SEQUENCE-LEVEL CERTAINTY REDUCES HALLUCINATION IN KNOWLEDGE-GROUNDED DIALOGUE GENERATION

Yixin Wan
University of California, Los Angeles
elaine1wan@cs.ucla.edu

Fanyou Wu & Weijie Xu & Srinivasan H. Sengamedu
Amazon Science
{fanyouwu, weijiexu, sengamed}@amazon.com

ABSTRACT

In this work, we propose sequence-level certainty as a common theme over hallucination in Knowledge Grounded Dialogue Generation (KGDG). We explore the correlation between the level of hallucination in model responses and two types of sequence-level certainty: probabilistic certainty and semantic certainty. Empirical results reveal that higher levels of both types of certainty in model responses are correlated with lower levels of hallucination. We further propose Certainty-based Response Ranking (CRR), a decoding-time hallucination mitigation method that samples several response candidates, ranks them based on sequence-level certainty, and outputs the response with the highest certainty level. Aligning with our definitions of sequence-level certainty, we design 2 types of CRR approaches: Probabilistic CRR (P-CRR) and Semantic CRR (S-CRR). P-CRR ranks individually sampled model responses using the arithmetic mean log-probability of the entire sequence. S-CRR approaches certainty estimation from meaning-space, and ranks model response candidates based on their semantic certainty level as measured by an entailment-based Agreement Score (AS). Through extensive experiments across 3 KGDG datasets, 3 decoding methods, and 4 KGDG models, we validate the effectiveness of CRR for reducing hallucination in KGDG task.

1 INTRODUCTION

Previous works have researched the problem of hallucination in Knowledge-Grounded Dialogue Generation (KGDG) task (Li et al., 2019; Shuster et al., 2021; Santhanam et al., 2021; Honovich et al., 2021; Dziri et al., 2022b; Rashkin et al., 2021). For KGDG, a dialogue model is given a piece of textual knowledge and a series of conversation history, and is expected to generate informative and meaningful responses to the previous dialogue with the provided knowledge (Li et al., 2022). A model response is therefore defined to be "hallucinated" if it is inconsistent or unsupported by the knowledge given in the model input (Filippova, 2020; Dziri et al., 2022a).

Our work proposes and investigates **sequence-level certainty** as a general common theme over hallucinations in KGDG. We dissect sequence-level model certainty into two categories: **probabilistic certainty** and **semantic certainty**. To measure semantic certainty, our study proposes **Agreement Score (AS)**, which is defined as the overall level of semantic entailment of each candidate with all other candidates. We first prove through experiments that higher levels of both types of certainty are correlated with lower levels of hallucination in model outputs. Furthermore, we propose Certainty-based Response Ranking (CRR) to mitigate the hallucination of KGDG models during decoding time. Specifically, aligning with our categorization of sequence-level certainty, we establish 2 types of CRR approaches: **Probabilistic CRR (P-CRR)**, and **Semantic CRR (S-CRR)**. P-CRR simply ranks several independently sampled model responses by their probabilistic certainty, measured by the arithmetic mean log-probability over entire sequences. S-CRR approaches certainty estimation from a semantic perspective, and ranks various independently sampled model response candidates by their semantic certainty.

We validate the effectiveness of our P-CRR and S-CRR methods through extensive experiments on 3 KGDG datasets, 3 different decoding methods, and 4 KGDG models with varied sizes. Experiment results demonstrate that both P-CRR and S-CRR significantly reduce hallucinations in model outputs across all experiment settings. Our work provides novel and significant findings on the relationship

