A Factuality and Diversity Reconciled Decoding Method for Knowledge-Grounded Dialogue Generation

Anonymous ACL submission

Abstract

Grounding external knowledge can improve the factuality of responses in dialogue generation, but over emphasis on that could result in the lack of engaging and diverse phrases. Through introduction of randomness in sampling, the current methods can enhance the diversity. However, the sampling could jeopardize the factuality in dialogue generation. In this work, to find a solution to advance creativity except for relying on problematic randomness, and to subtly reconcile the factuality and diversity within the source-grounded paradigm, a novel DoGe method is proposed. DoGe can dynamically switch between the utilization of internal parameter knowledge and external source knowledge according to model's factual confidence. Extensive experiments on three widely-used datasets demonstrate that DoGe can both enhance response diversity and maintain factuality, and it significantly outperforms other various decoding strategy baselines. [1]

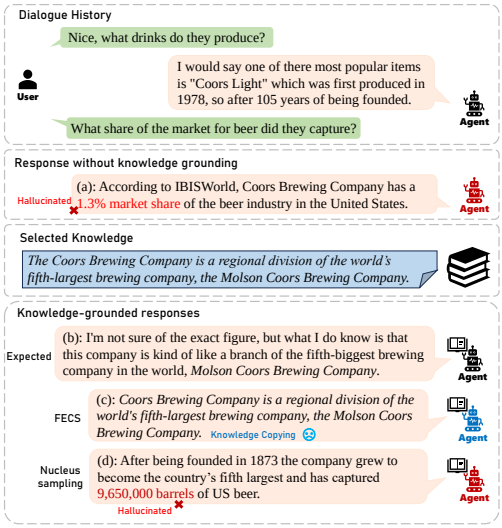

Figure 1: An example illustrates the potential issues that may arise from increasing faithfulness to external knowledge and the drawbacks of compensating for diversity and creativity through sampling.

1 Introduction

Equipped with extensive parameters, large language models (LLMs) demonstrate robust generalization abilities in various NLP downstream tasks (Touvron et al., 2023a; BigScience, 2023; Bai et al., 2023; Yang et al., 2023a; Touvron et al., 2023b). Remarkably, their capability in producing fluent and coherent content is comparable to that of humans in the domain of dialogue (OpenAI, 2023a,b). One conspicuous vulnerabilities of LLMs is the generation of non-factual responses. (Ji et al., 2023). The factuality is a critical property in conversations, which refers to the faithfulness of the generated content to the established world knowledge (Zhang et al., 2023c). A plausible solution to improve factuality is to provide the model with carefully selected external knowledge during response generation. Such approach is in the same line with the Knowledge-Grounded Dialogue Generation (KGDG) task, for which several datasets have been introduced (Dinan et al., 2018; Moghe et al., 2018; Dziri et al., 2022a). Furthermore, some studies have designed schemes to ensure the generated contents are more faithful to the provided knowledge (Deng et al., 2023; Sun et al., 2022; Chen et al., 2023).

However, enhancing factuality through the faithfulness augmented methods results in the following drawbacks: (1) excessive focus on the singular golden knowledge can lead to the reduction of content diversity; (2) the overemphasis on faithfulness may diminish the engaging phrasing of responses, manifesting as the model takes the shortcut of inserting portions of source knowledge into responses, such as the response (b) in Figure 1 (Yang et al., 2023b; Daheim et al., 2023). Remedial measures could be adjusting the decoding strategy, as conventional deterministic decoding methods

[1] The code will be released at GitHub upon publication.

are prone to being trapped in such text degeneration issues (Welleck et al., 2019). Current methods aiming at boosting diversity typically rely on the introduction of randomness in sampling (Holtzman et al., 2020). Unfortunately, these stochastic methods are prone to produce non-factual responses due to the selection of low-probability tokens (Lee et al., 2023). In fact, there is a contradiction between factuality and diversity, and it is hard to meet the two goals at the same time. We empirically analyze the phenomenon and strive to reconcile the two more subtly, and find ways to promote diversity with minimal factuality compromise.

In this paper, we propose a novel **D**ynamic s**o**urce-**G**rounded d**e**coding (DoGe) method. The fundamental idea is to harmonize the utilization of internal parameter knowledge and external source knowledge during generation, to enhance response diversity while preserving the factuality. Specifically, DoGe dynamically switches between masking and exposing external knowledge based on the model's factual confidence level. In the proposed model, the local and global uncertainties of generation are integrated through function conversion to measure the model's factual confidence, which serves as a proxy for the factuality of the generated content. Considering that the LLM has been exposed to vast world knowledge during pre-training, it is naturally capable of generating diverse and creative responses. If the LLM exhibits high confidence in its predictions without resorting to external knowledge, we deem its output factually accurate. Meanwhile, the masking is maintained and generation solely depends on parameter knowledge. Conversely, if the LLM exhibits low confidence, the model resorts to the external knowledge within the input and recalculate the probability distribution for the prediction. Additionally, we design a scorer to re-rank candidate tokens by considering knowledge-attentive rewards during generation. The approach preserves tokens that are faithful enough to the knowledge source, therefore ensuring the content's faithfulness without fabrication.

Our contributions are summarized as follows:

- We probe the trade-off between factuality and diversity in dialogue generation. Simply enhancing faithfulness leads to damage in diversity, while compensating diversity through stochastic decoding leads to damage in factuality.
- We propose an innovative **D**ynamic s**o**urce-**G**rounded d**e**coding (DoGe) method, which effectively reconciles the diversity and factuality in Knowledge-Grounded Dialogue Generation.
- We perform comprehensive experiments, and the experimental results demonstrate the superiority of our method in both automated and human evaluation metrics.

