of our approach.

## 1 Introduction

Large Language Models (LLMs) tend to produce hallucinations, wherein the generated text either contradicts the given source knowledge (intrinsic hallucination) or cannot be verified against it (extrinsic hallucination) (Maynez et al., 2020; Rawte et al., 2023). Despite the burgeoning enthusiasm for deploying Generative AI and LLMs in real-world applications, the issue of hallucinations poses significant concerns for downstream users. Consequently, the detection of hallucinations is paramount in enhancing the safety of LLM applications and in fostering trust among users of these technologies.

An effective hallucination detection system should be accurate, fast, and affordable. Cost-effectiveness is crucial because every check for hallucinations adds extra cost to the use of large language models (LLMs), which may already be substantially high. Moreover, the system must possess the flexibility to adapt to the rapidly evolving landscape of LLMs. As shown in Table 1, newer iterations of LLMs generally exhibit enhanced capabilities in mitigating hallucinations, thereby escalating the complexity of the detection challenge. Unfortunately, many current methodologies are either i) costly in terms of compute (Liu et al., 2023; Manakul et al., 2023b) or ii) depend on out-of-domain/external resources such as QA (Honovich et al., 2021; Fabbri et al., 2022) or NLI annotation (Laban et al., 2022; Honovich et al., 2022), potentially compromising performance.

Table 1: Performance evaluation of a GPT-3.5-based zero-shot hallucination detector across different generations of LLMs (see Appendix §E for prompt). This table illustrates a notable decline in detection efficacy when transitioning from older to more recent LLM iterations.

| Hallucination data | LLMs used in the data | F1 |
|---|---|---|
| MNBM ('20) | GPT, Bert, Rnn, ConvNet | 0.780 |
| FRANK ('21) | PointerNet, bertS2S, Bart | 0.694 |
| Seahorse (early '23) | T5, MT5, PALM | 0.576 |
| ScreenEval (late '23) | GPT-4, longformer | 0.130 |

In this study, we introduce a simple yet effective approach for automatically generating synthetic annotations to train hallucination detectors. Figure 1 shows an overview of our approach. The core of our method involves prompting a rewriting LLM to transform a given system response from the target LLM into both faithful and hallucinated versions, respectively. This technique distinguishes itself from existing methods (Gupta et al., 2021; Das et al., 2022b; Li et al., 2023a; Dziri et al., 2022a) in three significant ways. First, unlike traditional methods that rely on human-annotated examples of faithfulness, our strategy is entirely automated, eliminating need for manual annotation. Second, by directly altering responses from the target LLM, our trained detector aligns more closely with the response distribution of the target LLM, facilitating seamless adaptation to new LLMs. Lastly, while previous approaches require predefined information about the types of hallucinations for their generation process, our method operates without such assumptions. This allows



*Findings of the Association for Computational Linguistics ACL 2024*, pages 13321–13332

for the creation of a broader spectrum of hallucination types, enhancing the coverage and diversity of generated hallucinations.

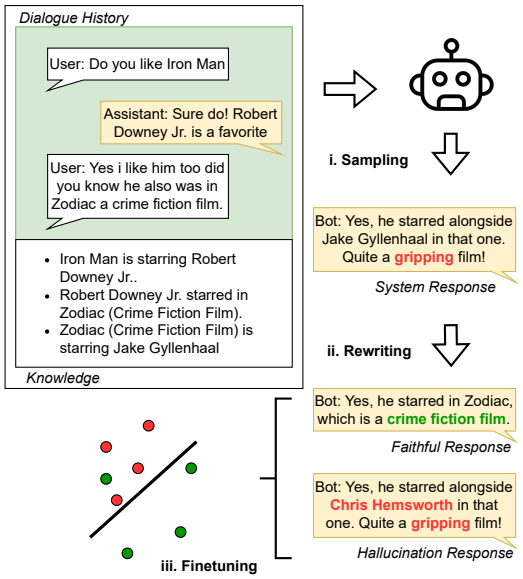



Figure 1: Overview of our automatic hallucination generation pipeline. Red and green highlights hallucinated and faithful claims.

Our experimental evaluations span two hallucination detection datasets, OpenDialKG (Moon et al., 2019) and BEGIN (Dziri et al., 2022b), where a T5-base model, fine-tuned with our novel data generation approach, significantly surpasses GPT-4 based methods in performance while achieving a tenfold increase in speed. Further analysis of the generated hallucinations uncovers previously unreported patterns, such as "adding attributes", expanding the discourse on hallucination beyond existing literature. Our code and data will be available at https://github.com/asappresearch/halugen.