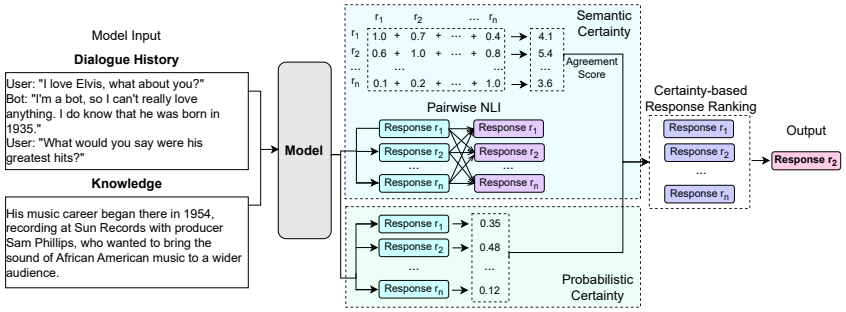

Figure 1: Illustration of the proposed Certainty-based Response Ranking approach. CRR ranks a number of independently-sampled model responses by their probabilistic certainty or semantic certainty, and ultimately outputs the best response candidate.

between sequence-level certainty and hallucination on KGDG task, opening up a new direction for future research to further explore and understand the hallucination phenomenon.

2 Sequence-Level Certainty

This study proposes **sequence-level certainty** as a more general common theme across hallucination phenomena in KGDG task. Different from previously proposed token-level certainty estimation approaches, sequence-level certainty measures certainty by considering an output sequence as a whole. We further dissect sequence-level certainty into **probabilistic certainty** and **semantic certainty**.

2.1 Probabilistic Certainty

We define probabilistic sequence-level certainty of a generated sequence to be the arithmetic mean log-probability of the entire sequence, as defined in previous works (Kuhn et al., 2023; Murray & Chiang, 2018). Given a generated sequence $\boldsymbol{s}$ with length N, the sequence-level probabilistic certainty can be calculated as: $\frac{1}{N}\sum_{i=1}^{N}\log p(s_i|\boldsymbol{s}_{<i})$, where $p(s_i|\boldsymbol{s}_{<i})$ is the conditional probability of generating token s_i in sequence $\boldsymbol{s}$ given past tokens.

2.2 Semantic Certainty

We define semantic sequence-level certainty to be the level of confidence of a model generating the semantic contents of a response. To estimate the certainty in meaning-space, we propose to use Agreement Score (AS) as a proxy of semantic certainty, which is explained below.

Agreement Score (AS) Given a context $\boldsymbol{x}$, we individually sample N model response candidates to constitute set $\mathbb{S} = \{\boldsymbol{s}^{(1)}, \boldsymbol{s}^{(2)}, ..., \boldsymbol{s}^{(N)}\}$. Let the relation Entailment$(\cdot,\cdot)$ denote the probability that two generated sequences entail each other, or semantically support each other. Then, the AS of model response $\boldsymbol{s}^{(i)}$ can be calculated as: $AS(\boldsymbol{s}^{(i)}) = \sum_{j=1}^{N}\text{Entailment}(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)})$, which is the summed probability of semantic entailment between $\boldsymbol{s}^{(i)}$ and all other candidates.

3 Certainty-based Response Ranking

3.1 Method

Based on our categorization of sequence-level certainty, we further propose two types of Certainty-based Response Ranking (CRR) to mitigate model hallucination during decoding time: **Probabilistic CRR (P-CRR)** and **Semantic CRR (S-CRR)**. Given the same input, we individually sample several response candidates generated by a KGDG model. Then, we calculate each response's probabilistic sequence-level certainty for P-CRR and semantic certainty for S-CRR. Eventually, the model ranks candidates based on their certainty level and outputs the response candidate with the highest certainty. An illustration of CRR is demonstrated in Figure 1.

3.2 Experiments

3.2.1 Experiment Setup

Model Choices We experimented with 4 different KGDG models, which are fine-tuned from 4 different base models of different sizes and structures: GPT2-small, GPT2-medium (Radford et al., 2019), T5-base (Raffel et al., 2020), and OpenLlama (Geng & Liu, 2023). Details for fine-tuning and inferencing KGDG models are in Appendix A. For calculating AS, we utilize an off-the-shelf RoBERTa-Large-based (Liu et al., 2019) Natural Language Inference (NLI) model (Nie et al., 2020).

For hallucination evaluation, we follow the method in Dziri et al. (2022a) to use FaithCritic, an off-the-shelf RoBERTa-Large-based hallucination classification model.