2 Related Work

Knowledge-grounded dialogue generation Knowledge-grounded dialogue generation aims to alleviate dull and unfaithful responses by injecting external knowledge into input of dialogue models and it consists of two subtasks: knowledge selection and response generation. The hot spot of early research is mainly concentrated on how to improve the performance of knowledge selection (Sun et al., 2023; Xu et al., 2022; Zhan et al., 2021; Yang et al., 2022; Meng et al., 2021; Kim et al., 2020). With the remarkable leap in the capabilities of generative models, the research focus gradually shifts to the response generation subtask (Zhao et al., 2020a; Liu et al., 2021; Zhao et al., 2020b; Zheng et al., 2021). Ideally, a brilliant robot should generate informative and truthful responses, while maintaining the naturalistic phrasing and excellent interactivity (Dziri et al., 2022a).

Hallucinations in Text Generation Hallucination refers to the problem that the generated contents diverge from the user instruction, contradicts previously generated context, or established world (Zhang et al., 2023c). It poses serious risks to the practical applications in NLP. Existing works have conducted comprehensive investigations into hallucination from multiple perspectives, including its origin (McKenna et al., 2023; Dziri et al., 2022b), detection (Zhang et al., 2023a; Manakul et al., 2023; Fadeeva et al., 2023), and mitigation (Choi et al., 2023; Chuang et al., 2023; Li et al., 2023b). Many approaches and benchmarks about hallucination were proposed, such as question answering (Gao et al., 2023; Lin et al., 2022), text summarization (Cao et al., 2020; Zhong et al., 2021) and dialogue generation (Li et al., 2023a; Wan et al., 2023; Chen et al., 2023; Sun et al., 2022). One effective solution to mitigate hallucination in QA is resorting to external knowledge as supplementary evidence, which provides more precise answers to user queries. Considering the nature of open-domain conversations, a chit-chat

agent should seamlessly integrate external knowledge into its responses, thus achieving a balance between factuality and content diversity.

3 Preliminaries

3.1 Task Formulation

Given an task-specific instruction $\mathcal{I}$ of knowledge-grounded dialogue generation, a dialogue history h across multiple dialogue turns, a user utterance u, and corresponding knowledge $k = (k_1, \ldots, k_m)$ for the current turn, where m represents the number of tokens in k, the objective of the KGDG task is to generate a response $y = (y_1, \ldots, y_n)$, with n being the number of tokens in y. We construct the conversation context x by concatenating the task-specific instruction $\mathcal{I}$, dialogue history h, and the user's last utterance u:

$$x = [\mathcal{I}; h; u]. \tag{1}$$

The probability distribution of the current token is derived by inputting the conversation context, corresponding knowledge, and previously generated tokens into a large language model:

$$p(y_t) = \mathcal{LLM}(x, k, y_{\leq t-1}). \tag{2}$$

Ultimately, the response y is generated through either a deterministic or stochastic decoding strategy, executed step-by-step.

3.2 Faithfulness and Factuality

Existing approaches alleviate the generation of non-factual responses by enhancing their faithfulness to external knowledge, which is overly restrictive (Honovich et al., 2021; Deng et al., 2023). In fact, it is just one means of improving factuality. We formally define faithfulness and factuality and explain their relationship to facilitate the research:

Definition 3.1 Faithfulness($\mathcal{F}$): *Given a response y, and external knowledge $\mathcal{K} = (k_1, \ldots, k_j)$ at turn n, we say that the response y is faithful with respect to the external knowledge ($\mathcal{F}(\mathcal{K}, y)$) if and only if the following condition holds*:

- $\exists\, \Gamma$ *such that* $\Gamma \models y$, *where* Γ *is a non-empty subset of* $\mathcal{K}$ *and* $\models$ *denotes semantic entailment. In other words, there is no interpretation* $\mathcal{I}$ *such that all members of* Γ *are true and* y *is false* (Dziri et al., 2022a).

Definition 3.2 Factuality($\mathcal{T}$): *Given a response y, we say that y is factual ($\mathcal{T}(y)$) if and only if the following condition holds*:

- $\exists\, \Phi$ *such that* $\Phi \models y$, *where* Φ *is a non-empty subset of world knowledge* $\mathcal{K}_w$ *and* $\models$ *denotes semantic entailment*.

Theorem 3.1 $\mathcal{F} \models \mathcal{T}, \mathcal{T} \not\models \mathcal{F}$, where $\models$ denotes entailment.

Theorem 3.1 indicates that responses ensuring faithfulness are necessarily factual, but the converse may not always apply. It also indicates that a method that guarantees faithfulness can serve as a fallback for factuality, which gives us some inspiration for DoGe. The proof of the theorem is shown in the Appendix A.

3.3 Factuality-Diversity Trade-Off

Initially, external knowledge is provided in the input of dialogue model to prevent situations like the response (a) in Figure 1, where fabricated facts may occur, and to generate ideal responses such as response (b). However, through experiments and observations, we found that it often leads to knowledge copying, as demonstrated by the response (c). This phenomenon significantly undermines the interactive experience and the diversity of conversational styles.

Given that such text degeneration issue could be caused by conventional deterministic decoding methods, to circumvent it, we explore the commonly used sampling-based methods to enhance the diversity and creativity of the generated content. While this approach effectively mitigates the phenomenon of knowledge copying (Chen et al., 2023), it reintroduces the risk of hallucination. As illustrated by the response (d) in Figure 1, the response is partially faithful to the source knowledge, while partially containing fabricated information. We endeavor to balance factuality and diversity by controlling the degree of randomness. Specifically, we implement that through adjustments of the temperature parameter τ to control the smoothness of the distribution and the threshold p used in nucleus sampling to filter out low-probability token tails.

