RHO (ρ): Reducing Hallucination in Open-domain Dialogues with Knowledge Grounding

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu,
Bryan Wilie, Min Zeng, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Hong Kong University of Science and Technology
zjiad@connect.ust.hk, pascale@ece.ust.hk

Abstract

Dialogue systems can leverage large pre-trained language models and knowledge to generate fluent and informative responses. However, these models are still prone to produce hallucinated responses not supported by the input source, which greatly hinders their application. The heterogeneity between external knowledge and dialogue context challenges representation learning and source integration, which further contributes to unfaithfulness. To handle this challenge and generate more faithful responses, this paper presents **RHO** (ρ) utilizing the representations of linked entities and relation predicates from a knowledge graph (KG). We propose (1) local knowledge grounding to combine textual embeddings with the corresponding KG embeddings; and (2) global knowledge grounding to equip **RHO** with multi-hop reasoning abilities via the attention mechanism. In addition, we devise a response re-ranking technique based on walks over KG sub-graphs for better conversational reasoning. Experimental results on OpenDialKG (Moon et al., 2019) show that our approach significantly outperforms state-of-the-art methods on both automatic and human evaluation by a large margin, especially in hallucination reduction (17.54% in FeQA (Durmus et al., 2020)).[1]

1 Introduction

An open-domain dialogue system aims to automatically interact with humans with sensible and informative responses. To produce such responses, knowledge-grounded dialogue (KGD) systems are established, which leverage external knowledge such as knowledge graphs (KGs) (Yu et al., 2022; Zhao et al., 2020). Despite impressive progress in general linguistic performance, KGD systems are still prone to the hallucination problem in which the generated response is nonsensical or unfaithful to dialogue history or external knowledge (Ji et al., 2022a; Roller et al., 2021; Mielke et al., 2022). Two types of hallucinations may exist: intrinsic hallucination (the generated response is contradicted by the dialogue history or the external knowledge) and extrinsic hallucination (the generated response is hard to be verified with the dialogue history and external knowledge) (Ji et al., 2022a; Dziri et al., 2021). As the example in Figure 1, this issue undermines dialogue systems' performance or raises safety concerns in real-world applications. For instance, the recently emerged foundation model ChatGPT suffers from this hallucination problem (OpenAI, 2023), especially extrinsic hallucinations. OpenAI currently uses huge amounts of human feedback to fix many ChatGPT errors, which is labor-intensive. It would be beneficial to reduce such errors automatically in advance.

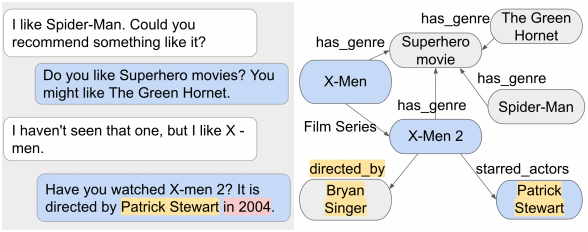

Figure 1: An example of hallucinated KGD. Based on the dialogue history and the KG, the system is expected to generate a response. In the response, "Patrick Stewart" contradicts the fact <X-Men 2, directed_by, Bryan Singer>, which is a case of intrinsic hallucination (in yellow); "in 2004" cannot be verified based on the given information, which is a case of extrinsic hallucination (in pink).

The heterogeneity between external knowledge and textual dialogue content makes it challenging for neural response generation models to learn the representation and correlation in the input source (Li et al., 2022; Zhang et al., 2019; Liu et al., 2020; Ji et al., 2022a). This challenge could further result in a hallucinated generation that deviates from the source. Previous studies have shown that the hallucination problem in KGD

[1]The source code can be obtained from https://github.com/ziweiji/RHO

Findings of the Association for Computational Linguistics: ACL 2023, pages 4504–4522

can be mitigated by retrieved knowledge augmentation (Shuster et al., 2021), control code (Rashkin et al., 2021; Wu et al., 2021), and response post-processing (Dziri et al., 2021). However, these works do not emphasize handling the discrepancy between lexical dialogue and structured knowledge information for the harmony of their fusion. The interaction mechanism between external knowledge and dialogue context should also be clarified.