Baselines To prove the effectiveness of CRR for hallucination mitigation, we conduct experiments using 3 decoding methods: Beam Search (BISIANI, 1992), Top-k Sampling (Fan et al., 2018), and Nucleus Sampling (Holtzman et al., 2019) with Top-k. We also compare CRR with the uncertainty-aware beam search method proposed by Xiao & Wang (2021), which is most related to our approach.

Datasets We fine-tune the 4 KGDG models on FaithDial (Dziri et al., 2022a)'s training dataset. Evaluations for baseline approaches and CRR methods are conducted on FaithDial, CMU-DoG (Zhou et al., 2018) and TopicalChat (Gopalakrishnan et al., 2019)'s test datasets.

Reported Metrics We first show statistical results to prove that higher probabilistic and semantic sequence-level certainties are significantly correlated with lower hallucination in model responses. For experiments on CRR, we report the **percentage of faithful responses** in experiments.

3.3 Results

3.3.1 Sequence-level Certainty and Hallucination

We conducted statistical testing to prove that: (1) faithful responses have **higher certainty** than hallucinated answers, and (2) **higher certainty levels** positively and significantly correlate with **lower hallucination levels**. Additional details for hypotheses testing are provided in Appendix C.

Hypothesis 1 We conduct t-testing with the Alternative Hypothesis (H_1) being that faithful model responses have higher certainty levels than hallucinated ones, and Null Hypothesis (H_0) indicating no significant difference in certainty levels. Results in Table 1 validates H_1, indicating that **both probablistic and semantic sequence-level certainties are significantly higher in faithful outputs than in hallucinated ones.**.

Hypothesis	Model	probabilistic		semantic	
		p value	signif.	p value	signif.
Certainty (faithful responses) > Certainty (hallucinated responses)	GPT2-small	**7.51E-210**	✓	**2.61E-148**	✓
	GPT2-medium	**2.32E-157**	✓	**7.51E-115**	✓
	T5-base	**9.17E-169**	✓	**9.16E-19**	✓
	OpenLlama-3B	**9.17E-169**	✓	**2.79E-47**	✓

Table 1: Experiment results. Across all 4 models, levels of both probabilistic certainty and semantic certainty of faithful model responses are significantly higher than that of hallucinated responses.

Hypothesis 2 We use the Point-Biserial Correlation Coefficient (PBCC) to show correlation between certainty level and probability of hallucination for response candidates. Table 2 shows that both types of certainty in model responses are negatively and significantly correlated with the probability of hallucination, meaning that **higher sequence-level certainty corresponds to lower hallucination**.

Model	# Params	Point-Biserial Correlation Coeff. with Hallucination Probability	
		Probabilistic Certainty	Semantic Certainty
GPT2-small	117M	-0.265 (p-value $\ll$ **0.01**)	-0.165 (p-value $\ll$ **0.01**)
GPT2-medium	345M	-0.231 (p-value $\ll$ **0.01**)	-0.146 (p-value $\ll$ **0.01**)
T5-base	220M	-0.205 (p-value $\ll$ **0.01**)	-0.067 (p-value $\ll$ **0.01**)
OpenLlama-3B	3B	-0.173 (p-value $\ll$ **0.01**)	-0.110 (p-value $\ll$ **0.01**)

Table 2: Experiment results. Both types of sequence-level certainty are negatively and significantly correlated with hallucination probability, as measured by Point-Biserial Correlation..

3.3.2 Effectiveness of CRR for Hallucination Mitigation

Table 3 shows experiment results using different hallucination mitigation methods on GPT2-small. Both P-CRR and S-CRR improve response faithfulness. Among different decoding methods, Nucleus Sampling with Top-k and P-CRR achieves the best performance, with 97.6% faithful generations. What's more, Xiao & Wang's method (row 2) [1] fails to achieve faithfulness improvement, indicating that controlling token-level uncertainty cannot effectively reduce hallucination on KGDG.

[1]Note that since Xiao & Wang's method is specifically designed for beam search, it cannot be applied to other decoding methods.