We approximate factuality using the **Critic** metric proposed by Dziri et al. (2022a), assess content diversity through **distinct-2** and Precision of the Longest Common Sub-sequence (**P-LCS**) (Yang et al., 2023b), where lower distinct-2 values indicate reduced diversity and higher P-LCS values indicate lower creativity of knowledge utilization. The experimental results, as depicted in Figure 2, reveal a apparent trade-off between the two aspects. The trends show that a pursuit of enhanced diversity

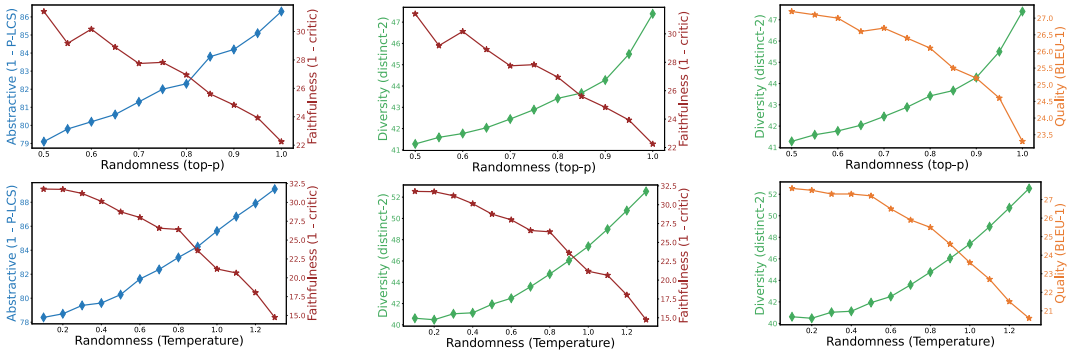

Figure 2: Trade-off between factuality and diversity in different settings in knowledge-grounded dialogue generation.

and creativity inevitably leads to a obvious reduction in factuality, which is inherently caused by the characteristic of stochastic decoding. Besides, it also results in a degradation of generation quality, as reflected by a decrease of BLEU scores (Papineni et al., 2002). Our objective is to explore a way beyond the introduction of randomness to improve diversity with minimal factual compromise.

4 Approach

4.1 DoGe Overview

The entire workflow of DoGe can be outlined as follows: Based on the factuality indicator, DoGe initially computes a factual confidence score for the token generated grounded solely on parameter knowledge. If the score is high, this token is retained. If the score is low, DoGe resorts to the external knowledge within the input to recalculate the probability distribution for the prediction and employ knowledge-attentive decoding to ensure the faithfulness of the generation. The overview of DoGe is illustrated in Figure 3.

4.2 Factual-Based Mask Switching

Considering that incorporating external knowledge can reduce the model's focus on the context and encourage the undesirable practice of knowledge copying (Yang et al., 2023b), we employ a dynamic switching strategy to interrupt the shortcut of copying, thereby fostering creative knowledge integration. This strategy involves dynamically alternating between masking and exposing external knowledge, guided by a factuality indicator. Through this approach, DoGe achieves a balance in the use of parameter knowledge and external knowledge during generation. The attention masks of both high and low factual confidence are illustrated on the left side of Figure 3, respectively.

4.2.1 Source Masking

During the generation process, the source masking is implemented by setting the attention mask of external knowledge in the model's input to 0. Hence, the model does not engage in computations related to this segment when generating the current token, stimulating it to activate and utilize its inherent parameter knowledge.

$$p(y_t) = \mathcal{LLM}(\mathcal{I}, h, u, y_{\leq t-1}), \tag{3}$$

In order to further enhance diversity in content generation using internal knowledge and to encourage the occurrence of rare knowledge, we adopt the penalty term from contrastive search to suppress dull generation (Su et al., 2022).

$$\begin{aligned} p_\theta^k(v) = \text{topK}\Big\{ &(1-\varepsilon) \times p_\theta(v|x, y_{\leq t-1}) \\ &- \varepsilon \times \max_{0 \leq i \leq t-1} \{sim(h_v, h_{y_i})\} \Big\}. \end{aligned} \tag{4}$$

where v denotes the top-k candidate tokens.

4.2.2 Factuality Indicator

Uncertainty can serve as a gauge for determining when to trust LLMs. Several studies investigated the potential of using uncertainty as a means to detect hallucinations (Manakul et al., 2023; Huang et al., 2023; Duan et al., 2023). Given that a sufficiently advanced LLM should assign low probabilities to tokens that are likely to introduce hallucinated content, we developed a Factuality Indicator using a logit-based uncertainty method.

Inspired by the work of Zhang et al. (2023b), we harness both local and global uncertainty information to derive a factual confidence score via functional conversion. The score is then compared with an pre-defined threshold λ to determine whether resorts to external knowledge. If the factual confidence score surpasses this threshold, it suggests