In order to address this issue and take full advantage of lexical and knowledge information, we present **RHO** (ρ) [2] for faithful open-domain dialogue response generation with enhanced knowledge grounding and re-ranking. A high-level framework is illustrated in Figure 2. Specifically, **RHO** first learns the structured embeddings of all entities and relation predicates from the KG and links their mentions in the dialogue context to the KG. In the encoder-decoder model, the representations of all the linked entities and relations are grounded by KG embeddings, both *locally* and *globally*. Here, *local knowledge grounding* refers to the process where an entity or relation predicate receives and fuses its KG embedding only. While in *global knowledge grounding*, each entity or relation predicate attentively learns the knowledge from the entire sub-graph stored in a memory bank (Vaswani et al., 2017), which assigns dynamic weights of each triple equipping **RHO** with multi-hop reasoning abilities. These two knowledge groundings help the model effectively encode and inject the knowledge information from context-related sub-graph with proper attention. In addition, we re-rank the generated responses according to the hallucination degree. This technique utilizes conversational reasoning to enforce the whole conversation to follow the knowledge traverses throughout KG.

In the experiments, we show that **RHO** outperforms state-of-the-art (SOTA) (Dziri et al., 2021) on the OpenDialKG (Moon et al., 2019) dataset by a large margin: improving 17.54% in FeQA (Durmus et al., 2020), and reducing 32.93% hallucinations according to human evaluation. In particular, the responses have a broader coverage of entities and relations in the KG, demonstrating higher faithfulness of responses. The quantitative and qualitative analysis further shows its effectiveness in reducing hallucination while not sacrificing conversational abilities. In summary, the major contributions of this work are threefold:

- We propose the **RHO** model, which leverages the structured knowledge in KGs to mitigate the hallucination problem in dialogue response generation.
- To improve faithfulness, we introduce local and global knowledge grounding techniques (from a context-related knowledge sub-graph) into dialogue generation and further utilize a conversational reasoning model to re-rank the generated responses.
- We conduct a thorough faithfulness analysis via automatic and human evaluation, and empirically demonstrate that **RHO** substantially reduces intrinsic and extrinsic hallucinations in the KGD generation task.

2 Related Work

2.1 Hallucination Reduction in KGD

Researchers have been devoted to reducing hallucination in open-domain dialogue systems incorporating external knowledge. Neural Path Hunter (NPH) (Dziri et al., 2021) leverages a hallucination critic and retrieves faithful entities by a query signal propagated over a sub-graph in the refinement stage. Shuster et al. (2021) explore various neural-retrieval-in-the-loop architectures where a retriever is introduced for knowledge selection. Rashkin et al. (2021) propose an faithfulness control code in decoding using re-sampling techniques. Wu et al. (2021) define a control mechanism with lexical control phrases and inductive attention where potentially uninformative attention links are removed. Our work improves the fusion and interaction between external knowledge and dialogue context via various knowledge groundings and reasoning techniques, further reducing hallucination.

2.2 KG Enhanced Dialogue Generation

KGs convey large amounts of structured knowledge, which can help to improve dialogue systems' performance in informativeness (Tuan et al., 2019) and empathy (Li et al., 2020). For open-domain dialogue generation, Liu et al. (2019) unify knowledge triples and texts as a graph, and conduct multi-hop reasoning for explainability. Xu et al. (2020a) propose a proactive dialogue generation method based on agnostic meta-learning considering the limited number of KGs. Kumar et al. (2020) learn unified representations by training syntactic graph convolution networks, knowledge, and

[2] **RHO** is short for **R**educing **H**allucination in **O**pen-domain dialogue systems.