## 2 Methodology

In this section, we detail our methodology for generating synthetic hallucinations that closely mimic those encountered in real-world applications of Large Language Models (LLMs). Prior approaches to hallucination generation have primarily relied on rewriting human-authored texts (Das et al., 2022b; Li et al., 2023a) or introducing perturbations to the knowledge source (Gupta et al., 2021; Dziri et al., 2022a; Zhang et al., 2023). However, these methods often yield outputs that diverge significantly from those produced by LLM systems, leading to a substantial discrepancy between the synthetic hallucinations and the genuine hallucinations observed in practice. To address this gap, our approach involves prompting a rewriting LLM to perturb the responses of the LLM system itself, rather than those written by humans. This strategy draws inspiration from the "Minor perturbation" technique described by Lucas et al. (2023), adapted to our context to ensure the synthetic hallucinations closely align with the expected data in real-world deployments.

To effectively train a hallucination detector, it is imperative to have access to both hallucinated and faithful responses. Unlike previous studies, where human-curated outputs served as the benchmark for faithful system outputs (Das et al., 2022b; Li et al., 2023a; Dziri et al., 2022a), the responses obtained directly from the target LLM system may contain a considerable proportion of non-faithful responses. To overcome this challenge, we employ the rewriting LLM to adjust the system's responses in a manner that promotes the generation of faithful outputs. The specific prompts utilized for inducing both hallucination and faithfulness are presented in Appendix §A. It is important to note that our process for generating hallucinations did not involve biasing the system with predefined categories of hallucination within the prompt, ensuring a more authentic and unbiased generation process.[1] For the rewriting LLM, we selected GPT-4 [2] due to its robust capabilities in text rewriting (Madaan et al., 2024). Leveraging a powerful rewriting LLM like GPT-4 enables the exploration of a wider array of hallucination categories, thereby enhancing the coverage of hallucinations that are likely to be encountered in real-world scenarios.

## 3 Experiments

### 3.1 Datasets

**OpenDialKG** is a dialogue dataset that was adopted by HaluEval (Li et al., 2023a), a recent benchmark for hallucination detection. OpenDialKG features human-generated dialogues exclusively with supporting knowledge sources from Freebase (Bollacker et al., 2008). In order to leverage the dataset for hallucination detection, we simulate a chatbot system by employing GPT-4 to generate responses grounded in both the provided knowledge and the preceding dialogue context. The

[1]These prompts have been designed with versatility in mind, allowing for straightforward adaptation to other NLP tasks such as question answering and summarization. However, our current investigation is focused exclusively on knowledge-grounded dialogues.

[2]We use gpt-4-1106-preview for our experiments.

specifics of the prompt template utilized for this simulation are detailed in Appendix B. To create a evaluation set on the generated responses, we employ Amazon Mechanical Turk annotators to evaluate whether the responses from the simulated chatbot system were fully supported by the dialogue history and the provided knowledge (for detailed annotation guidelines and interface, refer to Appendix D). Our collection (OpenDialKG-Eval) comprises 402 annotated responses. We designated responses with high-confidence labels as our test set and utilized the remainder for development purposes, resulting in 312 test responses and 90 for development. More details of OpenDialKG-Eval can be found in Appendix E. In addition, we simulate another 2000 responses from OpenDialKG for synthetic generation purpose.

**BEGIN** is a knowledge-grounded dialog dataset featuring 12k responses from four dialogue systems distributed over 3 document-scale knowledge domains – Wizard of Wikipedia (Dinan et al., 2018), TopicalChat (Gopalakrishnan et al., 2023) and DoG (Zhou et al., 2018) — all with mean knowledge snippets longer than OpenDialKG. In addition, there are three response categories in BEGIN: Fully attributable, Not fully attributable, Generic. Generic category refers to response that are vague and do not provide any new information. Therefore, in addition to faithful and hallucination generation, we also ask LLM to generate responses under "Generic" category. The detailed prompt can be found in Appendix A Table 9. Since BEGIN only released the Dev and Test split, we adopt 1,228 system responses from Dev for both synthetic generation and development while reporting results on Test split.

### 3.2 Baselines

**Zero-shot Detection** We compare with SelfCheckGPT (Manakul et al., 2023a), a consistency-based approach which samples system responses multiple times in temperature 1.0 and then leverage scores from NLI or QA to measure whether the target response is consistent with these samples.

Another baseline is G-Eval (Liu et al., 2023), which prompts GPT-4 with an annotation-rubric style prompt describing target variable and furthermore draws multiple samples at a higher temperature; emulating diverse multi-annotation by humans. Since both G-Eval and SelfCheckGPT can only output scores between 0 and 1 and BEGIN data has three output categories, we compare GPT-4 (Internal), our self-devised zero-shot detector, which prompts GPT-4 with an intuitive prompt to enable three-way outputs(Appendix§F) and does greedy decodes to generate a binary/ternary answer.

