

PICK: Polished & Informed Candidate Scoring for Knowledge-Grounded Dialogue Systems

Bryan Wilie, Yan Xu, Willy Chung,
Samuel Cahyawijaya, Holy Lovenia, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology
bwilie@connect.ust.hk

Abstract

Grounding dialogue response generation on external knowledge is proposed to produce informative and engaging responses. However, current knowledge-grounded dialogue (KGD) systems often fail to align the generated responses with human-preferred qualities due to several issues like hallucination and the lack of coherence. Upon analyzing multiple language model generations, we observe the presence of alternative generated responses within a single decoding process. These alternative responses are more faithful and exhibit a comparable or higher level of relevance to prior conversational turns compared to the optimal responses prioritized by the decoding processes. To address these challenges and driven by these observations, we propose Polished & Informed Candidate Scoring (PICK), a generation re-scoring framework that empowers models to generate faithful and relevant responses without requiring additional labeled data or model tuning. Through comprehensive automatic and human evaluations, we demonstrate the effectiveness of PICK in generating responses that are more faithful while keeping them relevant to the dialogue history. Furthermore, PICK consistently improves the system’s performance with both oracle and retrieved knowledge in all decoding strategies. We provide the detailed implementation in ¹.

1 Introduction

Knowledge-grounded dialogue (KGD) has been introduced as a means to ground conversation towards the provided knowledge, thereby enabling the generation of informative and engaging responses (Dinan et al., 2019; Zhou et al., 2018). Despite the advancements in training KGD systems to convincingly simulate human language on a linguistic plane, these systems still struggle with the challenge of producing responses that align with

¹<https://github.com/bryanwilie/pick>

Knowledge snippet
Due to his powerful and very large vocal range and energetic live performances, Rose has been named one of the greatest singers of all time by various media outlets, including "Rolling Stone" and "NME".
Dialogue History
Speaker 1: Didn't their guitarist slash leave the band?
Speaker 2: When did he leave did he release the six albums with them
Speaker 1: I heard Axl Rose was known for throwing tantrums
Response
Vanilla He was known for throwing tantrums
PICK He has been named one of the greatest singers of all time by various media outlets

Table 1: PICK empowers models to generate more faithful to the knowledge snippet and serve as a more appropriate reply to the conversational context. In this sample, the response prioritized by PICK is more grounded in external knowledge (highlighted in blue) than the optimal response prioritized by the decoding process (i.e., beam search) and is also more relevant to the dialogue context. Here, the vanilla response repeats the dialogue history (highlighted in red).

those human-preferred qualities. Such deficits can be attributed to various issues, e.g., hallucination as well as the lack of coherence and engagingness in the generated responses (Fu et al., 2022; Shuster et al., 2022; Rashkin et al., 2021b; Zhao et al., 2020a).

Numerous methodologies have been investigated to leverage the potential of various training and decoding methods to address these identified issues. For instance, the recent human quality alignment methods, such as Ouyang et al. (2022), hinge on collecting extensive human annotations, followed by the reward model fine-tuning to approximate human preference. This process then guides the optimization of the language model (LM) through reinforcement learning. While this approach has demonstrated promising results, it is noteworthy that accumulating such a significant volume of manual human data is highly resource-intensive in terms of both time and human labor.

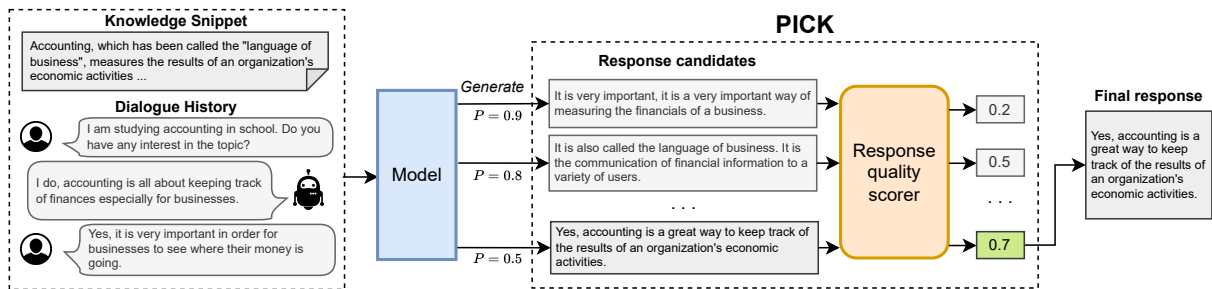


Figure 1: Overview of PICK. Instead of taking the response with the highest joint probability over the generated tokens, we select the response with the highest overall response quality score on faithfulness and relevance from the top- r responses. We propose to assess the response candidates’ quality based on the dialogue history and the corresponding knowledge without further tuning. Simple yet effective, PICK ensures better relevance and coherence of the generated response.

Through analyzing various LM generations, we observe that within one decoding process, there exist alternative generated responses that are more faithful and relevant to prior conversational turns. These candidates, however, are overlooked by the decoding processes as they are not prioritized as the optimal responses. Driven by these observations, we propose a straightforward yet effective human-aligned re-ranking framework to direct model responses closer to KGD qualities.

We introduce Polished & Informed Candidate Scoring (PICK), a generation re-scoring framework for KGD tasks, which empowers models to generate optimal dialogue responses that are more faithful to the knowledge provided and relevant to the dialogue history without requiring additional model tuning. The proposed framework is also model-agnostic; thus, it can be applied to various LMs with different architectures and sizes. Furthermore, it circumvents the need for supplementary labeled data by exploiting off-the-shelves metrics that correlate well with human judgment. While considering its contextual relevance to the dialogue history, utilizing these metrics allows the model to produce better responses. However, to enable the generation of responses that are more faithful and relevant, it is essential to condition the response on the dialogue history and accurate knowledge grounding. To do so, we explore various metrics that ensure the response is aligned with the knowledge and utilize the existing automatic metrics that correlate well with human judgment. Our experiments and human evaluation show that PICK enables models to produce responses more faithful to the provided knowledge and relevant to the dialogue history.

Our contributions to this work are three-fold. (1)

We propose PICK, a generation re-scoring framework for KGD that empowers models to generate dialogue responses that are more faithful to the provided knowledge and relevant to the dialogue history. The proposed framework is simple yet effective; it does not require further model tuning and additional labelled data for language modelling alignment. (2) We analyze the improvement from PICK-reranked responses in the systems with both oracle and retrieved knowledge and show that PICK consistently improves the performance in all decoding strategies. (3) We investigate the impact of diverse scoring metrics and decoding settings on generation quality. Then, we present the best scoring and decoding configurations for PICK on KGD tasks.