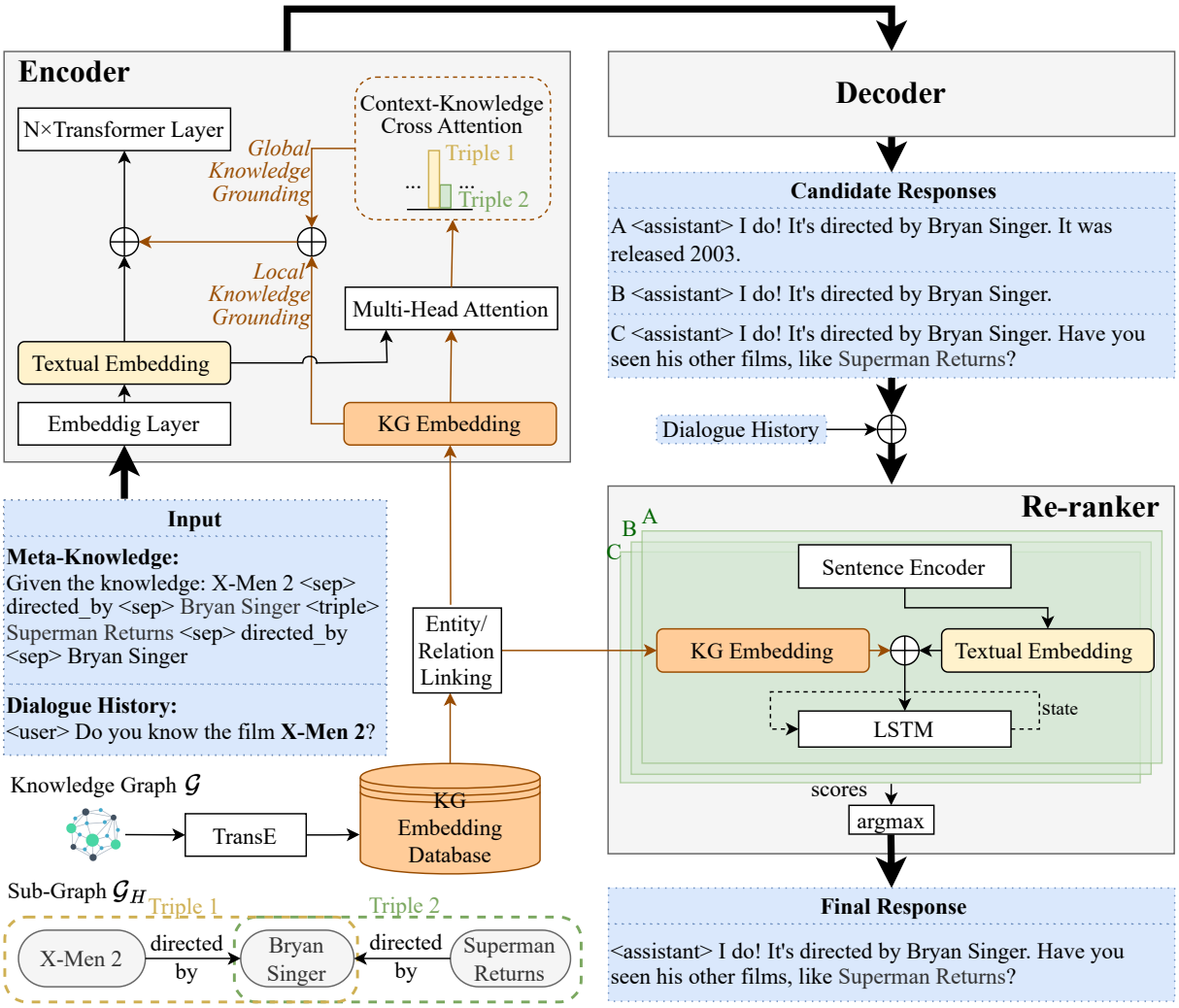

Figure 2: The overview of the proposed **RHO** framework. The input follows a knowledgeable task guidance template, including meta-knowledge and dialogue history (§3.2). To facilitate knowledge grounding, we first employ TransE to gain KG embeddings. For local knowledge grounding, we adapt entity/relation linking to recognize mentions in the dialogue context. The corresponding KG embeddings are locally fused into textual embeddings (§3.3). For global knowledge grounding, we aggregate the entire knowledge sub-graph in a memory bank via the attention mechanism so that the textual embeddings receive all context-related knowledge with emphasis (§3.4). After that, the encoder-decoder model generates several candidate responses. During post-processing, a re-ranker trained by traversal over a knowledge sub-graph conditioned on the dialogue context selects the most faithful response as the final output (§3.5).

memory module with triplet loss. Xu et al. (2022); Zhou et al. (2018); Zhang et al. (2020) explore and demonstrate how commonsense KG facilitates language generation in dialogue systems. Besides, Yang et al. (2020); Rony et al. (2022); Chaudhuri et al. (2021) are committed to incorporating KG into task-oriented dialogue models. Different from the above literature, our work employs the factoid knowledge paths from KG to improve the *faithfulness* of open-domain dialogue systems.