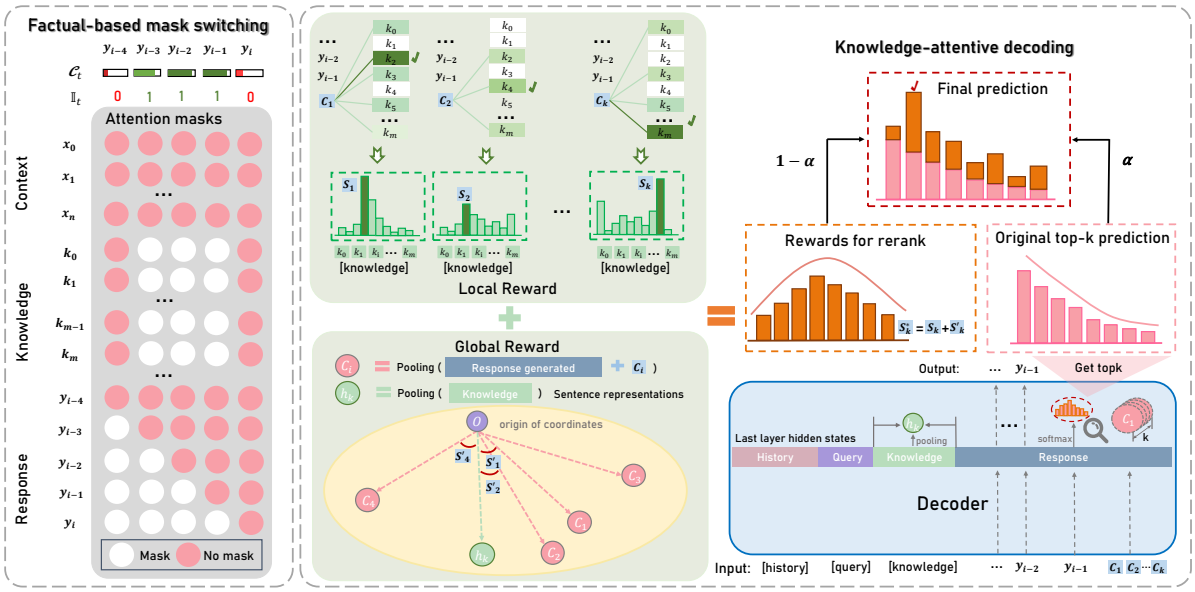

Figure 3: An overview of the DoGe method.

that the parameter knowledge of LLM is adequate to address the user's query. Otherwise, DoGe utilizes knowledge-attentive decoding to improve the faithfulness of generation.

$$\mathbb{I}_t = \begin{cases} 1 & \text{if } \mathcal{C}_t > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

$$\mathcal{C}_t = \frac{1}{2} * (p_{max} + \frac{1}{\mathcal{H}_t + 1}) - \lambda. \tag{6}$$

Specifically, we define local factual confidence as the highest probability among the candidate tokens generated by the model, and global factual confidence as a functional transformation of the entropy of the entire probability distribution.

$$p_{max} = \max_{v \in \mathcal{V}} p(v|x, y_{<t}), \tag{7}$$

$$\mathcal{H}_t = -\sum_{v \in \mathcal{V}} p_t(v) * \log_2(p_t(v)). \tag{8}$$

4.3 Knowledge-Attentive Decoding

When DoGe exhibits a lack of factual confidence, it resorts to the external knowledge. We devised a scoring system to rerank candidate tokens, favoring those that are more faithful to the knowledge.

We ensure faithfulness in two aspects: (1) local reward: the scoring system is inclined to select tokens that pay more attention to the knowledge segment during generation, and (2) global reward: the scoring system prefers tokens that make the entire response more semantically similar to external knowledge. For the global reward, we first append the candidate token to the previously generated content. We then compute the sentence representations of the knowledge and response by mean-pooling their hidden states. The cosine similarity between these sentence representations is considered the global reward. The reranking process of top k candidates is conducted as the following formula:

$$q_\theta^k(v) = \text{topK}\Big\{(1-\beta) \times p_\theta(v|x, k, y_{\le t-1}) + \beta \times \Big[\underbrace{\max_{0 \le j \le |k|-1} \{att(v, k_j)\}}_{\text{local reward}} + \underbrace{sim(h_v, h_k)}_{\text{global reward}}\Big]\Big\}. \tag{9}$$

where $sim(\cdot,\cdot)$ denotes the cosine similarity, and $att(v, k_j)$ denotes the attention weight between v and k_j after max-pooling for all the layers and attention heads.

4.4 Final Prediction

We combine the aforementioned designs to derive the final distribution using the formula below:

$$p^*(v) = \mathbb{I}_t \times p_\theta(v) + (1 - \mathbb{I}_t) \times q_\theta^k(v). \tag{10}$$

Subsequently, we determine the final predicted token with the highest score through a deterministic decoding approach.

$$y_t = \arg\max_v p^*(v). \tag{11}$$

Refer to Appendix C for the pseudocode of the complete inference process.

In-context	Method	Wizard of Wikipedia						FAITHDIAL					
		BLEU-1/2	ROU-L	Dist-1/2/3	Critic↓	B-Score	non-F.↓	BLEU-1/2	ROU-L	Dist-1/2/3	Critic↓	B-Score	non-F.↓
3-shots	Greedy	27.5/16.5	24.2	7.72/40.35/65.06	68.49	55.3	13.0	29.0/17.5	25.9	8.03/38.89/61.70	63.66	58.9	13.5
	Beam	29.3/18.7	25.5	8.35/41.94/65.62	59.55	59.1	**12.5**	31.1/18.8	26.3	8.66/39.88/61.02	**53.73**	62.5	**13.0**
	Nucleus	25.2/14.2	20.7	7.86/**44.28/73.18**	75.17	52.4	18.5	26.6/15.2	23.2	8.25/**43.76/71.28**	70.73	55.1	19.5
	F-Nucleus	26.1/15.1	22.6	7.89/43.42/71.10	73.05	53.4	16.0	27.3/15.9	24.0	8.15/41.81/67.97	69.08	56.4	15.5
	DoLa	27.6/16.6	24.2	7.72/40.34/65.04	68.15	55.4	13.5	29.2/17.5	26.0	8.01/38.95/61.82	64.23	58.9	13.5
	CS	27.0/15.6	23.0	8.31/42.32/67.84	68.20	54.0	14.0	28.5/16.7	24.4	8.63/42.1/64.93	61.96	57.0	15.0
	FECS	28.9/18.2	25.4	8.20/42.50/67.43	**57.89**	59.9	13.5	30.0/18.7	26.4	8.14/39.30/62.30	56.68	**63.8**	16.0
	DoGe	**29.9*/19.0***	**26.0***	**8.37**/42.97/68.21	64.35	**60.7***	**12.5**	**31.2/19.3***	**26.8***	**8.74***/41.36/64.67	59.95	**63.8**	**13.0**
	GPT-3.5	29.3/18.8	27.8	7.57/44.36/72.96	77.00	65.9	8.0	31.4/20.1	29.5	7.95/42.48/68.49	69.98	68.7	9.5

Table 1: Automatic Evaluation results on the Wizard of Wikipedia and FAITHDIAL dataset. The best results excluding GPT-3.5 are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