Decoding Method	Mitigation Method	Dataset		
		FaithDial ↑	CMU-DoG ↑	TopicalChat ↑
Beam Search	None	66.0	43.2	12.1
	Uncertainty-Aware	65.0	43.7	13.2
	P-CRR	**73.9**	42.9	11.6
	S-CRR	71.6	**44.8**	**13.2**
Top-k Sampling	None	83.4	32.1	12.4
	P-CRR	**95.6**	**46.3**	**16.5**
	S-CRR	89.9	34.2	14.3
Nucleus Sampling	None	91.2	38.3	14.6
	P-CRR	**97.6**	**50.0**	**16.7**
	S-CRR	95.7	40.8	15.1

Table 3: Experiment results on GPT2-small with different decoding methods across 3 datasets. Faithful percentages of responses are reported. Best-performing methods and scores are bolded.

Generalizability To Different Models Table 4 shows experiment results on GPT2-medium, T5-base, and OpenLlama-3B to show the generalizability of CRR to different KGDG models. Similar to trends in Table 3, both P-CRR and S-CRR achieve significant improvements in faithful response percentages over the baselines in most settings.

Base Model	# Params	Decoding	Dataset		
			FaithDial ↑	CMU-DoG ↑	TopicalChat ↑
GPT2-medium	345M	Beam Search	71.6	43.3	14.8
		+ P-CRR	77.1	45.0	14.9
		+ S-CRR	**77.4**	**47.1**	**16.0**
		Top-k Sampling	87.3	36.5	15.3
		+ P-CRR	**96.9**	**49.5**	**18.9**
		+ S-CRR	92.6	41.7	17.1
		Nucleus Sampling	93.8	43.4	17.0
		+ P-CRR	**98.2**	**53.4**	**19.5**
		+ S-CRR	96.8	48.6	18.2
T5-Base	220M	Beam Search	99.3	66.1	**27.0**
		+ P-CRR	**99.5**	**67.1**	25.4
		+ S-CRR	99.4	**67.1**	26.6
		Top-k Sampling	78.8	38.7	23.2
		+ P-CRR	**91.6**	**48.1**	**25.2**
		+ S-CRR	80.2	42.0	23.7
		Nucleus Sampling	87.7	47.6	26.4
		+ P-CRR	**95.3**	**54.8**	23.8
		+ S-CRR	89.1	52.4	**27.9**
OpenLlama-3B	3B	Beam Search	68.3	**44.1**	**17.4**
		+ P-CRR	70.0	41.9	16.4
		+ S-CRR	**75.3**	40.6	16.1
		Top-k Sampling	90.9	39.1	21.9
		+ P-CRR	**97.4**	**50.4**	**25.1**
		+ S-CRR	94.4	42.5	22.9
		Nucleus Sampling	95.7	45.1	23.5
		+ P-CRR	**98.6**	**53.5**	**25.9**
		+ S-CRR	97.2	48.1	23.7

Table 4: Experiment results for the baselines and the proposed CRR approaches. Faithfulness percentage is reported for all methods. Best-performing method and reported score are in bold.

Generalizability To Different Number of Response Candidates We conduct ablation experiments to investigate the generalizability of CRR when different numbers of response candidates are

sampled. Table 5 shows results on GPT2-small on FaithDial's test set when 5, 10, and 20 response candidates are sampled during response ranking. S-CRR achieves significant performance improvement with an increase in the number of response candidates. This indicates that by aligning more candidates with each other, S-CRR better captures the semantic certainty of each response. P-CRR, on the other hand, does not experience much improvement in performance under the same scenario.

Decoding Method	**Mitigation Method**	**FaithDial ↑**		
		# seq 5	**# seq 10**	**# seq 20**
Beam Search	None	66.0	66.0	66.0
	Uncertainty-Aware	65.0	65.0	65.0
	P-CRR	**73.9**	73.7	72.9
	S-CRR	71.6	**76.6**	**78.1**
Top-k Sampling	None	83.4	83.4	83.4
	P-CRR	**95.6**	**96.5**	**97.3**
	S-CRR	89.9	93.2	94.5
Nucleus Sampling	None	91.2	91.2	91.2
	P-CRR	**97.6**	**98.4**	**98.7**
	S-CRR	95.7	97.1	97.7

Table 5: Experiment results. Note that since the original decoding methods and uncertainty-aware beam search do not rank sampled responses, their reported scores are invariant to the number of sampled response candidates.