2 Related Work

Knowledge-Grounded Dialogue Dinan et al. (2019) develop a large dataset with conversations directly grounded on knowledge retrieved from Wikipedia. Alongside the work, recent works aim to build dialogue models that could conduct faithful and relevant knowledgeable discussions on open-domain topics (Li et al., 2022; Liu et al., 2021; Xu et al., 2022, 2023). Aiming to improve informativeness, a knowledge selection process is introduced to determine which specific elements of knowledge are informative to the dialogue (Kim et al., 2020; Zhao et al., 2020b). Further, Li et al. (2020) propose learning how knowledge is expressed to improve coherence and knowledge relevance. Shuster et al. (2021) utilize neural-retrieval-in-the-loop architectures to develop models that maximize knowledgeability while retaining conversational ability. Rashkin et al. (2021b) use the gold knowledge and

control the model to generate faithful and relevant responses. The PICK framework is orthogonal to recent works but similarly focused on enabling the model to generate relevant responses faithful to the provided knowledge.

Alignment of Dialogue Response Quality To exhibit alignment with dialogue response quality in the model dialogue response, several works implemented the concept of reinforcement learning from human feedback (Christiano et al., 2017) explicitly or implicitly in the dialogue domain. Jaques et al. (2019) evaluate responses for coherence and engagement using a supervised conversational evaluator with human-annotated labels. Yi et al. (2019) collect human interaction data as implicit human feedback. Hancock et al. (2019) develop an agent that would ask for feedback to improve its dialogue abilities further. Those works accumulate manual human data and are highly resource-intensive in terms of time and human labor. On the other hand, there are also works that re-rank response candidates to improve the dialogue response quality. Mei et al. (2017) utilize the Latent Dirichlet Allocation (LDA) method to learn document-level latent topics to select the best continuation based on document-level topic-level matching. Welleck et al. (2019) improve the consistency by re-rank utterances using an NLI model trained on a Dialogue NLI dataset that they created for the purpose. Ko et al. (2019) train four classifiers on synthetically generated data to re-rank plausible sentences. Unlike those works, we leverage off-the-shelf automatic metrics that correlate well with human judgment on conversation-level qualities; hence it does not require additional labeled data for the language modeling alignment. Furthermore, we devise a framework that doesn’t require further model tuning to promote faithful and relevant candidates.

3 Methodology

Knowledge-grounded dialogue (KGD) systems are built to be informative teachers. Such systems must be faithful to one or more source documents we implicitly trust and serve as an appropriate reply to the conversational context (Rashkin et al., 2021a; Zhan et al., 2021; Honovich et al., 2021). In KGD systems, a model is trained to generate a response based on the dialogue utterances with the user and ground to the knowledge snippet. We denote a KGD dataset as $\{\mathcal{D}^n\}_{n=1}^N$. At every turn t we have dialogue history at turn t denoted as

$\mathcal{D}_t = \{(U_i, S_i)\}_{i=1}^t$, where U_t is the user utterance and S_t the system response. Each of these S_t responses is grounded to knowledge snippets K_t that are retrieved from a knowledge base \mathcal{A} as illustrated in Figure 1, our proposed framework takes in input $X_t = (T_t, K_t, \mathcal{D}_{t-1}, U_t)$ with T_t resembling the conversation topic at turn t , to a fine-tuned model f_θ to generate a relevant and faithful response sequence \hat{S}_t . The concatenation of \mathcal{D}_{t-1} and U_t is a dialogue history.

3.1 Re-ranking Framework

Beam search and nucleus sampling decoding methods allow the model to generate multiple responses (i.e., hypotheses) to the same inputs. Instead of selecting the response with the highest probability from the model, we propose a re-ranking method that ensures better relevance and faithfulness of the generated responses without further tuning. Our approach treats all of the r hypotheses as a pool of r response candidates $C = \{C_1, \dots, C_r\}$ to be further ranked based on their qualities. We evaluate each response candidate with ready-to-use scorers to get its quality score of μ . Our goal is to select the best scoring candidate according to their associated scores $C_\mu = \{\mu(C_1), \dots, \mu(C_r)\}$, that is, to identify the best dialogue response candidate \hat{S}_t according to the metrics, which is given by:

$$\hat{S}_t = \arg \max_{C_j \in C} \{\mu(C_1), \dots, \mu(C_r)\}$$

3.2 Decoding Strategy

To produce the top- r response candidates $C = \{C_1, \dots, C_r\}$, we take the same input X_t to the fine-tuned model f_θ and perform generations with the number of return sequences to be r , with r being larger than 1. Each response candidate C_j is an independently computed returned sequence from the search hypotheses or random sampling. Although by both paradigms, the last $(r - 1)$ hypotheses are seen as inferior response candidates, we will later show that it is not the case and that by evaluating their qualities using the ready-to-use automatic metric, we can let the same fine-tuned model f_θ reach a more optimal response quality.

3.3 Response Quality Scorer

KGD aims to ground the conversation by generating responses that are faithful to the provided knowledge and relevant to the dialogue history. To achieve this, we leverage off-the-shelf automatic

Split	# Wizard responses	
	seen	unseen
Train	74092	-
Dev	3939	3927
Test	3865	3924

Table 2: Statistics of Wizard of Wikipedia.

metrics to evaluate the quality of response candidates. These metrics allow us to assess the faithfulness and the relevance of the responses without the need for additional labeled data. We consider the qualities of the response candidate w.r.t the dialogue history μ_D and the input knowledge snippet μ_K to construct the final quality score of μ . We elaborate on the metric corresponding to each aspect score in Section 4.2 and 4.3.