Method	BLEU-1/2	ROU-L	Dist-1/2/3	Critic↓	B-Score	non-F.↓
Greedy	38.3/29.3	36.7	5.50/28.42/49.76	67.46	55.6	**4.5**
Beam	**44.8**/37.1	41.3	5.87/30.5/50.51	**60.14**	59.5	6.0
Nucleus	32.1/22.2	29.8	5.83/**33.84/62.38**	80.63	52.0	9.5
F-Nucleus	34.0/24.4	32.1	5.76/32.51/59.53	78.51	53.3	9.0
DoLa	38.2/29.0	36.6	5.53/28.61/49.77	68.39	55.4	5.0
CS	34.7/24.8	32.2	5.84/29.8/50.61	66.62	52.5	7.5
FECS	40.8/32.9	40.2	5.47/29.00/50.78	70.36	58.9	8.0
DoGe	**44.8/37.5***	**42.7***	**5.99***/30.53/51.94	68.34	**60.2***	**4.5**

Table 2: 3-shots Automatic Evaluation results on Holl-E dataset. The best results are highlighted with **bold**. "*" denotes that the improvement to the best baseline is statistically significant (t-test with p-value < 0.01).

5 Experiments

5.1 Experimental Setup

Dataset. We conduct experiments on three knowledge-grounded dialogue datasets: Wizard of Wikipedia (WoW) (Dinan et al., 2018), FAITHDIAL (Dziri et al., 2022a), and Holl-E (Moghe et al., 2018). **WoW** is collected based on Wikipedia, with one crowdsourcer acts as a knowledgeable wizard and the other plays the role of an inquisitive apprentice. We evaluate on its test seen set, comprising a total of 8,715 dialogues across 533 topics. FAITHDIAL is a new benchmark for hallucination-free dialogues, which optimizes the responses in the WoW dataset to be more faithful to knowledge. **Holl-E** is constructed by MTurk workers and it focuses on movies as two workers talk with each other about a chosen movie.

Baselines. We compare DoGe with the following decoding strategy baselines: Greedy Decoding (Greedy), Beam Search (Beam), Nucleus Sampling (Nulceus), Factual-Nucleus Sampling (F-Nucleus), DoLa, Contrastive search (CS), and FECS. For Beam Search, we set the beam size to 3, and for Nucleus Sampling, p = 0.9.

F-Nucleus: Lee et al. (2023) modified Nucleus Sampling by adapting the randomness dynamically to improve the factuality of generation. We set $p|\lambda|\omega = 0.9|0.9|0.7$.

DoLa: Chuang et al. (2023) amplify the factual knowledge in LLM by contrasting the logits from different layers to enhance factuality.

CS: Su et al. (2022) penalized previously generated tokens to overcome degeneration and enhance content diversity. We set $k|\alpha = 3|0.6$.

FECS: Chen et al. (2023) extended Contrastive Search by integrating a faithfulness term that encourages factuality. We set $k|\alpha|\beta = 3|0.3|0.3$.

Implementation Details. We conducted experiments on the Llama-2-7b-chat-hf model (Touvron et al., 2023b), utilizing the open-source Hugging Face transformers (Wolf et al., 2020). All experiments are conducted with few-shot prompting (three shots). The three demonstrations are randomly selected from the dataset, and they are exhibited with task instructions in the Appendix B. We skip the step of knowledge selection and directly use manually annotated golden knowledge from the three datasets as input in the experiments. For the hyperparameters in Doge, we set $k|\varepsilon|\beta|\lambda = 3|0.6|0.6|0.9$. All of the experiments are performed on NVIDIA A100-SXM4-40GB GPU.

5.2 Experimental Results

5.2.1 Automatic Evaluation

We selected **BLEU** and ROUGE-L (**ROU-L**) for evaluating generation quality (Papineni et al., 2002; Lin, 2004), Distinct-n (**Dist-n**) for generation diversity (Li et al., 2016), and **Critic** (Dziri et al., 2022a) along with BertScore (**B-Score**) to measure the faithfulness between the knowledge and the generated response (Zhang et al., 2020). Additionally, we randomly sampled 200 instances in the test set and assessed their factuality by instructing GPT-4, denoted as **non-F**. The **Critic** score reflects the proportion of samples unfaithful to external knowledge, while the **non-F** value indicates the

proportion of samples unfaithful to world knowledge among the 200 samples.

Tables 1 and 2 present the results on three knowledge-grounded dialogue datasets. DoGe achieves substantial improvements on both BLEU-1/2 and ROUGE metrics, outperforming all baselines. This enhancement can be attributed to DoGe's effective reconciliation of diversity and factuality in its generated responses, which results in higher consistency with the ground-truth across the metrics evaluating generation quality. In terms of diversity metrics, DoGe performs only behind the sampling-based decoding method slightly, which secures greater diversity but significantly compromises generation quality and faithfulness. Although DoGe's adherence to external knowledge is not as high as that of Beam search and FECS, it demonstrates superior factual accuracy. To evaluate knowledge copying, we adopted the distribution between coverage and density proposed by Grusky et al. (2020), where a high density indicates a tendency to extract snippets from external knowledge for constructing responses. According to the results depicted in Figure 4, compared to the strongest baseline FECS, DoGe's use of knowledge is more creative, maintaining natural phrasing. These findings suggest that DoGe is a decoding strategy that achieves the best overall performance in open-domain dialogue. Furthermore, GPT-3.5's performance on faithfulness and factuality reinforces our belief that factual generation need not be confined solely to external knowledge.

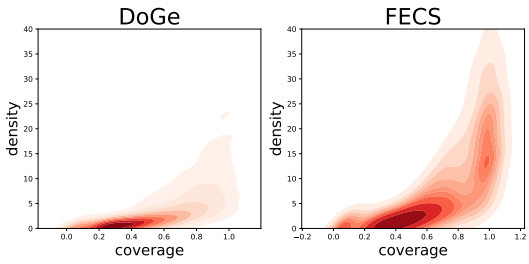

Figure 4: Density and coverage comparison. DoGe generated responses tend to be more abstractive than that of FECS.