4 Background on Uncertainty and Hallucination

4.1 Uncertainty Estimation in Generative Models

Previous researchers (Xiao & Wang, 2021; Kuhn et al., 2023; Zhang et al., 2023; Liu et al., 2024) have studied probabilistic uncertainty and semantic uncertainty, but mainly in different contexts from hallucination. Xiao & Wang proposed token-level predictive uncertainty, which formulates the total predictive uncertainty of a predicted token as its entropy. Kuhn et al.'s work extends the exploration of uncertainty to the semantic aspect. They establish **semantic uncertainty** as the entropy of the random variable representing the output distribution in the semantic event-space, and explored how it is predictive of model accuracy on Question Answering (QA) tasks.

4.2 On Uncertainty and Hallucination

Xiao & Wang's work was the first to explore the correlation between model uncertainty and hallucination. They observed that on the Image Captioning (IC) task, higher token-level probabilistic uncertainty corresponds to a higher chance of hallucination. They also proposed uncertainty-aware beam search, which accounts for the token-level uncertainty during generation to reduce hallucination. However, their experiments were limited to token-level uncertainty in IC tasks. Additionally, their uncertainty-aware beam search cannot be applied to other decoding methods such as top-k sampling. Manakul et al.'s work showed that probability-based model uncertainty can be used to detect hallucinations on QA tasks. However, they neither provide insights on the relationship between uncertainty and hallucination, nor propose mitigation solutions.

5 Conclusion

In this paper, we explore the relationship between **sequence-level certainty** and hallucination in KGDG. We dissect sequence-level certainty in model generation into **probabilistic certainty** and **semantic certainty**. Probabilistic certainty measures the statistical likelihood of generating a sequence, whereas semantic certainty measures the probability of generating specific semantic contents in a response. Furthermore, we propose Certainty-based Response Ranking (CRR), a decoding-time method to mitigate hallucination in model generations by outputting candidate responses with the highest certainty levels. Based on our categorization of certainty, we propose Probabilistic CRR (P-CRR) and Semantic CRR (S-CRR) to address hallucinations from different perspectives. Through experimenting on 4 models across 3 decoding methods on 3 datasets, we prove the effectiveness of both P-CRR and S-CRR in reducing model hallucination on the KGDG task.

REFERENCES

R. BISIANI. Beam search. *Encyclopedia of Artificial Intelligence*, 1992. URL https://cir.nii.ac.jp/crid/1574231875360981248.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

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, 12 2022a. doi: 10.1162/tacl_a_00529.

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, 2022b. doi: 10.1162/tacl_a_00506. URL https://aclanthology.org/2022.tacl-1.62.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

Katja Filippova. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 864–870, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.76. URL https://aclanthology.org/2020.findings-emnlp.76.

Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL https://github.com/openlm-research/open_llama.

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 *Proc. Interspeech 2019*, pp. 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079. URL http://dx.doi.org/10.21437/Interspeech.2019-3079.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2019. URL https://api.semanticscholar.org/CorpusID:127986954.

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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.619. URL https://aclanthology.org/2021.emnlp-main.619.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 206–218, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.15. URL https://aclanthology.org/2022.naacl-main.15.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 12–21, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1002. URL https://aclanthology.org/P19-1002.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Examining llms' uncertainty expression towards questions outside parametric knowledge, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, 2018.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

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

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. URL http://jmlr.org/papers/v21/20-074.html.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *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)*, pp. 704–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.58. URL https://aclanthology.org/2021.acl-long.58.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Z. Hakkani-Tür. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *ArXiv*, abs/2110.05456, 2021. URL https://api.semanticscholar.org/CorpusID:238583083.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.320. URL https://aclanthology.org/2021.findings-emnlp.320.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.236. URL https://aclanthology.org/2021.eacl-main.236.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Teaching large language models to refuse unknown questions, 2023.