The relevance score μ_D is calculated given the response candidates and the dialogue history, $\mu_D(C_j, (\mathcal{D}_{t-1}, U_t))$, which the faithfulness score μ_K is calculated regarding the input knowledge snippet, $\mu_K(C_j, K_t)$. In this work, we consider both qualities of the response candidate equally important; thus, we derive the final quality score μ based on the sum of μ_D and μ_K . Our proposed method allows more randomness in the decoding process, which may cause high meaningless repetition in some r hypotheses. To filter this, we remove the hypotheses that contain repetitive words. We also filter hypotheses that contain a word more than 30 characters long since words that long are not likely to occur in an English general text.²

4 Experiments

4.1 Dataset and Models

We use Wizard of Wikipedia (WoW) (Dinan et al., 2019), a large-scale corpus of multi-turn knowledge-grounded dialogues between an “apprentice” and a “wizard”, to conduct our experiments in developing the KGD systems. We use the same split in (Dinan et al., 2019) as stated in (Shuster et al., 2020). We aim to produce better responses, thus we focus on only modeling “wizard” response utterances in the dialogue where they respond to the “apprentice” utterances. The data statistics of WoW are shown in Table 2. We adopt a pre-trained GPT-2 (Radford et al., 2019), and T5-small (Raffel et al., 2020) as the backbones. We

²https://en.wikipedia.org/wiki/Longest_word_in_English

fine-tune both models and limit the maximum sequence length to 512. Maximizing our GPU (RTX 2080Ti) capacity, we train the GPT-2 model in a batch size of 4 and the T5 model in a batch size of 8 for 10 epochs with early stopping patience of 3. We train using all of the training data and use the dev (seen topics) split and monitor the model’s loss in this split to do the early stopping and to choose the best model to use in the experiment. More training details are provided in §4.5.

4.2 Faithfulness Score

Faithfulness problem can also be considered an intrinsic hallucination problem for KGD tasks. Following Shuster et al. (2021), we leverage Knowledge F1 (KF1)³, calculated based on the unigram overlap between the generated response and the input knowledge snippet, to assess the faithfulness of responses. There are also other alternative n-gram-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), entailment measurement from a state-of-the-art natural language inference (NLI) model (Liu et al., 2019), and the similarity measurement from BLEURT (Selam et al., 2020). We further investigate the impact of different faithfulness scorers in Section 6.1 and find out that KF1 shows its distinct effectiveness in ensuring both faithfulness and overall performance.

4.3 Relevance Score

Overlap-based automatic evaluation metrics are known to be ineffective in distinguishing the relevance between the generated response with the dialogue history due to the one-to-many nature of dialogue (Zhao et al., 2017; Yeh et al., 2021). Therefore, we explore reference-free model-based metrics on top of them. Specifically, we utilize the FED metric (Mehri and Eskenazi, 2020) as the relevance scorer. FED is an unsupervised evaluation metric that uses DialoGPT (Zhang et al., 2020) to measure 18 fine-grained turn- and dialogue-level qualities of dialogue. It calculates the likelihood of manually designed follow-up utterances to measure multiple qualities of dialogue. Moreover, it is proven to correlate well with human judgment.

We follow the hierarchical groupings from (Phy et al., 2020) and separate the fine-grained metrics in FED between basic (w.r.t understandability) and further (w.r.t likeability) response qualities, both at the turn- and dialogue-level. At the turn level (TL),

³<https://github.com/facebookresearch/ParLAI>

Models	Test (seen topics)				Test (unseen topics)			
	BLEU-4	ROUGE-L	F1	KF1	BLEU-4	ROUGE-L	F1	KF1
Baselines								
MemNet (w/ aux loss) (Dinan et al., 2019)	1.5	-	35.5%	-	0.3	-	32.2%	-
dodecaDialogue (Shuster et al., 2020)	10	-	<u>38.4%</u>	-	9.7	-	-	-
Controlled GPT-2 (Rashkin et al., 2021b)	8.9	-	-	-	8.4	-	-	-
Controlled T5 (Rashkin et al., 2021b)	8.4	-	-	-	8.7	-	-	-
PLUG-Golden Knowledge (Li et al., 2022)	11.5	31.1	36.0%	47.8%	8.8	29.0	33.4%	46.0%
GPT-2								
Greedy	12.4	29.9	32.6%	48.6%	12.1	29.9	32.3%	46.8%
Beam search ($n = 5, r = 1$)	15.0	33.1	35.6%	64.5%	13.9	32.2	34.5%	60.3%
+ PICK ($n = 5, r = 5$)	16.6	34.1	37.0%	73.7%	15.6	33.7	36.4%	71.0%
Beam search ($n = 10, r = 1$)	15.4	33.4	35.8%	68.9%	14.3	32.5	34.7%	64.5%
+ PICK ($n = 10, r = 10$)	16.7	34.5	37.4%	80.4%	16.0	34.5	37.1%	78.2%
Top- k sampling ($k = 3, r = 1$)	8.7	26.3	29.0%	39.7%	8.1	25.5	28.2%	37.8%
+ PICK ($k = 3, r = 10$)	14.9	33.0	36.2%	67.6%	14.2	32.7	35.6%	64.6%
Top- p sampling ($p = 0.5, r = 1$)	11.5	28.3	31.2%	46.2%	10.4	27.6	30.1%	43.5%
+ PICK ($p = 0.5, r = 10$)	16.0	34.1	37.2%	72.7%	15.2	34.0	36.9%	70.2%
T5								
Greedy	14.7	33.0	35.6%	56.0%	14.4	32.4	35.0%	56.2%
Beam search ($n = 5, r = 1$)	16.3	34.8	37.7%	77.8%	15.6	34.7	37.4%	78.8%
+ PICK ($n = 5, r = 5$)	16.3	34.9	37.8%	79.6%	15.6	34.8	37.6%	80.2%
Beam search ($n = 10, r = 1$)	16.2	34.8	37.7%	81.8%	15.5	34.7	37.5%	82.7%
+ PICK ($n = 10, r = 10$)	16.1	34.8	37.7%	84.3%	15.4	34.8	37.6%	84.8%
Top- k sampling ($k = 3, r = 1$)	11.7	30.3	33.3%	49.3%	11.4	29.9	32.8%	49.8%
+ PICK ($k = 3, r = 10$)	15.9	34.2	37.3%	71.8%	15.2	34.0	37.0%	72.4%
Top- p sampling ($p = 0.5, r = 1$)	13.9	32.3	35.1%	55.3%	14.0	31.9	34.6%	55.6%
+ PICK ($p = 0.5, r = 10$)	16.9	34.9	38.0%	74.6%	16.4	34.7	37.5%	74.7%

Table 3: Overall performance comparisons. PICK significantly improves the performances of all models and decoding methods, even on the top- k and top- p sampling that gained low automatic metrics scores on their vanilla responses. The best performances in each section are in **bold**, while the overall best is underlined.

we group semantically appropriate, understandable, and fluent as turn-level metrics that measure the basic qualities of responses. We see the additional qualities as the ones that make the response more likeable and group the interesting, engaging, specific, relevant, and correct measurements into one. Similarly, at the dialogue level (**DL**), we group coherent, error recovery, consistent, and diverse as the dialogue-level basic qualities of responses. At the same level, we also group depth, likeable, understandable, flexible, informative, and inquisitive as dialogue-level metrics that measure the further qualities of responses. On top of that, we also experiment with combining each level of metrics and all of the measurements to find the best combination to produce responses relevant to the dialogue history.