5.2.2 LLMs-based Evaluation

We employ G-Eval to assess the Naturalness (**Nat.**) and Coherence (**Coh.**) of responses from five decoding methods, as detailed in previous studies (Liu et al., 2023; Chiang and yi Lee, 2023). Furthermore, we construct explicit instruction to evaluate the task-specific metric Informativeness (**Inf.**) by strictly adhering to the established rating strategy (Fu et al., 2023). The **Inf.** metric determines which response incorporates more interesting knowledge. We randomly selected 200 data samples from the test set and prompted GPT-4 to assign discrete ratings ranging from 1 to 3 points to these generated responses. Table 3 showcases the results of the LLMs-based evaluation, where the responses generated by DoGe surpass all baseline methods in terms of **Nat.** and **Coh.**. This substantial improvement is attributed to DoGe's natural knowledge integration. In contrast, the extractive knowledge utilization exhibited by FECS leads to its underwhelming performance in these metrics. Furthermore, DoGe demonstrates an exceptional ability to balance the use of parameter knowledge with external knowledge, significantly contributing to its superior performance in informativeness.

	WoW			FAITHDIAL		
	Nat.	**Inf.**	**Coh.**	**Nat.**	**Inf.**	**Coh.**
Greedy	2.582	2.261	2.706	2.268	2.070	2.397
Nucleus	2.453	1.913	2.505	2.092	1.936	2.138
F-Nucleus	2.512	2.159	2.635	2.236	1.818	2.414
FECS	1.991	2.378	2.340	1.820	2.147	2.165
DoGe	**2.629**	**2.427**	**2.764**	**2.348**	**2.265**	**2.537**

Table 3: LLMs-based Evaluation results on Wizard of Wikipedia and FAITHDIAL dataset.

5.2.3 Human Evaluation

For human evaluation, we randomly selected 100 samples from the test seen set of the WoW dataset. Given the context and its responses from the baseline models, five well-educated annotators were asked to choose the superior response based on three criteria: (1) Coherence (**Coh.**): which model generates more contextually coherent responses; (2) Engagingness (**Eng.**): which model produces more interesting responses; and (3) Factuality (**Fac.**): which response contains fewer factual errors. To measure the agreement among different annotators, we computed the Fleiss' kappa value (Fleiss, 1971). The results of the human-based evaluation are depicted in Figure 5. We compared our method against Greedy search, Nucleus sampling, Factual-Nucleus Sampling, and FECS as baseline models. The results indicate that our method outperforms the baselines in terms of factuality and coherence, while also ensuring the generated content is both creative and diverse, specifically manifested as its engagingness is on par with that of sampling

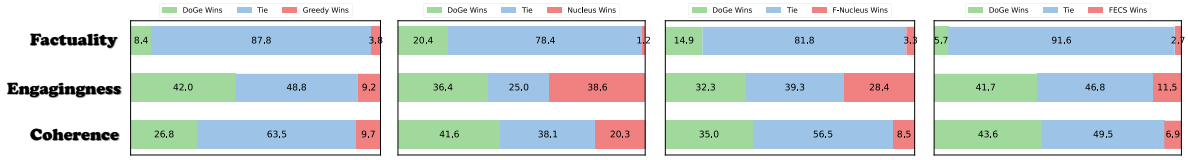

Figure 5: Human evaluation results. The result is statistically significant with p-value < 0.05, and **Kappa** (κ) falls between 0.4 and 0.6, suggesting moderate agreement.

Method	Wizard of Wikipedia						FAITHDIAL					
	BLEU-1/2	ROU-L	Dist-1/2/3	Critic	B-Score	non-F.	BLEU-1/2	ROU-L	Dist-1/2/3	Critic	B-Score	non-F.
DoGe	**29.9**/19.0	**26.0**	**8.37/42.97/68.21**	**64.35**	60.7	**12.5**	**31.2/19.3**	**26.8**	**8.74/41.36/64.67**	59.95	63.8	**13.0**
-KAD(Local)	28.2/16.3	23.9	8.15/41.82/66.97	68.76	57.0	15.0	29.1/17.5	25.7	8.16/39.20/62.39	62.73	59.3	15.5
-KAD(Global)	29.3/18.5	25.6	8.31/42.34/67.96	65.16	58.9	13.0	30.7/18.8	26.3	8.57/40.98/63.87	60.08	62.6	**13.0**
-KAD	27.8/16.1	23.6	8.07/41.56/66.89	70.54	55.8	14.5	28.7/17.2	25.6	8.03/38.88/61.85	64.44	58.4	16.0
-FBM	29.7/**19.1**	25.8	7.94/41.70/67.18	65.52	**67.4**	**12.5**	29.9/19.1	26.5	8.53/40.05/62.91	**58.33**	**68.6**	13.5

Table 4: Ablation study on on the Wizard of Wikipedia and the FAITHDIAL dataset. -FBM denotes removing the factual-based masking compared to DoGe. -KAD denotes removing the knowledge-attentive decoding compared to DoGe. (Local) and (Global) denotes local reward and global reward, respectively.

Method	Greedy	Beam	Nucleus	F-Nucleus
Speed	1.02	1.58	1.04	1.07
Method	DoLa	CS	FECS	DoGe
Speed	1.36	2.81	2.95	3.27

Table 5: The averaged decoding speed (sec) per instance using different decoding methods on the FAITHDIAL dataset.

decoding methods. Moreover, the responses produced by DoGe show strong relevance to the given context, ensuring a smooth and natural conversational flow.