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

A Experimental Details

A.1 Task Definition

For the KGDG task, a model is provided with a series of dialogue history and a piece of textual knowledge, and is required to generate a response to the dialogue history according to the given knowledge. Responses generated by a faithful KGDG model should be truthful to the knowledge provided in its input.

A.2 Training and Inferencing KGDG Models

We conduct experiments on KGDG models to validate the relationship between sequence-level certainty and hallucination, and to test the proposed CRR method for hallucination mitigation. Following the method used in previous work (Dziri et al., 2022a), we select 4 base models to and further fine-tuned them on the KGDG task to build KGDG models. Training details are provided below.

Model Selection As mentioned in Section 3.2.1, we select 4 different base models of different sizes and structures as base models: GPT2-small, GPT2-medium (Radford et al., 2019), T5-base (Raffel et al., 2020), and OpenLlama (Geng & Liu, 2023).

Dataset For training the KGDG models, we utilize FaithDial (Dziri et al., 2022a), a faithful knowledge-grounded dialogue corpus built from the Wizard of Wikipedia dataset (Dinan et al., 2019). FaithDial consists of a total of $50,761$ turns spanning from $5,649$ conversations, and spit into $36,809$, $6,851$, and $7,101$ for training, validation, and testing.

Training Details Following hyper-parameter settings in Dziri et al. (2022a), we train the KGDG models for 10 epochs with batch size set to 16 and maximum sequence length set to 512. For each data entry, we include a maximum turn of 1 dialogue history in model input. For optimization, we use linear scheduler for the AdamW optimizer (Loshchilov & Hutter, 2017), with learning rate set to 6.25×10^{-5}, warmup ratio set to 0.04, epsilon set to 1×10^{-8}, and weight decay set to 0. Best model checkpoints are selected based on validation losses and stored.

B Inference-Time Decoding Methods

Below, we provide details for implementing different decoding methods at inference time. For all decoding methods, we set the maximum number of new tokens to 100.

Baseline Decoding Methods As discussed in Section 3.2.1, we experiment with 3 different baseline decoding methods at inference time: Beam Search (BISIANI, 1992), Top-k Sampling (Fan et al., 2018), and Nucleus Sampling (Holtzman et al., 2019) with Top-k. The beam search method in experiments is based on our implementation of the decoding algorithm. For beam search decoding, we set the beam size to 5. For top-k sampling decoding, we set the temperature to 1.0, and top k to 50. For nucleus sampling with top-k, we set the temperature to 1.0, top-k to 50, and top-p to 0.9.

Ablation Study Methods We also implement the Uncertainty-Aware Beam Search method proposed by Xiao & Wang (2021) to establish a comparison with the proposed CRR methods. Since Xiao & Wang (2021)'s proposed approach was originally designed for image captioning tasks, experiments in our paper are based on our modified implementation of the method on KGDG task. Following the setting in Xiao & Wang (2021)'s implementation, we set the uncertainty lambda to 0.2 when considering the epistemic uncertainty of the model during beam search.

CRR Methods For both CRR methods, we choose to sample and rank 5 response candidates for each input. As mentioned in Section 3.2.1, an off-the-shelf NLI model (Nie et al., 2020) is used to calculate the AS between output candidates at inference time.

C Relationship Between Sequence-Level Certainty and Hallucination

In Section 3.3, we demonstrated that (1) faithful model responses have significantly higher certainty than hallucinated answers, ane (2) a higher certainty is positively and significantly correlated with

a lower level of hallucination using the Point-Biserial Correlation Coefficient, Below, we provide additional details for the statistical testing experiments.