The last zero-shot baseline we compare with is SCALE (Lattimer et al., 2023), NLI-based approach which first decomposes the supporting context into chunks, calculate NLI scores on the chunk level using FlanT5 (Chung et al., 2022), then use the maximum score as the final prediction of factual consistency.

**Detection with End-to-end Finetuning** We use T5-base, an encoder-decoder LM with 223M parameters, as the base model of the detector and fine-tune it on multiple synthetic datasets.[3] We make our best efforts to conduct apple-to-apple comparison among different synthetic data. On OpenDialKG-Eval benchmark, we compare with FADE (Das et al., 2022b) and HaluEval (Li et al., 2023a), where we adopted their existing synthetic hallucinations as negative and human written responses from OpenDialKG as positive data for training. On BEGIN dataset, we compare with Aug-WOW (Gupta et al., 2021) and BEGIN-Adv. (Dziri et al., 2022a), both are synthetic generation baselines and their performances on BEGIN has been reported by (Dziri et al., 2022a). For more details of these synthetic data generation baselines, please refer to Section 5

### 3.3 Results

Table 2: Macro-F1 and latency of hallucination detection methods over OpenDialKG-Eval.

| | F1 | Latency |
|---|---|---|
| Zero-shot Detection | | |
| SelfCheckGPT (QA) (Manakul et al., 2023a) | 0.536 | 60.59 sec |
| SelfCheckGPT (NLI) (Manakul et al., 2023a) | 0.579 | 0.93 sec |
| G-Eval (Liu et al., 2023) | 0.608 | 2.79 sec |
| SCALE$_{XL}$ (Lattimer et al., 2023) | 0.687 | 0.22 sec |
| T5-base Finetuned over Synthetic Data | | |
| FADE (Das et al., 2022b) | 0.625 | 0.20 sec |
| HaluEval (Li et al., 2023a) | 0.702 | 0.20 sec |
| Our approach | 0.762 | 0.20 sec |

Table 2 shows the performance of hallucination detection and latency per response on OpenDialKG-Eval. Latencies are profiled over AWS g5.xlarge instances with no batching sae for G-Eval which requires OpenAI API access. From

[3]For more experimental details, please refer to Appendix H.

the results, our approach not only out-performs T5 detectors finetuned over previous hallucination generation baselines, but more interestingly, it out-performs state-of-the-art zero-shot detection methods. Besides performance, finetuned models achieve significantly lower latency than all zero-shot baselines. We also show the results on BEGIN data. The results can be found in Table 3, where similar observation can be found.

Table 3: Macro-F1 and latency of hallucination detection over BEGIN test split with three-class classification.

| | F1 | Latency |
|---|---|---|
| Zero-shot Detection | | |
| GPT-4 (Internal) | 0.323 | 1.13 sec |
| T5-base Finetuned over Synthetic Data | | |
| AugWow (Gupta et al., 2021) | 0.378 | 0.20 sec |
| BEGIN-Adv. (Dziri et al., 2022a) | 0.459 | 0.20 sec |
| Our approach | 0.473 | 0.20 sec |

Lastly, average cost per synthetic response generation is 0.008 USD on OpenDialKG and 0.006 USD on BEGIN, using *gpt-4-1106-preview*. In comparison, average cost of human annotation per example for OpenDialKG-Eval is 0.20 USD.

### 3.4 Ablation Study

To analyze the significance of both hallucination and faithful response generation, we conduct an ablation study to replace one of the generation using system response. Results are shown in Table 4. Results show that both categories of synthetic data are necessary to effectively fine-tune the detector.

Table 4: Results of ablation study. "pos-F1" and "neg-F1" represents F1 performance over faithful and Hallucination labels separately.

| Approach | pos-F1 | neg-F1 | F1 |
|---|---|---|---|
| Our approach | 0.812 | 0.713 | 0.762 |
| w/o faithful generation | 0.747 | 0.618 | 0.683 |
| w/o hallucination generation | 0.517 | 0.502 | 0.509 |

## 4 Hallucination Pattern Analysis

### 4.1 Hallucination Pattern Analysis

Previous work usually predefined hallucination patterns such as replacing or swapping entities (Das et al., 2022a; Li et al., 2023a). We randomly sample 144 hallucinations generated by our method over OpenDialKG dataset, and manually annotate these into a taxonomy of 6 distinct pattern-driven categories characterizing the pattern surfaced in the hallucination, further described in Appendix §C.