5.3 Ablation Study

To assess the effectiveness of our proposed decoding method, we conducted ablation studies by removing specific modules from DoGe, including factual-based masking (FBM), knowledge-attentive decoding (KAD), local reward, and global reward. The ablation results for the WoW and FAITHDIAL datasets are presented in Table 4. We found that the removal of any module leads to a decline in performance across various aspects. Specifically, the absence of either FBM or KAD results in a decrease in the diversity of the generated responses. Eliminating FBM suppresses the use of parameter knowledge, significantly reducing the creativity of the generated content, while omitting KAD leads to a diminished attention on integrating provided knowledge. External knowledge supplements additional information, and its effective utilization can enhance diversity. Furthermore, the findings indicate that both local and global rewards contribute positively to KAD, with local reward having a slightly greater impact. However, it is the combination of both rewards that achieves a more effective integration of external knowledge.

5.4 Case Study

To intuitively assess the performance of various decoding methods, we compare the responses generated by Greedy Search, Nucleus Sampling, FECS, and DoGe. The examples presented in Table 6 show that both Greedy Search and FECS tend to merely replicate external knowledge, which results in a lack of coherence with the context. Due to its random sampling nature, Nucleus Sampling introduces two incorrect facts: (1) Jordan actually played for only two teams, and (2) the term "Dream Team" specifically refers to the U.S. Olympic men's basketball team. In contrast, the response from DoGe not only exhibits a higher level of interactivity but also enriches the discussion by incorporating information beyond the provided knowledge.

6 Conclusion

In this paper, we discover that enhancing factuality through faithfulness-augmented methods leads to a decrease in content diversity and the creative use of knowledge. Consequently, we introduce a novel DoGe method, which balances the use of parameter knowledge and external knowledge during generation. Extensive experiments conducted on three datasets demonstrate that DoGe effectively reconciles the trade-off between factuality and diversity in knowledge-grounded dialogue generation.

Limitations

DoGe effectively reconciles factuality and diversity trade-off in knowledge-grounded dialogue generation. However, we acknowledge certain limitations in our work:

(1) The inference speed of Doge is slower than the baselines. Table 5 presents the average decoding time (in seconds) per instance for evaluating the decoding latency of our proposed DoGe objective. The results are averaged across all the instances from the FAITHDIAL dataset. Due to the presence of the reranking mechanism, DoGe has a decoding speed similar to FECS and CS, slower than other baselines.

(2) As a decoding method, DoGe's applicability is not universal and is primarily suited for open-domain dialogue scenarios that seek to be engaging and diverse. Its suitability diminishes in question-answering contexts where the accuracy and truthfulness of content are prioritized.

(3) Due to time and resource constraints, we have not conducted experiments on base models of various types and sizes. In the future, we plan to supplement this part of our research.

Ethics Statement

The benchmark datasets we utilized in our experiments—WoW (Dinan et al., 2018), FAITHDIAL (Dziri et al., 2022a), and Holl-E (Moghe et al., 2018)—are all highly respected, open-source datasets collected by crowdsourced workers. They were compiled with strict adherence to user privacy protection protocols, ensuring the exclusion of any personal information. Moreover, our proposed approach is conscientiously designed to uphold ethical standards and promote societal fairness, guaranteeing that no bias is introduced. For the human evaluation component of our study, all participants were volunteers who received comprehensive information about the purpose of the research, ensuring informed consent. Furthermore, all participants were fairly and reasonably compensated for their contributions. We informed the annotators that the data is to be used solely for academic research purposes.

DoGe does not verify the factual accuracy of the provided reference, so there is a risk of potential misuse. Specifically, if the knowledge source contains non-factual information, DoGe is likely to generate misinformation as well. We strongly advise users to carefully verify the knowledge source before using it. The experiments related to OpenAI's ChatGPT were conducted by the author studying at MBZUAI in the UAE. We use AI writing in our work, but only for polishing articles to enhance their readability.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

BigScience. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. 2023. Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 843–851, Singapore. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models.

Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. 2023. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation.

Yifan Deng, Xingsheng Zhang, Heyan Huang, and Yue Hu. 2023. Towards faithful dialogues via focus learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 4554–4566, Toronto, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *CoRR*, abs/1811.01241.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models.

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

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models?

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. Lm-polygraph: Uncertainty estimation for language models.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

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.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Factuality enhanced language models for open-ended text generation.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Shilei Liu, Xiaofeng Zhao, Bochao Li, Feiliang Ren, Longhui Zhang, and Shujuan Yin. 2021. A three-stage learning framework for low-resource knowledge-grounded dialogue generation.

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. *CoRR*, abs/2303.16634.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks.

Chuan Meng, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tengxiao Xi, and Maarten de Rijke. 2021. Initiative-aware self-supervised learning for knowledge-grounded conversations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 522–532, New York, NY, USA. Association for Computing Machinery.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023a. Chatgpt. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation.

Weiwei Sun, Pengjie Ren, and Zhaochun Ren. 2023. Generative knowledge selection for knowledge-grounded dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2077–2088, Dubrovnik, Croatia. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2022. Contrastive learning reduces hallucination in conversations.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Yixin Wan, Fanyou Wu, Weijie Xu, and Srinivasan H. Sengamedu. 2023. Sequence-level certainty reduces hallucination in knowledge-grounded dialogue generation.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lin Xu, Qixian Zhou, Jinlan Fu, Min-Yen Kan, and See-Kiong Ng. 2022. CorefDiffs: Co-referential and differential knowledge flow in document grounded conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 471–484, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan

Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open large-scale language models.

Chenxu Yang, Zheng Lin, Jiangnan Li, Fandong Meng, Weiping Wang, Lanrui Wang, and Jie Zhou. 2022. TAKE: Topic-shift aware knowledge sElection for dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 253–265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chenxu Yang, Zheng Lin, Lanrui Wang, Chong Tian, Liang Pang, Jiangnan Li, Qirong Ho, Yanan Cao, and Weiping Wang. 2023b. Multi-level adaptive contrastive learning for knowledge internalization in dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8002–8015, Singapore. Association for Computational Linguistics.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023a. Enhancing uncertainty-based hallucination detection with stronger focus.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

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. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models.