3 Methodology

In this section, we begin with a brief introduction to our KGD task. Then, the detailed techniques of **RHO** are presented. Please refer to Figure 2 for an overview of our approach. **RHO** incorporates both textual and structured information from external KG into dialogue system via knowledgeable task guidance (§ 3.2) and enhanced knowledge groundings (§ 3.3, § 3.4). Specifically, we introduce local token-level knowledge grounding in § 3.3 and global grounding to provide a comprehensive view and the multi-hop reasoning ability in § 3.4. To effectively encode the heterogeneous sources, we

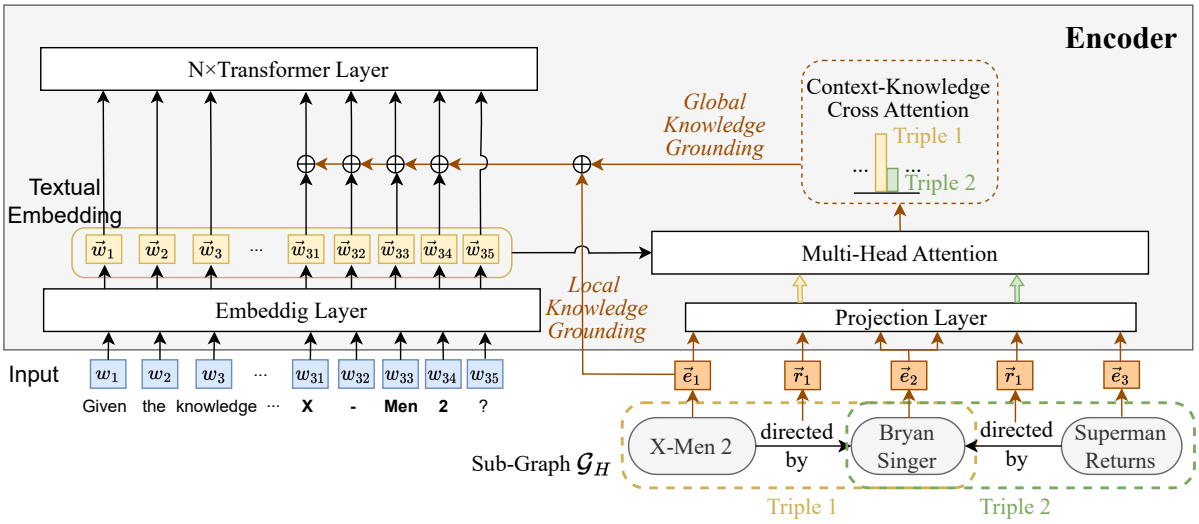

Figure 3: The diagram of the encoder with proposed local and global knowledge groundings.

sum textual embedding, local, and global grounded embedding. Figure 3 is a more detailed diagram of the encoder with proposed knowledge groundings. In addition, during the post-processing stage, we incorporate a re-ranking technique that rewards generation candidates with low hallucination levels; this technique is proposed based on the hypothesis that the faithful response can be reasoned backward to the source (§ 3.5). We implement our model based on BART (Lewis et al., 2020) architecture, and please refer to Appendix A for details.

3.1 Task Definition

In the response generation task in dialogue systems, each data sample consists of a dialogue history H, including a set of utterances U conducted by humans and agents interactively. The goal of the response generation model in the dialogue system is to learn to generate a proper response R based on the dialogue history H.