We also explore another reference-free model-based metric, USL-H (Phy et al., 2020), for comparison. USL-H combines three models trained to

determine whether a response is valid and grammatically correct and to evaluate the sensibleness and the likelihood of a given response. The analysis is included in Section 6.1.

4.4 Baselines

We select baseline models that utilize gold knowledge snippets in their generation process. We take the performances of MemNet (Dinan et al., 2019), dodecaDialogue (Shuster et al., 2020), GPT-2 and T5 with control code and resampling (Rashkin et al., 2021b), and PLUG-Golden Knowledge (Li et al., 2022) as our baselines. We also experiment with PICK in the settings where the provided knowledge is retrieved instead of using the oracle knowledge. We leverage KnowledGPT Zhao et al. (2020b) and perform a similar procedure elaborated in Section 3, to select responses that are more relevant to the dialogue history and the ones that are more faithful to the retrieved knowledge. We di-

Inference methods	Test (seen topics)				Test (unseen topics)			
	BLEU-4	ROUGE-L	F1	KF1	BLEU-4	ROUGE-L	F1	KF1
Greedy	<u>5.8</u>	20.3	22.0%	56.4%	4.7	19.0	20.4%	53.6%
Beam search ($n = 10, r = 1$)	2.5	12.6	14.4%	47.3%	2.1	12.0	13.6%	45.6%
+ PICK ($n = 10, r = 10$)	2.5	12.8	14.7%	47.9%	2.1	12.1	13.9%	46.1%
Top- k sampling ($k = 3, r = 1$)	4.4	18.5	20.5%	46.7%	3.7	17.6	19.4%	45.3%
+ PICK ($k = 3, r = 10$)	5.4	19.5	21.3%	70.4%	4.7	18.5	20.1%	67.9%
Top- p sampling ($p = 0.3, r = 1$)	5.6	20.1	21.8%	56.0%	4.6	18.8	20.3%	52.9%
+ PICK ($p = 0.3, r = 10$)	5.8	20.4	22.1%	68.9%	4.9	19.3	20.9%	66.2%

Table 4: Results of KnowledGPT (Zhao et al., 2020b) with PICK and different inference methods. Although other decoding methods underperform the greedy baseline, our method still improves over each of them. Here, KF1 is calculated w.r.t the retrieved knowledge instead of the gold knowledge. The best performances in each section are in **bold**, while the overall best is underlined.

rectly utilize the codes and models provided⁴ and adjust the generation parameters as required.

We note the **vanilla** responses performance as our lower bound: the responses performance on each decoding method without PICK (i.e. taking the top-1 hypotheses from beam search).

4.5 Training details

During training, the concatenated utterances are delimited using speaker ID of either <speaker1> or <speaker2>, and the concatenations of the topic, knowledge snippet, and the utterances are separated by a separator token \n. We experiment using learning rates (lr) of $1e-5$, $5e-4$, $1e-4$, $5e-4$ to fine-tune the models, and then we pick the models with the lowest loss on the dev (seen topics) split to be the models that we will be using throughout the experiments. Ultimately, the best GPT-2 model is fine-tuned with lr of $1e-5$ and the T5 with $5e-4$.

4.6 Evaluation

Automatic Metrics We evaluate the final response qualities by comparing them to the gold responses. We perform the automatic evaluation using BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and unigram-F1³. We implement the BLEU measurements following Rashkin et al. (2021b). To make a fair comparison with the previous work, we utilize BLEU-4 scoring as it is implemented in Rashkin et al. (2021b) and ROUGE-L scoring (the mean F1 measures) as it is implemented in ⁵. Further, we also use KF1 as stated in Section 4.2 for faithfulness measurement.

⁴<https://github.com/zhaoxlpu/KnowledGPT.git>

⁵<https://huggingface.co/spaces/evaluate-metric/rouge>

Human Evaluation We conduct manual evaluations to measure the qualities of the generated responses from two aspects: *Faithfulness* and *Relevance*. We take 100 random generation samples from the test (unseen topics) split and ask crowd-sourced annotators⁶ to evaluate on a 4-point Likert scale from 1 (low quality) to 4 (high quality). We ask for three level-1 (all-kinds) contributors and three level-3 (experienced only) contributors and report their average scores. The complete annotation guideline is attached in Appendix E.

5 Results

5.1 Results with Oracle Knowledge

Overall, as observed in Table 3, the proposed method achieved significantly better performances than the baselines from the previous works, especially in the comparison of BLEU-4 scores. In this table, the PICK responses are all re-ranked based on the sum of FED turn-level basic metrics and KF1. The proposed method significantly improves the performances of all models and decoding methods in all of the BLEU-4, ROUGE-L, F1, and KF1 metrics. Interestingly, for the top- k and top- p sampling decoding that previously gained low scores BLEU-4, ROUGE-L, F1, and even KF1 on their vanilla responses, there exist alternative responses that have a more similar quality to the gold response and our proposed re-ranking and scoring framework promotes that.

All the re-ranked responses also obtained a huge increase of KF1 compared to their vanilla baseline, especially in the top- k and the top- p sampling, where the vanilla KF1 is far much lower than the re-ranked response’s KF1. Although the scores of

⁶<https://appen.com/>

Models	Faithfulness	Relevance
Gold responses	2.64	2.09
GPT-2		
Beam search ($n = 10, r = 1$)	2.26	1.48
+ PICK ($n = 10, r = 10$)	2.49*	1.86*
T5		
Beam search ($n = 10, r = 1$)	2.51	1.87
+ PICK ($n = 10, r = 10$)	2.56	2.01*

Table 5: PICK enables models to produce responses more faithful to the provided knowledge and relevant to the dialogue history, as shown from the human evaluation on responses from GPT-2 and T5 models. In each section, * indicates that this result is significantly better (p -value < 0.05) from their respective baseline comparison. See Figure B for a detailed visualization of the Likert score distribution.