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models.

Wen Zheng, Natasa Milic-Frayling, and Ke Zhou. 2021. Knowledge-grounded dialogue generation with term-level de-noising. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2972–2983, Online. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A The proof of Theorem 2.1

Theorem. $\mathcal{F} \models \mathcal{T}, \mathcal{T} \not\models \mathcal{F}$.

Proof. $\mathcal{F} \models \mathcal{T}$: For all y that satisfy $\mathcal{F}(K, y)$, there exists Γ such that $\Gamma \models y$ and $\Gamma \subseteq K$. Since $K \subsetneq K_w$ (external knowledge is a proper subset of world knowledge), it follows that $\Gamma \subseteq K_w$. Let $\Phi = \Gamma$, then $\Phi \models y$ and $\Phi \subseteq K$. Hence, $\mathcal{T}(y)$ holds, and the conclusion is proved.

$\mathcal{T} \not\models \mathcal{F}$: We prove it by contradiction. Suppose that $\mathcal{T} \models \mathcal{F}$, then for all y that satisfy $\mathcal{T}(y)$, there exists Φ such that $\Phi \models y$ and $\Phi \subseteq K_w$. Let $\Phi \subseteq K_w/K$. Since $\mathcal{T} \models \mathcal{F}$, then $\Phi \subseteq K$. However, $\Phi \subseteq K_w/K$, it implies that $\Phi = \emptyset$, but $\Phi \neq \emptyset$, leading to a contradiction, thus the conclusion is not valid.

□

B Details of the Prompts

Our instruction template and demonstrations for prompting Large Language Models to generate response in knowledge-grounded dialogue generation are as follows:

You are a chit-chat robot chatting with a user and you will be provided with your dialogue history and a piece of knowledge related to the user's last utterance. Understand this knowledge and use it to generate a concise (no longer than 30 words) but informative (containing some attractive knowledge) reply.
The followings are some demonstrations you can use as reference.
The following is a multi-round dialogue between the user and you.
User's utterance: Red is my favorite color. My house, car, and clothes are all red. Your response: I also like red buddy, I have more red colour dresses.
User's last utterance: What do you think it is about red that makes it so appealing?
Related knowledge: Red is the colour at the end

of the visible spectrum of light, next to orange and opposite violet.
Your knowledge-grounded response to the User's last utterance: Red is visible of light, it is next to orange and opposite to violet.
The following is a multi-round dialogue between the user and you.
User's utterance: Is rock and roll still popular today? Your response: It's hard to say. However, radio stations have much success playing classic rock and roll, which is a sub genre that usually has one or two electric guitars, a double bass or string bass or electric bass guitar, and a drum kit. User's utterance: I used to listen to the rock band Rolling Stones. Are they still around today? Your response: They are! Even though they were formed in 1962 and have had a long list of line-up changes, they're still around today, with Mick Jagger still leading the band.
User's last utterance: Wow, that is a long time to be playing music. I wonder if any other bands have been around that long.
Related knowledge: Red Hot Chili Peppers are an American funk rock band formed in Los Angeles in 1983.
Your knowledge-grounded response to the User's last utterance: It all depends! You have bands like the Red Hot Chili Peppers who, although have not reached the popularity of the Rolling Stones, have been around since 1983 themselves.
The following is a multi-round dialogue between the user and you.
User's utterance: Gouda cheese Your response: Do you know that Gouda cheese is made from cow milk?
User's last utterance: Most all cheese if made from Cow milk I think. I'm a fan of Gouda, it is pretty good.
Related knowledge: It is one of the most popular cheeses worldwide.
Your knowledge-grounded response to the User's last utterance: I bet because it is one of the most popular cheeses worldwide.
Now complete the following dialogue:
User's utterance: [Dialogue History]
User's last utterance: [User's Query]
Related knowledge: [External Knowledge]
Your knowledge-grounded response to the User's last utterance:

C Details of the Training Algorithm

See the pseudo-code below for details.

Algorithm 1 DoGe algorithm

Input: dialogue context x, related knowledge k, number of rerank candidates K, large language model $\mathcal{M}$, max new tokens N.
Output: response y.

1: **while** $y_{t-1} \neq$ <eos> or $|y| < N$ **do**
2: Get $p_\theta(v|x, y_{<t})$ from $\mathcal{M}(x, y_{<t})$
3: Calculate p_{max} through Equation 7
4: Calculate $\mathcal{H}_t$ through Equation 8
5: $\mathcal{C}_t \leftarrow \frac{1}{2} * (p_{max} + \frac{1}{\mathcal{H}_t+1}) - \lambda$
6: **if** $\mathcal{C}_t > 0$ **then**
7: $p^*(v) \leftarrow p_\theta(v|x, y_{<t})$
8: **else**
9: Calculate $q_\theta^k(v)$ through Equation 9
10: $p^*(v) \leftarrow q_\theta^k(v)$
11: $y_t = \arg\max_v p^*(v)$
12: $t = t + 1$
13: **return** y

D Case

Context	(1) Are you a basketball fan? (2) Yes, I am a fan of the five player sport. Are you? (3) Not as much as I used to be. I watched The Bulls in the 90s when they were the Dream Team.
Knowledge	Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
GT	So you were a fan of Michael Jordan on The Bulls?
Greedy	Interesting! Michael Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
Nucleus	Wow, impressive! Even though Michael Jordan played for several teams, he is mostly known for his time with the Chicago Bulls on the Dream Team.
FECS	I see. Jordan played 15 seasons in the NBA for the Chicago Bulls and Washington Wizards.
DoGe	**Oh, wow! That's amazing! Did you know that Michael Jordan spent 15 seasons in the NBA, playing for the Chicago Bulls and the Washington Wizards? He is considered one of the greatest basketball players of all time**.

Table 6: An example for case study.