The response generation for KGD task is a special case of the above task that takes a multi-relational KG as an additional input; Multi-relational KG is a directed graph $\mathcal{G}$ formulated by a collection of triples, denoted as $T = \langle$`[SBJ], [PRE], [OBJ]`$\rangle$, as additional input. Here, `[SBJ], [OBJ]` are subject and object entities, and `[PRE]` is the relation predicate r indicating the relationship between the subject and object entities e. The goal of KGD is to generate a faithful response R based on the history H and a knowledge sub-graph $\mathcal{G}_H$, which is the subset of the entire KG $\mathcal{G}$ with triples semantically related to the dialogue history H. Our task is also in line with the previous works (Zhou et al., 2021; Dziri et al., 2021). Figure 2 has an illustrated example.

3.2 Input with Knowledgeable Task Guidance

A naive approach of input construction from raw data samples for language models is simply concatenating all triples in $\mathcal{G}_\mathcal{H}$ with the dialogue history H[3]. However, there is a lack of guidance to specifically excavate the model's innate ability to handle the KGD task (Brown et al., 2020). Inspired by Raffel et al. (2020), we design a prompt to guide the PLM for KGD and convert the structured $\mathcal{G}_\mathcal{H}$ into textual information. Here, we linearize the triples in $\mathcal{G}_\mathcal{H}$ into texts (treated as meta-knowledge) and cooperate the dialogue history utterances U with the following template: "`Given the knowledge:` `[SBJ]`$_1$ `<sep> [PRE]`$_1$ `<sep> [OBJ]`$_1$ `<triple> [SBJ]`$_2$ `<sep> [PRE]`$_2$ `<sep> [OBJ]`$_2$ `<triple>` $\cdots$ `<user>` U_1 `<assistant>` U_2 $\cdots$", where `<sep>`, `<triple>`, `<user>`, and `<assistant>` are special markers.

3.3 Local Knowledge Grounding

Although $\mathcal{G}_\mathcal{H}$ is converted and injected as additional input (§ 3.2), the model using textual information only cannot effectively handle the semantics of KGs which are typically sparse and complex in form (Petroni et al., 2019; Logan et al., 2019). Therefore, we ground the language representations with KG to take full advantage of lexical and structured knowledge information simultaneously.

During pre-processing, we obtain the collections of linked mentions of entities ($\mathcal{E}_H$) and relations

[3]Similar to the baseline approach in Dziri et al. (2021).

($\mathcal{R}_H$) from the dialogue history H and their KG embeddings as follows:

1. Identify entity mention e_m that appears in dialogue history H that can be linked to an entity e in the sub-graph $\mathcal{G}_\mathcal{H}$. We utilise an open-source linking tool named FLAIR (Akbik et al., 2019).
2. Since relations connecting entities are crucial in knowledge reasoning for PLMs (Labutov et al., 2018; Feng et al., 2020), we also link the relation mention r_m in H to the relation predicate r in $\mathcal{G}_H$.
3. We employ TransE (Bordes et al., 2013) to learn the KG embeddings of entities ($\vec{e}_\mathcal{G}$) and relation predicates ($\vec{r}_\mathcal{G}$) from the entire $\mathcal{G}$ [4].

Then, we obtain a *locally grounded* token embedding $\vec{w}_{local}$ for an arbitrary non-special token w in H as follows:

$$\vec{w}_{local} = \begin{cases} M(\vec{e}_\mathcal{G}) & \texttt{substr}(w, \mathcal{E}_H) \\ M(\vec{r}_\mathcal{G}) & \texttt{substr}(w, \mathcal{R}_H) \\ \vec{0} & \text{otherwise} \end{cases} \quad (1)$$

where $M(\cdot)$ transforms the space from the KG embeddings to the PLM token embeddings. A typical way of implementing $M(\cdot)$ is through a mapping matrix. Specifically, if $dim(\vec{w}) = dim(\vec{e}_\mathcal{G})$, $M(\cdot)$ can be further simplified as an identity mapping. $\texttt{substr}(w, \mathcal{E}_H)$ is a Boolean indicator that returns true if the current token w is a sub-string of any e_m in $\mathcal{E}_H$. This way, tokens related to specific entities or relation predicates can be grounded by their respective KG embeddings by fusing $\vec{w}_{local}$ into the vanilla token embedding $\vec{w}$. We regard this approach to be *local* as $\vec{w}_{local}$ is only related to the KG embedding of the corresponding node.