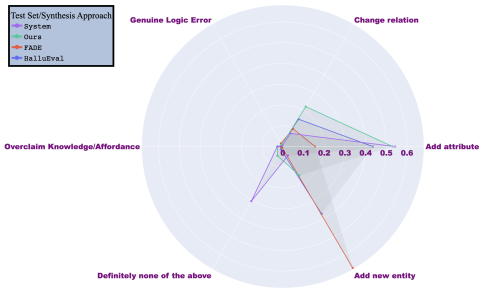



Figure 2: Spiderplot spider-web traces visualizing how the synthesized hallucinations from our approach (in *green*) + two baselines (HALUEVAL,FADE, in *red* and *blue*) as well as the system response distribution (SYSTEM,in *purple*) distribute over the 6 qualitative categories as laid out in §4.1. Both HALUEVAL (blue) and FADE (red) show a marked skew towards "Add new entity", while OURS (green) shows a closer alignment with the SYSTEM (purple).

Table 5: Hallucination patterns appeared in OpenDialKG-Eval and our synthetic generated data for finetuning.

| Pattern name | System | HaluEval | FADE | Ours # |
|---|---|---|---|---|
| Adding attribute to an entity | 0.540 | **0.435** | 0.156 | 0.530 |
| Adding or updating relation | 0.070 | 0.150 | 0.099 | 0.220 |
| Addding new entities | 0.050 | **0.370** | **0.675** | 0.160 |
| Overclaim knowledge/affordance | 0.027 | 0.011 | 0.010 | 0.025 |
| Inference error beyond above | 0.004 | 0.011 | 0.018 | 0.010 |
| None of the above | 0.310 | 0.016 | 0.042 | 0.050 |
| KL(●, System) | - | 0.671 | 1.527 | 0.340 |

Pattern distributions are both listed in Table 5 and Figure 2. From the pattern distribution, it is interesting to see that our method has fewer hallucinations from entity replacing/swapping, the most dominant hallucination type is adding unverifiable attributes to an entity. This indicates that our methods generate responses which conform tighter to the real hallucination distribution in contrast to prior approaches. The KL Divergence between the categorical pattern distribution of our method and the system response based distribution is 0.3395, compared to the much greater 0.6706 (and 1.52) between the distribution of HaluEval (and FADE) vs the latter.

### 4.2 Quality Analysis of Synthetic Data Generation

To more closely evaluate the effectiveness of rewriting, we did human annotation over 100 randomly sampled system responses along with our synthetically generated responses based on these system responses. Table 6 shows the portion of faithful data within each type of responses:

Upon reviewing the annotations, our method

Table 6: Human analysis of faithfulness of our generated synthetic responses in comparison to system outputs.

| | Faithfulness |
|---|---|
| System output | 41% |
| Faithful generation | 51% |
| Hallucination generation | 5% |

demonstrates a significant reduction in the generation of unfaithful responses (hallucinations) compared to system outputs. Our approach also yields a higher number of faithful responses compared to the baseline system outputs, aligning with our objectives. Specifically, out of 59 instances of hallucinations identified in system outputs, our faithful generator converts 11 into faithful responses, achieving a conversion rate of approximately 19%. Conversely, among the 41 faithful responses generated by the system, our method inadvertently transformed only one into a hallucination.

## 5 Related Work

Research on generating synthetic annotations for hallucination detection has explored various strategies. Some approaches, like FADE (Das et al., 2022b) and HaluEval (Li et al., 2023a), manipulate human-written texts by altering entities or applying predefined hallucination criteria, respectively. These methods only focus on introducing hallucinations and ignore faithfulness augmentation, assuming human-generated content to be inherently accurate, which maybe untrue. Other studies focus on modifying the knowledge source before response generation. AugWow (Gupta et al., 2021) introduces hallucinations by using irrelevant or no evidence, while BEGIN-Adv (Dziri et al., 2022a) alters subjects, objects, named entities, or verbs in the source material, prompting a GPT2-based system (Radford et al., 2019) for response regeneration. These techniques, however, might lead to predictable hallucination patterns due to their reliance on predefined rules.

More recently, ICD (Zhang et al., 2023) mitigates LLM hallucinations by finetuning a model on non-factual samples, aiming to down-weight factually weak predictions. Despite its novelty, the reliance on entity perturbation for generating non-factual samples could limit the coverage of detected hallucinations. HaluEval-Wild (Zhu et al., 2024) aims to evaluate LLM hallucination in human-LLM interactions. Their approach first collects challenging user queries which can lead to hallucinated LLM responses. Faithful reference responses are generated using GPT4 with retrieval augmentation. While the generated data is challenging, it is not obvious how to adapt the approach for customized tasks. Interestingly, Li et al. (2023b) observed that the effectiveness of the LLM-generated synthetic data in supporting model training is negatively correlated with the subjectivity of the target task.

## 6 Conclusions

In this work, we aim to address the prevalent challenge of training data for hallucination detection being either unavailable or expensive to curate. We hypothesize that this can be addressed via a framework that automatically synthesizes both hallucinated and faithful responses using a prompt-based method. Our experimental results on two datasets verify effectiveness of our approach and show it compares favourably against several baselines, including those using prompt-based synthesis.