BLEU-4, ROUGE-L, and F1 are comparable in the inference of the T5 model using beam search, we can still see the improvement in KF1, signifying that the proposed method better addresses the use of knowledge in its response. Lastly, PICK also closes the performance gap between unseen and seen evaluation, as it is especially observed in the GPT-2 model performances. We provide samples of the responses in Table A1.

5.2 Results with Retrieved Knowledge

We reproduce the greedy decoding performance similar to what was reported in (Zhao et al., 2020b). Here, we use the same scoring metrics on PICK as in §5.1. Although other decoding methods’ underperform in comparison to greedy in the knowledge retrieval setting (see Table 4), PICK still shows improvement over each decoding method taken individually as the BLEU-4, ROUGE-L, F1, and KF1 scores all increase. Interestingly, although our method improves the generated response’s performance compared to its vanilla counterpart, the top performances are still comparable with the greedy decoding performance, except for the significantly better KF1 on the responses re-ranked by PICK. It is important to note that the KF1 here is calculated w.r.t the retrieved knowledge instead of the gold knowledge. We also investigate the underperformance of KnowledGPT being inferred through decoding methods other than greedy, and we found that the issue persists through repeated trials in different settings. We leave this issue out of the scope of this paper.

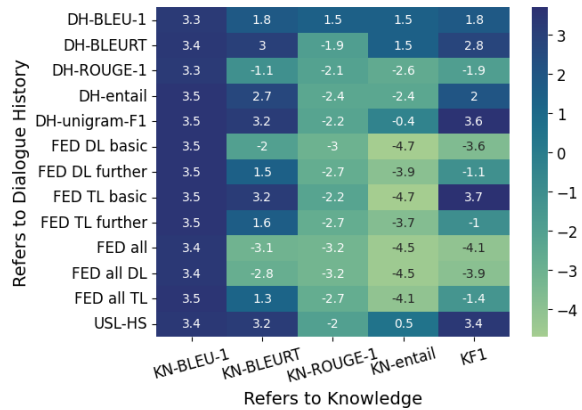


Figure 2: Comparing combinations of automatic evaluation of response qualities w.r.t dialogue history and knowledge snippet, PICK with FED turn-level basic metrics and KF1 produced responses with the best qualities. The comparisons are shown here as a heatmap of the sum of mean-normalized BLEU-4, ROUGE-L, F1, and KF1 w.r.t gold responses. The x-axis labels the knowledge-oriented metrics, and the y-axis labels the dialogue-history-oriented metrics used in the comparisons. TL denotes turn level, and DL denotes dialogue level.

5.3 Human Evaluation

We conduct the manual evaluation on responses from GPT-2 and T5 models decoded with beam search 10. We compare both the vanilla ($r=1$) and the PICK responses ($r=10$) and evaluate the quality of the responses in the aspect of *Relevance* and *Faithfulness*. Table 5 shows that PICK responses from GPT-2 and T5 models are more faithful and relevant than the vanilla generations. These findings correlate well with the automatic results shown in Table 3, where for both of the responses, PICK responses achieved higher BLEU-4, ROUGE-L, F1, and KF1 scores in comparison to the vanilla responses. We attach a detailed visualization of the Likert score distribution in Figure B.

It is also known from previous works that attempts to make the system more faithful usually lead to trade-offs between the response’s relevance and faithfulness scores (Rashkin et al., 2021b), either due to the response being not quite as pertinent to the previous conversation turns or it gives overly extractive responses. This result also showcases the merit of the proposed re-ranking method to improve the faithfulness of the responses without a trade-off on the response’s relevance towards the previous conversation turn.

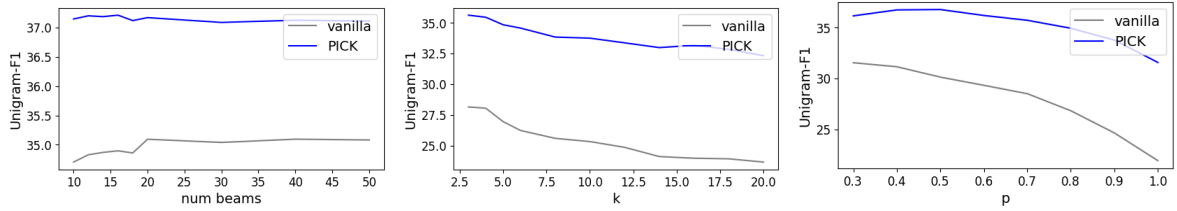


Figure 3: Performance comparison with unigram-F1 while varying the number of beams, k and p in respectively beam search, top- k and top- p sampling with r kept at 10. Extending the number of beams does not help, but k and p threshold help to mitigate a bad response formation.

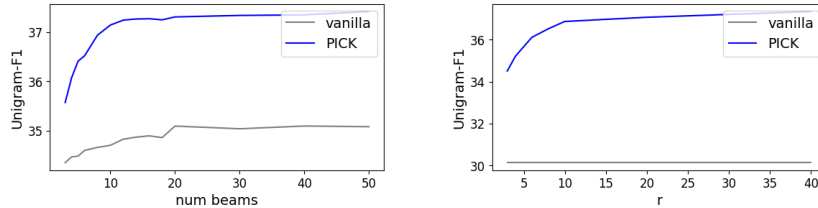


Figure 4: Performance comparison with unigram-F1 while varying r to follow the number of beams (left figure) and varying r on nucleus sampling with $p = 0.5$ (right figure). Increasing r in both the search and sampling decoding experiments promotes the existence of responses that are more similar to the gold response.

6 Analysis and Discussion

6.1 Rescoring Metrics

We study the effectiveness of automatic metrics explored in Section 4.2 and 4.3. We employ GPT-2+PICK with each automatic metric as the scoring method by normalizing the performances on each of their mean and standard deviations for a fairer comparison. We report the normalized comparison as a heatmap in Figure 2. The best performance is achieved by using FED turn-level that measures the basic qualities of responses w.r.t the dialogue history (FED turn-level basic) alongside KF1.

6.2 Decoding Strategy

We perform ablation of GPT-2+PICK and vanilla on the decoding methods to find the optimum configurations with our proposed method. For beam search, we increase the number of beams starting from 10. For the top- k and nucleus sampling, we increase the threshold of k and p to perform the sampling, starting from $k = 3$ and $p = 0.3$. For each experiment, we keep $r = 10$, and we note the comparison of performances w.r.t the gold as a probe towards the degree of the responses being in a certain desired quality range that the gold responses reflect (Figure 3).