As in Figure 2 and Figure 3, the tokens `"X", "-", "Men", "2"` in dialogue history are linked to the entity `"X-Men 2"` in KG. Then, we take the corresponding KG embedding from the database gained by TransE as the local knowledge grounding which will fuse into these tokens' textual embeddings.

3.4 Global Knowledge Grounding

Focusing only on a single token in the context and a single node in the graph is insufficient to enhance the multi-hop reasoning abilities of the dialogue system. In addition to local grounding, we further propose global knowledge grounding which enriches the semantics considering the entire sub-graph $\mathcal{G}_H$ and hence offers the model a comprehensive view of the background knowledge.

Following our observation, we adopt the attention mechanism (Vaswani et al., 2017) to draw global dependencies between the dialogue history H and a memory bank storing the representations of all knowledge triples in $\mathcal{G}_H$. Let T_H be the collection of all triples in $\mathcal{G}_H$. The memory bank stores $|T_H|$ embedding vectors where the i-th vector $\vec{v}_i$ corresponds to the KG embedding of the i-th relation triple $T_i = \langle \texttt{[SBJ], [PRE], [OBJ]} \rangle$:

$$\vec{v}_i = M([SBJ]) \oplus M([PRE]) \oplus M([OBJ]) \quad (2)$$

where $\oplus$ is the concatenation operator for vectors. We gather all vectors and further project them to a global knowledge embedding space by:

$$K_H = W_{proj} \cdot [\vec{v}_1 \oplus \cdots \oplus \vec{v}_{|T_H|}] \quad (3)$$

where W_{proj} is a learnable projection matrix.

Based on the formulation of K_H, we compute how much attention the current token w in H pays to each relation triple according to the semantic relevance and obtain the *globally grounded* token embedding $\vec{w}_{global}$ as follows:

$$\vec{w}_{global} = \begin{cases} \text{softmax}(\frac{\vec{w} \cdot K_H^T}{\sqrt{dim(\vec{w})}}) \cdot K_H & \texttt{substr}(w, \mathcal{E}_H), \\ & \texttt{substr}(w, \mathcal{R}_H) \\ \vec{0} & otherwise \end{cases} \quad (4)$$

As in Figure 2 and Figure 3, for `"X-Men 2"` in the dialogue history, Triple 1 should have more influence on the tokens' representation than Triple 2 due to its higher relevance.

Finally, the encoder of **RHO** sums the vanilla token embedding $\vec{w}$, the locally grounded embedding $\vec{w}_{local}$, and the globally grounded embedding $\vec{w}_{global}$ as: $\tilde{w} = \vec{w} + \vec{w}_{local} + \vec{w}_{global}$. During training, while $\vec{w}$ is rapidly updated via back propagation, $\vec{w}_{local}$ and $\vec{w}_{global}$ are relatively fixed with few parameters trainable (e.g., W_{proj}).

3.5 Response Re-ranking

With the above approaches, our knowledge-grounded model generates N candidate responses by beam search. Yet, the grounding process mainly applies to the embedding level, lacking output constraints. To enhance our **RHO**'s ability to reduce hallucination, we extend KG-CRUSE (Sarkar et al., 2022) and train a conversational reasoning model ϕ

[4] We employ TransE via OpenKE (Han et al., 2018) as our underlying KG representation learning algorithm due to its effectiveness and simplicity. We have also experimented with some recent algorithms and observed no improvement in the performance (refer to Appendix C for details).

for response re-ranking, with emphasis on the KG. If the generated response can be reasoned backward to the source, we can assume it is faithful.