## 7 Limitations

In this work, the quality of the synthetically generated data is partially determined by the capability of prompted LLM. However, this issue is not severe since our goal is to facilitate the fine-tuning process of the hallucination detection model rather than using the data for evaluation.

In addition, note that hallucinations generated in this work, may still be different from those exist unintentionally in system outputs. One promising future work is to explore unintentional hallucination/faithful output generation, such as evaluating output faithfulness via sampling, or perturbing the prompt so that it becomes more or less likely to induce hallucinations.

Lastly, since we are encouraging the LLM to generate hallucinations, there is a risk of introducing misinformation into the real world data, which is also a common issue for large language model generation in general. We encourage people to follow policies and strategies with regarding to data sourcing, fact checking, etc. in order to mitigate such issue.

## References

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human

knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Souvik Das, Sougata Saha, and Rohini Srihari. 2022a. Diving deep into modes of fact hallucinations in dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 684–699, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

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

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

Alexander Richard Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601.

Joseph L Fleiss, B Levin, and MC Paik. 1981. Statistical methods for rates and proportions. john wiley & sons. *New York*, 870.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. $q^2$: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Barrett Lattimer, Patrick CHen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1691–1703.

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

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023b. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.

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

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee.

2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. 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.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023b. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv e-prints*, pages arXiv–2303.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

Zhiying Zhu, Yiming Yang, and Zhiqing Sun. 2024. Halueval-wild: Evaluating hallucinations of language models in the wild.

## A Prompt Template for Synthetic Response Generation

Table 7 and Table 8 include the prompt templates to generate hallucinated responses and faithful responses.

For BEGIN dataset, we also create a prompt to generate "Generic" responses, as shown in Table 9

## B Prompt Template for Simulating Chatbot on OpenDialKG

Table 10 contains the prompt template that we use to prompt GPT-4 for system responses on OpenDialKG.

## C Rubric/Typology for Qualitative Annotation

For the qualitative annotation in Table 4 of the main body, we use the rough definitions/guidelines below. We formulate these types based on prior work on hallucination and hallucination typology such as FRANK.

- Type No 1 : Adding attribute to entity, or adding new value to a known entity.
- Type No 2 : Changing or misspecifying the relation between two entities, or interchanging and swapping their roles w.r.t the same relation.
- Type No 3 : Adding new entities in place of an existing entity, or even otherwise, and mentioning any information about them leaving aside one that purely expresses a no-information stance
- Type No 4 : Mistakenly claiming knowledge or committing to action about something that the model doesnt really know or cannot act upon
- Type No 5 : A genuine error in the logic and inference beyond just new entities, misattributed or swapped roles and relations.
- Type No 6: Definitely none of the above, it is something else

## D AMT Annotation Guidelines, Setup and Template

This section describes the AMT annotation guidelines for OpenDialKG-Eval.

A snapshot of the template instructions as seen for an actual example can be viewed in Figure 3. Furthermore, we enclose the complete annotation template [including rules and illustrative examples in its contents] in the form of a single .html file included in the Supplementary Materials along with this submission.

Annotators were restricted to be from Anglophone countries (USA, UK, Australia and New Zealand) to ensure a good likelihood of them being native speakers. Further, annotators were restricted to be from among those with a prior approval rate of atleast 98%.

Annotators were compensated fairly at a rate of 9.3$ per HIT per hour which is well over the minimum wage of 7.25$ per hour in the U.S.A as per Department of Labour estimates for 2023.

We also provide due warning to the annotators not to even inadvertently share any PII or personal information and this is in no way required for our task. We also assure them that time taken etc [nothing beyond the task pertinent annotation] will be used or shared. The disclaimer we include in the template is "Important Disclaimer: Please avoid sharing any personal details or information including PII or demographics anywhere in this study. We will also not be sharing how much time you took to solve this, or what your individual experience profile was. We will merely be using the judgements made about aspects of generated output in relation to input. No other data implicitly or explicitly collected will be shared."

## E Statistics of OpenDialKG-Eval

OpenDialKG-Eval comprises 180 faithful generations and 132 hallucinations, while the development set contains 39 faithful generations and 51 hallucinations. The inter-annotator agreement, measured by Cohen's, stands at 0.583, indicating a moderate level of agreement according to Fleiss's guidelines (Fleiss et al., 1981) on interpreting Cohen Kappa magnitude.

For the annotation process, we utilized a scale ranging from -2 to +2, excluding 0. Here, -2 represents strong hallucination, and +2 signifies strong faithfulness. Based on this scale, instances scoring greater than 1 or less than -1 were allocated to the test set, with the remaining instances assigned to the development set.