From Figure 3, We observe that increasing the number of beams does not improve the performance of the generated responses, while as the p and k increase, the performances degrade. We

conjecture that loosening the respective p and k sampling thresholds weakens the mitigation of bad continuations and, in turn, produces worse generations (i.e., generations with considerably low perplexity (Holtzman et al., 2019)). The two sampling strategies show different optimum thresholds, as top- k sampling responses start to deteriorate with k larger than 3, while for the nucleus sampling, we see benefits in relaxing the p threshold to 0.5. We conjecture that due to the probability selection of the nucleus sampling, top- p tokens within $p \leq 0.5$ could still be reliable, as it produces generations with considerably low perplexity as mentioned in (Holtzman et al., 2019).

6.3 Number of Return Sequences (r)

We also perform further ablation of GPT-2+PICK and vanilla on varying r . We extend r on the nucleus sampling with $p = 0.5$, and on the beam search, we set the r to follow the number of beams n to retain the top n choices when a new token in the sequence is generated. The performance comparisons are noted in Figure 4.

Our observation shows that increasing r in both the search and sampling decoding experiments promotes the existence of responses that are more similar to the gold response. With the gold response holding the desired qualities we aim to achieve, these findings also indicate that increasing the r could help increase the response qualities generated

by the same model to some extent. Figure 4. shows that PICK response performance begins to saturate around $n = 10$ and $r = 10$ in the beam search and nucleus sampling experiment, most likely because the best candidate response is consistently found within that range of return sequences.

6.4 Error Analysis

To better understand our method’s limitation, we provide manually sampled study cases with GPT-2+PICK using beam search ($n = 50$, $r = 50$), in which better responses are not selected by the scorer. We observe three kinds of errors. First, the current metric fails to promote the selection of better responses. We conjecture that this happens due to the low correlation of the automatic metrics towards the human judgments (Yeh et al., 2021); hence, implementing better human preference metrics will aid in better response promotion.

Second, in some cases, substandard responses are selected due to their high overlap with the knowledge snippet. This could be because the metrics used rely on the spurious correlation between attribution and word overlap and thus do not reliably distinguish attributable abstractive responses (McCoy et al., 2019; Dziri et al., 2022). We perform a further ablation study on this error in Appendix D. Third, the knowledge snippets provided are irrelevant to the dialogue history. We provide these entries of case study samples in Table A2.

7 Conclusion

This work investigates the alignment of KGD responses to faithfulness and relevance. We propose PICK, a straightforward yet effective generation re-ranking framework for KGD. PICK is model-agnostic, does not require further model tuning, nor requires additional labelled data for the language modelling alignment. Experimental results show that the proposed method enables models to produce better responses that are more faithful to the provided knowledge and relevant to the dialogue history.

Acknowledgements

We thank the anonymous reviewers for their valuable and constructive comments. We thank Tiezheng Yu for the insightful discussions. This work has been partially funded by the PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong, and the Hong Kong

Fellowship Scheme by the Hong Kong Research Grants Council (RGC).

Limitations

While our proposed method, PICK, is model-agnostic and can be adopted by various model architectures, our exploration in this work is limited to GPT-2 and T5. Generating multiple alternative responses within a single decoding process also can increase the computational overhead of the KGD system, thus ways to increase PICK’s efficiency would be beneficial in the future. Additionally, while PICK improves the faithfulness and relevance of responses, it may not address other challenges in knowledge-grounded dialogue systems such as self-consistency, engagingness, long-term coherence, and more. Further research is needed to explore these limitations and develop more comprehensive approaches for generating better responses. We leave these explorations open for future work.

Ethics Statement

In this paper, we propose a re-ranking framework, targeting better correlating the final generation with some concrete attributes of the response. However, our work has a broader impact given the current popularity of ChatGPT. ChatGPT replies on a reward model to model human feedback for reinforcement learning. However, the training of the reward model requires a huge amount of human annotations, which is time- and resource-consuming. Then it comes to a question -what is the expression of human preference and whether it is possible to model human preference without heavy human annotations? Though far from perfect, we take an initial step in this direction by exploring the usage of automatic metrics to re-rank the responses. We believe it is a promising and valuable research topic.

References

- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

- Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.
- Tingchen Fu, Xueliang Zhao, Chongyang Tao, Ji-Rong Wen, and Rui Yan. 2022. There are a thousand hamlets in a thousand people’s eyes: Enhancing knowledge-grounded dialogue with personal memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3901–3913.
- Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.
- Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: 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.
- Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.
- Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *International Conference on Learning Representations*.
- Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3456–3466.
- Linxiao Li, Can Xu, Wei Wu, Yufan Zhao, Xueliang Zhao, and Chongyang Tao. 2020. Zero-resource knowledge-grounded dialogue generation. *Advances in Neural Information Processing Systems*, 33:8475–8485.
- Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. 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, pages 206–218*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the 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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2262–2272.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.
- Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235.
- Hongyuan Mei, Mohit Bansal, and Matthew Walter. 2017. Coherent dialogue with attention-based language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- 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.
- Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178.
- 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.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits

- of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021a. Measuring attribution in natural language generation models. *arXiv preprint arXiv:2112.12870*.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. 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)*, pages 704–718.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.
- Kurt Shuster, Da Ju, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2453–2470.
- Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. *Findings of the Association for Computational Linguistics: EMNLP 2022*.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.
- Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 93–107.
- Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. Kilm: Knowledge injection into encoder-decoder language models. *arXiv preprint arXiv:2302.09170*.
- Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33.
- Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75.
- Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.
- Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.
- Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao, and Rui Yan. 2020a. Low-resource knowledge-grounded dialogue generation. In *International Conference on Learning Representations*.
- Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020b. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390.
- Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

A Dialogue samples

We provide the dialogue samples in Table A1.