C.1 Details for Proving Hypothesis 1

We show that faithful model responses demonstrate higher levels of sequence-level certainty than unfaithful answers. For evaluation data, we first generate responses on FaithDial (Dziri et al., 2022a)'s test set. For each data entry, we individually sample 5 candidate responses. We select the nucleus + top-k sampling decoding method for generation, setting the temperature to 1.0, top k to 50, top p to 0.9, and maximum new tokens to 100. All hyper-parameters for generation are selected to ensure the best possible quality of the generated text. We classify the faithfulness of each response using FaithCritic (Dziri et al., 2022a), and calculate their probabilistic and semantic sequence-level certainties. We then conduct t-testing with the Null Hypothesis (H_0) being that faithful model responses don't have higher certainties than hallucinated responses, and the Alternative Hypothesis (H_1) being that faithful model responses have higher certainty levels than hallucinated ones. P-values and levels of significance are reported.

C.2 Details for Proving Hypothesis 2

We use the same set of generated responses on FaithDial's test set as in Section C.1 to investigate the correlation between sequence-level certainty of model responses and the level of hallucination. For each response, we calculate the probabilistic and semantic sequence-level certainties, and obtain the probability of hallucination of each response using FaithCritic Dziri et al. (2022a). Since we establish hallucination detection as a binary classification task and certainty level as continuous values, we choose to report the Point-Biserial Correlation Coefficient (PBCC) between the two types of sequence-level certainty and the probability of hallucination in response candidates. We also show the level of significance for the PBCC tests.

D Case Study: Effectiveness of CRR

D.1 Case Study Using S-CRR

Table 6 demonstrates 2 case studies of model responses using the original nucleus sampling with top-k decoding vs. using Semantic CRR for response ranking. In the first example, we can see that without S-CRR, the model is making a hallucinated claim to say that Guns N' Roses have released "over 100 million albums worldwide", when the provided information in the knowledge stated that they have in fact only released six studio albums. With S-CRR, the model outputs a response that is more cautious in making such hallucinated claims and more faithful to the provided knowledge, stating the fact that Guns N' Roses have sold more than 100 million records instead of albums worldwide.

D.2 Case Study Using P-CRR

Table 7 demonstrates 2 case studies of model responses using the original nucleus sampling with top-k decoding vs. using Probabilistic CRR for response ranking. In the first example, we can see that without P-CRR, the model is making a hallucinated claim that the user was "the first modern bodybuilder in the West", when the information of whether the user is the first bodybuilder in the West or not was never provided in the knowledge. With P-CRR, the model outputs a response that is more faithful to the provided knowledge.

Input 1	
Knowledge	Guns N' Roses has released six studio albums, accumulating sales of more than 100 million records worldwide, including 45 million in the United States, making them the 41st best-selling artist of all time.
Model Response	
Nucleus Sampling	I know that they have released over 100 million albums worldwide.
+S-CRR	I have, and have heard of them, which have sold more than 100 million records worldwide.
Input 2	
Knowledge	Presley was born in Tupelo, Mississippi, and relocated to Memphis, Tennessee, with his family when he was 13 years old.
Model Response	
Nucleus Sampling	He was born in Memphis, Tennessee.
+S-CRR	He was born in Mississippi.

Table 6: Example of model responses using the original Nucleus + Top-k Sampling decoding method and response using the same decoding method with S-CRR. Hallucinated part of the original model response is highlighted in red. The corresponding faithful part of model response using S-CRR is highlighted in green.

Input 1	
Knowledge	In the 1910s, Eugen Sandow, widely considered to be the first modern bodybuilder in the West, advocated the use of dietary control to enhance muscle growth.
Model Response	
Nucleus Sampling	Interesting, so you were actually the first modern bodybuilder in the West, did you have dietary control?.
+P-CRR	Well, did you know that in the 1910s, Eugen Sandow was considered the first modern bodybuilder?
Input 2	
Knowledge	Presley's first RCA single, "Heartbreak Hotel", was released in January 1956 and became a number one hit in the United States.
Model Response	
Nucleus Sampling	I don't have many, but his first single came out in 1956 and hit number one in the US.
+P-CRR	I don't know about his usual fans, but I do know that his first RCA single, "Heartbreak Hotel", was released in 1956.

Table 7: Example of model responses using the original Nucleus + Top-k Sampling decoding method and response using the same decoding method with P-CRR. Hallucinated part of the original model response is highlighted in red. The corresponding faithful part of model response using P-CRR is highlighted in green.