Table 7: Prompt template to generate hallucinated responses.

| |
|---|
| Take a deep breath and work on this problem step by step.<br>I want you act as a chatbot in a conversation with human. Your job is to edit a detail in the True Response and generate a Hallucinated Response that is inconsistent with the Dialogue History and Knowledge.<br>- Valid edit actions include removing, replacing or adding a short piece of information to the True Response.<br>- If the True Response is faithful, please edit it to generate a Hallucinated Response.<br>- If the True Response has already contained hallucination, please edit it to generate an adversarial Hallucinated Response that are more difficult to be detected.<br>- The generated Hallucinated Response should be ambiguous or complex or non-trivially implicit to be detected by a human who has access to all the Knowledge and Dialogue History.<br>- The generated Hallucinated Response should contain similar number of words as the True Response. Do not make it lengthy.<br><br>#Knowledge#: {Instructional prompt for target system}<br>#Dialogue History#: {dialogue history}<br>#True Response#: {system output}<br>Now, please generate your hallucinated response:<br>#Hallucinated Response#: |

Table 8: Prompt template to generate faithful responses.

| |
|---|
| Take a deep breath and work on this problem step by step.<br>I want you act as a chatbot in a conversation with human.<br>Given a Response that contains hallucination, your job is to edit the Response lightly and generate a faithful Response that is fully supported by with the Dialogue History and Knowledge.<br>- Valid edit actions include removing or replacing a short piece of information to the Response.<br>- Every token of the generated Response should be strictly verifiable by the Knowledge and Dialogue History. Even commonsense information needs to be verifiable.<br>- Please keep the similar writing style as the Response. Do not make your response lengthy.<br><br>#Knowledge#: {Instructional prompt for target system}<br>#Dialogue History#: {dialogue history}<br>#Response#:{system output}<br><br>Now, please generate your faithful response:<br>#Faithful Response#: |

## F Prompt for Motivating Zero-Shot Detector Experiment in Intro Table 1

Prompts are shown in Table 11.

## G Prompt for GPT-4 (Internal) Zeroshot Approach

Prompts are shown in Table 12

## H Experimental Details

During finetuning, we use batchsize = 4, apply AdamW gradient descent (Loshchilov and Hutter, 2018) and tune learning rates from the range of [1e-3, 1e-4, 1e-5]. We run 5 epochs for OpenDialKG and 1 epoch over the BEGIN dataset. We evaluate our model, HaluEval-based and FADE-based finetuned models using Dev set, choose the best performing learning rate and report the performances on test set. In addition, we adopt Low Rank Adaptation (Hu et al., 2022) with r=16, $\alpha$=32, and target_modules=["q", "v"] during optimization. Our experiments are base on

BEGIN-Adv. has 8k unreleased data for training, while there are only 1.2k data in BEGIN dev that we can leverage for fine-tuning. In order to generate similar amount of training data, we generate 3 synthetic responses per category for each example in BEGIN Dev set. We adopt temperature 0.5 to avoid repeat generation.

Wherever pertinent, we provide mean results over two random runs.

## I Additional Arguments for Cost Efficiency of our Method

We re-emphasize some key additional points here regarding relative cost efficiency of our approach:

1. Inference Time Efficiency: Our findings, presented in Table 2 and Table 3, demonstrate that a fine-tuned T5-base classifier offers significantly lower latency compared to other base-

Table 9: Prompt template to generate 'Generic' responses for BEGIN dataset.

| |
|---|
| Take a deep breath and work on this problem step by step.<br>I want you act as a chatbot in a conversation with human.<br>Given a Response, your job is to rewrite it such that it is ostensibly about the same topic as the Response but becomes vague and does not contain any factual statement.<br>Examples of rewritten Response includes but not limited to back-channeling, expressing uncertainty, or diverting the conversation from ambiguous or controversial topics. Do not make your response lengthy.<br><br>#Knowledge#: {Instructional prompt for target system}<br>#Dialogue History#: {dialogue history}<br>#Response#: {system output}<br>Now, please generate your faithful response:<br>#Rewritten Response#: |

Table 10: Prompt template to simulate the chatbot system for OpenDialKG.

| |
|---|
| Take a deep breath and work on this problem step by step.<br>Given a Dialogue History and Knowledge, your job is to follow instructions in the Knowledge and generate a faithful Response based on the Knowledge and Dialogue History.<br>#Knowledge#:<br>You are a chatbot. Your goal is to continue the conversation by responding to user's last utterance.<br><br>You have the following knowledge that can be used to generate your response:<br>{KG knowledge}<br>#Dialogue History#:<br>{dialogue history}<br>Now, please generate your response:<br>#Response#: |