Knowledge snippet	
Due to his powerful and very large vocal range and energetic live performances, Rose has been named one of the greatest singers of all time by various media outlets, including "Rolling Stone" and "NME".	
Dialogue History	
Speaker 1:	Didn't their guitarist slash leave the band?
Speaker 2:	When did he leave did he release the six albums with them
Speaker 1:	I heard Axl Rose was known for throwing tantrums
Response	
Vanilla	He was known for throwing tantrums
PICK	He has been named one of the greatest singers of all time by various media outlets
Gold	Maybe that was due to his powerful and large vocal range
Knowledge snippet	
His music career began there in 1954, recording at Sun Records with producer Sam Phillips, who wanted to bring the sound of African American music to a wider audience.	
Dialogue History	
Speaker 1:	I like to listen to Elvis! Have you heard him?
Speaker 2:	Yeah my favorite song of his is "Hound Dog" written by Jerry Leiber.
Speaker 1:	Yes, I love that song! I really like his version of "My Way"
Response	
Vanilla	He started his career in 1954, and recorded his first album in the early 60s.
PICK	His music career began in 1954, he was a producer and he wanted to bring the sound of African American music to a wider audience.
Gold	Elvis started his music career in 1954 recording at Sun Records with producer Sam Phillips.
Knowledge snippet	
, the books have sold more than 500 million copies worldwide, making them the best-selling book series in history, and have been translated into seventy-three languages.	
Dialogue History	
Speaker 1:	I liked it because of all the different animals, I am in love with animals, and the ones they made were so unique!
Speaker 2:	Harry Potter has attracted a wide adult audience, as well as younger readers. If you go to Universal you will see whole families dressed up!
Speaker 1:	I know everyone was young when everyone was into it, it was huge when I was young, like all the rage
Response	
Vanilla	It's hard to imagine how much the books have sold worldwide!It must be so popular now!
PICK	I agree, the books have sold more than 500 million copies worldwide, and they have been translated into seventy-three languages.
Gold	The books have sold 500 million copies worldwide, making it the best-selling book series in history. Wow that is impressive!

Table A1: Samples of the responses from the beam search (n=10), top- k sampling, and nucleus sampling generated by GPT-2 using the vanilla, re-ranked, and oracle methods. We also attach the gold responses for each dialogue. The vanilla response is not as faithful w.r.t. the knowledge snippet, especially compared to the other 3 responses. The re-ranked responses, however, are more faithful and yet still relevant to the dialogue history as they are not a verbatim copy of the knowledge snippet.

B Details of Human Evaluation Results

We further show the distribution of Likert scores in Figure A1 in addition to the average reports.

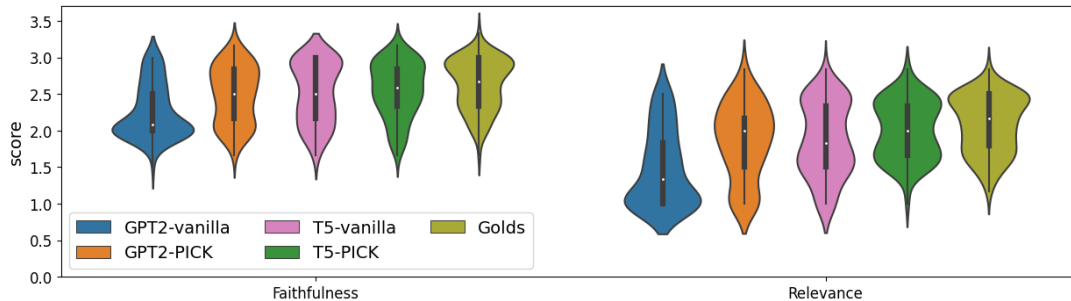


Figure A1: Distribution of human evaluation's Likert score on Faithfulness and Relevance

C Error analysis

We provide the dialogue samples of our error analysis in Table A2.

Knowledge snippet	
The album has sold approximately 30 million copies worldwide, including 18 million units in the United States, making it the best-selling debut album of all time in the US, as well as the eleventh best-selling album in the United States.	
Dialogue History	
Speaker 1: I totally forgot about that song! Didn't listen to it for a while!	
Speaker 2: You know that album sold 30 million copies!	
Speaker 1: Really? That's insane! I didn't know that	
Response	
PICK	The album has sold approximately 30 million copies worldwide, including 18 million units in the United States, making it the best-selling debut album
Other candidate	It is the best selling debut album of all time!
Gold	Maybe that was due to his powerful and large vocal range
Knowledge snippet	
Thierry Daniel Henry (; born 17 August 1977) is a retired French professional footballer who played as a forward and is the second assistant manager of the Belgium national team.	
Dialogue History	
Speaker 1: No i didn't know that! that's pretty cool!	
Speaker 2: Yes, he actually played for Monaco, Barcelona, New York, and Arsenal, where he was the club's all-time record goalscorer!	
Speaker 1: Wow! I bet he is really good at what he is doing. How old is he?	
Response	
PICK	Thierry Henry is a retired French professional footballer who played as a forward and is the second assistant manager of the Belgium national team
Other candidate	He was born on August 17, 1977.
Gold	He was born on 17 August 1977 which makes him 40 years old. Do you like watching soccer?
Knowledge snippet	
Dylan Lauren (born May 9, 1974) is an American entrepreneur.	
Dialogue History	
Speaker 1: I love chocolate bars. Have you ever heard of Dylan's Candy Bar?	
Speaker 2: I have heard of it, the daughter of Dylan Lauren owns it, but she is a fashion designer	
Speaker 1: Who is Dylan Lauren?	
Response	
PICK	Dylan Lauren (born May 9, 1974) is an American entrepreneur.
Other candidate	She is an american entrepreneur!
Gold	She is an american entrepreneur!
Knowledge snippet	
The first match of American football was played on November 6, 1869, between two college teams, Rutgers and Princeton, under rules based on the association football rules of the time.	
Dialogue History	
Speaker 1: Same here. My favorite team is the Saints!	
Speaker 2: The first football game was played on November 6, 1869.	
Speaker 1: Wow! I did not know it was that old. Do you know who the teams were?	
Response	
PICK	The first match of American football was played on November 6, 1869 between two college teams, Rutgers and Princeton, under rules based on the association football rules of the time.
Other candidate	The teams were Rutgers and Princeton.
Gold	The teams were Rutgers and Princeton. Two college teams.
Knowledge snippet	
The maximum score is 300, which is achieved by getting 12 strikes in a row.	
Dialogue History	
Speaker 1: That's cool! So I can just get started again. :) Do you know much about bowling?	
Speaker 2: I am knowledgeable. In bowling your target is to knock all pins in the lane.	
Speaker 1: Oh yeah? What happens if you don't hit any pins?	
Response	
PICK	The maximum score is 300 which is achieved by getting 12 strikes in a row
Other candidate	If you don't hit any pins, you get 12 strikes in a row.
Gold	If you will make a strike when you don't hit any pins.
Knowledge snippet	
To Kill a Mockingbird is a novel by Harper Lee published in 1960.	
Dialogue History	
Speaker 1: Oh wow, a Jack of all Trades it seems. Do you have a favorite novel? I'd love a recommendation!	
Speaker 2: His first, "A Time to Kill" is older, it was published in 1989 but he spent four years writing it and in my opinion, it's his best.	
Speaker 1: 4 years is an incredible amount of time to spend on a book, I bet it is good. I will definitely check it out! I imagine it's a thriller about murder?	
Response	
PICK	It was published in 1960 by Harper Lee.
Other candidate	I'm not sure, but I do know that to Kill a Mockingbird is a book by Harper Lee.
Gold	Yes, another book I like is To Kill a Mockingbird but that was written by Harper Lee. That is also about a crime

Table A2: Six study case samples from the GPT-2+PICK using beam search ($n = 50$, $r = 50$). The first two samples show errors caused by the current metric failing to promote the selection of better responses seen in the other candidate. The next two samples show errors caused by the selection of worse responses due to high overlap towards the knowledge snippet. The last two sample errors were because of the utilization of irrelevant provided knowledge snippets. The other candidate is one of the best response alternatives that are not selected either by vanilla or even by PICK.

D Re-ranking responses w.r.t the dialogue history, the knowledge, or both

We also perform an ablation study of doing the re-ranking with different considerations on the scoring method: only considering the response quality w.r.t the dialogue history (using FED turn-level basic metrics), the knowledge (using unigram-F1 to the knowledge), or both (see Table A3). Re-ranking the response candidates considering only the qualities w.r.t the dialogue history results in generations that are less knowledgeable and perform worse across the board. On the contrary, considering only how much of the response refers to the knowledge snippets promotes more responses that are direct copies of the knowledge. While considering both the response candidate’s qualities w.r.t the dialogue history and the knowledge snippet produces similar response quality per their automatic metrics performances, the responses contain far less direct knowledge copies within them. This is a desirable result since the direct knowledge copies responses could be considered responses less relevant to the dialogue as they would be perceived as very rigid non-human-like responses. This is also an interesting finding, that the combination of both qualities in consideration can alleviate the failure of overlap-based metrics to rank abstractive but faithful responses as previously observed in (Dziri et al., 2022).

Models	Test (seen topics)					Test (unseen topics)				
	BLEU-4	ROUGE-L	F1	KF1	% Kn-copy	BLEU-4	ROUGE-L	F1	KF1	% Kn-copy
Beam search	15.4	33.4	35.8%	68.9%	24.3%	14.3	32.5	34.7%	64.5%	15.3%
+ PICK w.r.t dialogue history	10.9	31.7	33.7%	54.7%	0.8%	10.4	30.9	32.8%	52.2%	0.6%
+ PICK w.r.t knowledge	16.9	34.5	37.4%	81.1%	36.1%	16.0	34.4	37.1%	78.8%	26.6%
+ PICK w.r.t both	16.7	34.5	37.4%	80.4%	15.5%	16.0	34.5	37.1%	78.2%	13.7%

Table A3: Ablation on PICK w.r.t dialogue history, knowledge, or both. PICK that considers both references produce similar performances but with responses containing less direct knowledge copies within them (i.e. 15.5% vs 36.1%) compared to only conditioning it to the knowledge. We use beam search with ($n = 10$, $r = 10$). The best performances in each section are in **bold**.

E Annotation guideline

We gave the following annotation instructions, guidelines, and examples to the human evaluator for them to follow.

Given the same dialogue context (external knowledge + conversation history) between the user and assistant, the AI assistant generates a dialogue response.

Please read the context and response and determine whether the response is relevant, faithful to external knowledge, and whether you liked the response if you are the user.

Please also click the "[Show job instructions](#)" and read the examples and the reasoning carefully.

External knowledge

Conversation history

User :

Assistant :

User :

Response

(a)

QUESTION | Multiple Choice

Is this response a relevant reply to the previous utterance in the conversation?

- The response is irrelevant and does not address the dialogue context
- The response is somewhat relevant but may lack important details or partially address the dialogue context
- The response is relevant and addresses the dialogue context well.
- The response is mostly relevant and thoroughly addresses the dialogue context

QUESTION | Multiple Choice

How faithful is the response?

- The response has verifiable knowledge content, but it is not supported by the knowledge document at all
- The response has no verifiable knowledge content
- The response is partly supported or partly conveys the information in the knowledge document
- The response is mostly supported and mostly conveys the information in the knowledge document

(b)

Figure A2: Main instruction (a) and annotation guidelines (b) for the human evaluation.

Job Instructions

Overview

In this job, you will be presented with dialogue context (external knowledge + conversation history) between the user and assistant, and the response generated by the AI assistant. Review the response to determine the relevancy of the response, whether the response is faithful to the external knowledge, and whether you liked the response if you are the user.

Please check the reference examples of conversations below.
Please also see the answers and reasons below.

Steps

1. Read the external knowledge and the conversation history
2. Read the response
3. Determine if the response is relevant to the conversation history by seeing and choosing the criteria available in the options
4. Determine if the response is faithful to the external knowledge by seeing and choosing the criteria available in the options

Examples

(a)

Examples

External knowledge

Presley was born in Tupelo, Mississippi, and relocated to Memphis, Tennessee, with his family when he was 13 years old.

Conversation history

User: yes...who doesn't love Elvis Presley and his music?

Assistant: Seriously. I have been planning a vacation to Graceland myself. I have family in Tennessee so it could be good.

User: I would love to tour his home. Which one of his songs is your favorite?

Response

He was born in Tupelo, Mississippi and moved to Memphis, Tennessee when he was 13 years old.

Questions

Is this response a relevant reply to the previous utterance in the conversation?

- The response is somewhat relevant but may lack important details or partially address the dialogue context

Reason:The user and assistant are discussing some things about Elvis Presley. The response is somewhat relevant since it continues the discussion about touring Elvis Presley's home, but it only partially address the dialogue context as it doesn't respond to the part that discussing about his songs.

How faithful is the response?

- The response is partly supported or partly conveys the information in the knowledge document

Reason:The response is supported by external knowledge, but it partially conveys the part that Elvis Presley relocated "with his family" when he was 13.

How much do you like the response if you are the user?

- The response is acceptable

Reason:The response engages the user through the information it offers, but it is not that interesting of a response since it ignores the question in the discussion.

(b)

Figure A3: Additional instructions of human evaluation (a) and the examples of annotation (b).