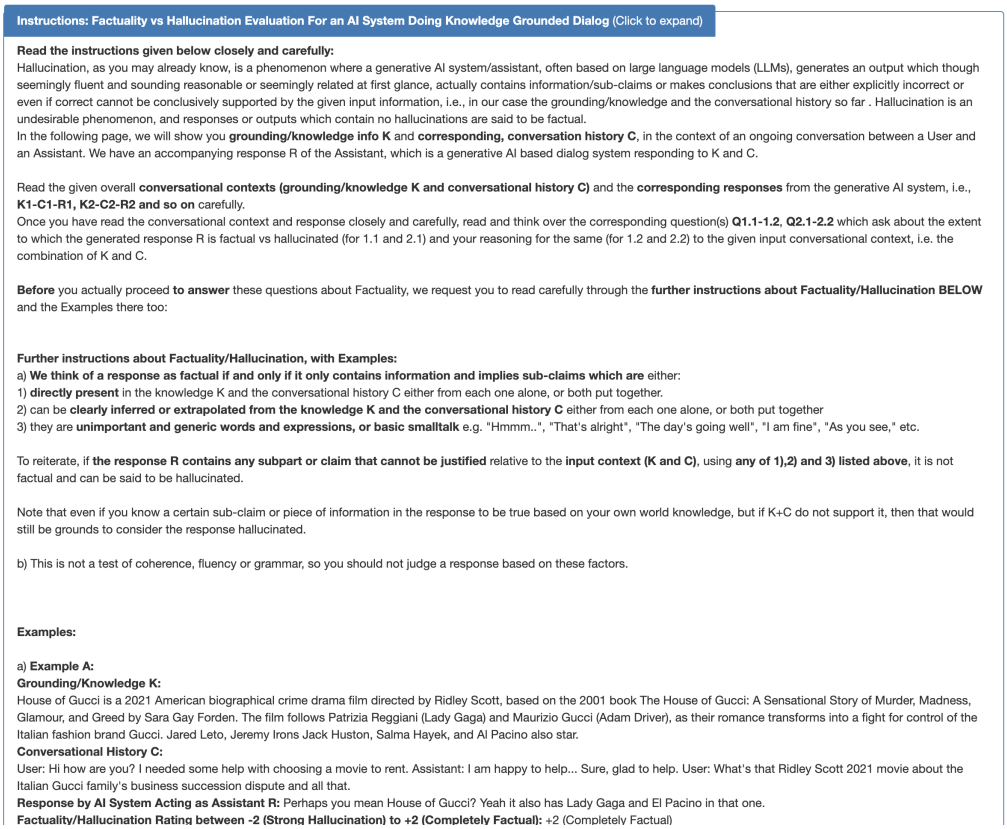
line methods. This is crucial for applications where high latency is not viable.

2. An additional pragmatic aspect bolstering our relative cost efficiency angle in the long term is the steadily decreasing per-token costs of API-based colossal LLMs like GPT-4, driven by widening adoption, faster accelerators, better compression/distillation and quantization etc. Specifically for GPT-4, we have in the past year seen three major price decrease events, each by a factor of 2+, on June 13th, November 11th and Jan 25th, causing a price decrease of over 8 fold. In contrast, human annotation costs in dollar terms rise albeit very gradually.

## J Out of Domain Evaluation

We conducted additional experiments to evaluate our model's performance in out-of-distribution scenario. Specifically, we utilized the best-performing T5 models, which were finetuned using our dataset generated from OpenDialKG, and applied the model to the BEGIN test set. And we compared with our internal GPT-4 baseline.

Given that the BEGIN dataset features significantly longer knowledge contexts in natural language text, as opposed to the list of triplets from the Freebase knowledge graph found in OpenDialKG, this setup simulates a substantial out-of-distribution evaluation. To adapt our binary classifiers for use with BEGIN, we collapse the "generic" and "not fully attributable" labels into a single "hallucination" category, resulting in a modified binarized test set. The results are as follows in Table 13, which indicates that our approach leads to significantly better performance in comparison to a naive random baseline, and shows competitive performance with GPT-4 based zero-shot detection in out-of-distribution scenario, emphasizing its generalization ability.



**Instructions: Factuality vs Hallucination Evaluation For an AI System Doing Knowledge Grounded Dialog** (Click to expand)

**Read the instructions given below closely and carefully:**
Hallucination, as you may already know, is a phenomenon where a generative AI system/assistant, often based on large language models (LLMs), generates an output which though seemingly fluent and sounding reasonable or seemingly related at first glance, actually contains information/sub-claims or makes conclusions that are either explicitly incorrect or even if correct cannot be conclusively supported by the given input information, i.e., in our case the grounding/knowledge and the conversational history so far . Hallucination is an undesirable phenomenon, and responses or outputs which contain no hallucinations are said to be factual.
In the following page, we will show you **grounding/knowledge info K** and **corresponding, conversation history C**, in the context of an ongoing conversation between a User and an Assistant. We have an accompanying response R of the Assistant, which is a generative AI based dialog system responding to K and C.

Read the given overall **conversational contexts (grounding/knowledge K and conversational history C)** and the **corresponding responses** from the generative AI system, i.e., **K1-C1-R1, K2-C2-R2 and so on** carefully.
Once you have read the conversational context and response closely and carefully, read and think over the corresponding question(s) **Q1.1-1.2**, **Q2.1-2.2** which ask about the extent to which the generated response R is factual vs hallucinated (for 1.1 and 2.1) and your reasoning for the same (for 1.2 and 2.2) to the given input conversational context, i.e. the combination of K and C.

**Before** you actually proceed **to answer** these questions about Factuality, we request you to read carefully through the **further instructions about Factuality/Hallucination BELOW** and the Examples there too:

**Further instructions about Factuality/Hallucination, with Examples:**
a) **We think of a response as factual if and only if it only contains information and implies sub-claims which are** either:
1) **directly present** in the knowledge K and the conversational history C either from each one alone, or both put together.
2) can be **clearly inferred or extrapolated from the knowledge K and the conversational history C** either from each one alone, or both put together
3) they are **unimportant and generic words and expressions, or basic smalltalk** e.g. "Hmmm..", "That's alright", "The day's going well", "I am fine", "As you see," etc.

To reiterate, if **the response R contains any subpart or claim that cannot be justified** relative to the **input context (K and C)**, using **any of 1),2) and 3) listed above**, it is not factual and can be said to be hallucinated.

Note that even if you know a certain sub-claim or piece of information in the response to be true based on your own world knowledge, but if K+C do not support it, then that would still be grounds to consider the response hallucinated.

b) This is not a test of coherence, fluency or grammar, so you should not judge a response based on these factors.

**Examples:**

a) **Example A:**
**Grounding/Knowledge K:**
House of Gucci is a 2021 American biographical crime drama film directed by Ridley Scott, based on the 2001 book The House of Gucci: A Sensational Story of Murder, Madness, Glamour, and Greed by Sara Gay Forden. The film follows Patrizia Reggiani (Lady Gaga) and Maurizio Gucci (Adam Driver), as their romance transforms into a fight for control of the Italian fashion brand Gucci. Jared Leto, Jeremy Irons Jack Huston, Salma Hayek, and Al Pacino also star.
**Conversational History C:**
User: Hi how are you? I needed some help with choosing a movie to rent. Assistant: I am happy to help... Sure, glad to help. User: What's that Ridley Scott 2021 movie about the Italian Gucci family's business succession dispute and all that.
**Response by AI System Acting as Assistant R:** Perhaps you mean House of Gucci? Yeah it also has Lady Gaga and El Pacino in that one.
**Factuality/Hallucination Rating between -2 (Strong Hallucination) to +2 (Completely Factual):** +2 (Completely Factual)

Figure 3: A snapshot of how the initial instructions and examples section of the template would appear to an annotator doing a HIT for our annotation task.

Table 11: Prompt template of zero-shot hallucination detector GPT-3.5-turbo for Table 1.

| |
|---|
| <DocumentGivenToAISystem>: {Input/Document}</DocumentGivenToAISystem><br><SummaryByAISystem>: {System Output}</SummaryByAISystem><br>Is the output Summary generated by the AI System Faithful to the Document given to it?<br>Or is it Hallucinated? (Answer with +1 for Faithful or -1 for Hallucinated): |

Table 12: Prompt for GPT-4 (Internal) Zeroshot Approach (The Ternary version with Generic, the binary one omits the part concerned with Generic class)

| |
|---|
| <PromptGivenToExtBot>: {Knowledge}</PromptGivenToExtBot><br><ConvHistoryBetweenUserAndExtBot>: {System Output}</ConvHistoryBetweenUserAndExtBot><br><ResponseByExtBot>: {System Output}</ResponseByExtBot><br>The Response here can be either Faithful to the Context (Prompt and ConvHistory) OR it ecan be hallucinated/contain hallucinations (says something that is contradictory or not<br>entirely or close to likely supported by the context).<br>A third possibility is that it says something really generic and not really having a relevant truth value or sufficient relatibility to context, such as smalltalk, obviously<br>true statements amongst other things.<br>Thus a Response can be Faithful, Hallucinated or Generic w.r.t the Prompt given to it and the ConvHistory.<br>Is the output Response given by the ExtBot Faithful to the Prompt given to it and the ConvHistory between User and ExtBot so far?<br>Or is it Hallucinated? Or is it Generic?<br>(Answer with 2 for Faithful, 1 for Generic or 0 for Hallucinated): |

Table 13: Detection Results on Out-of-Distribution Data.

| Approach | Macro-F1 |
|---|---|
| Random baseline | 0.455 |
| Our approach | 0.518 |
| GPT-4 (Internal) | 0.543 |