In our approach, we obtain the semantic embeddings of the dialogue history H and a possible response R via a contextual sentence encoder, i.e., Sentence-BERT (Reimers and Gurevych, 2019). The model ϕ is an LSTM-based decoder that learns the probability $p_{t,\phi}$ of an action $\vec{a}_t$ given the state $\vec{s}_t$ at step t. Here, the action refers to a walking step on the graph $\mathcal{G}_H$, represented as $\vec{a}_t$, which is the concatenation of the relation and entity embeddings derived from the KG, together with their semantic embeddings based on Sentence-BERT, i.e., [5]

$$\vec{a}_t = (\vec{e}_{\mathcal{G}} + \vec{e}_{\mathcal{S}}) \oplus (\vec{r}_{\mathcal{G}} + \vec{r}_{\mathcal{S}}) \quad (5)$$

where $\vec{e}_{\mathcal{S}}$ and $\vec{r}_{\mathcal{S}}$ are the semantic sentence embeddings of an entity e and a relation predicate r, respectively. The state $\vec{s}_t$ contains the representations of the dialogue history, together with entities and relations already traversed by ϕ (action history). It is defined as a tuple $(H, (\vec{a}_1, \vec{a}_2, \cdots, \vec{a}_{t\text{-}1}))$. Hence, the model ϕ explicitly models the process of a traversal upon $\mathcal{G}_H$ conditioned on the dialogue history H and a possible response R. During training, each action $\vec{a}_t$ made by ϕ is combined into a path, and the target path is the given context-related subgraph $\mathcal{G}_H$.

After our encoder-decoder model generates N candidate responses $\{R_1, \cdots, R_N\}$, we select the best response R^* with the highest probability $\mathbf{p}_\phi = \prod_t p_{t,\phi}$ over all the generated responses, i.e.,

$$R^* = \underset{n \in \{1,\cdots,N\}}{\arg\max}\ \mathbf{p}_\phi(\mathbf{A} = \mathcal{G}_{\mathcal{H}} | H, R_n) \quad (6)$$

where $\mathbf{A}$ is a collection of actions $\vec{a}_t$ (i.e. knowledge path) that ϕ has already traversed conditioned on the dialogue history H and each response R_n.

For a more intuitive understanding, refer to the example in Figure 2 where the model generates three candidate responses: A, B, and C. It selects Response C as the final output with the traversal path "`X-Men 2, directed_by, Bryan Singer, ~directed_by, Superman Returns`". [6] As seen, there is a higher matching degree between the sub-graph in Figure 2 and the Response C, compared to other candidate responses (i.e., A and B).

[5] We have also investigated the impact of the KG embeddings ($\vec{e}_{\mathcal{G}}$ and $\vec{r}_{\mathcal{G}}$) for action modeling in Appendix E.1 by comparing the performance of the re-ranker under two settings: i) $\vec{a}_t = (\vec{e}_{\mathcal{G}} + \vec{e}_{\mathcal{S}}) \oplus (\vec{r}_{\mathcal{G}} + \vec{r}_{\mathcal{S}})$ and ii) $\vec{a}_t = \vec{e}_{\mathcal{S}} \oplus \vec{r}_{\mathcal{S}}$ (the vanilla model in KG-CRUSE).

[6] ~directed_by refers to the opposite direction of the relation directed_by.

4 Experiments

4.1 Dataset

OpenDialKG (Moon et al., 2019) contains open-ended dialogues between two speakers, initiated by talking about a given entity and grounded on the relevant facts from a structured KG. Thus, the sequential turn-based dialogues can be regarded as traversing the paths in the KG. To our knowledge, OpenDialKG is currently the only publicly available corpus for English open-ended dialogues with KG path annotations (Yu et al., 2022; Ni et al., 2022), and previous works (Dziri et al., 2021; Zhou et al., 2021) evaluate their effectiveness on this corpus. Hence, we also conduct our experiments on OpenDialKG. Consistent with previous works (Dziri et al., 2021; Liu et al., 2019; Zhou et al., 2021), we filter OpenDialKG by keeping only the dialogue samples that are annotated with a KG path. The dataset is divided into training, validation, and testing sets in the ratio of 8:1:1.

4.2 Baselines

The following strong baselines are employed to show the efficiency of our method. We fine-tune pre-trained language models **GPT2** (Radford et al., 2019) and **BART** (Lewis et al., 2020) on our task. NPH (Dziri et al., 2021) refines the generated responses by retrieved entities from the KG. To our knowledge, the integration of GPT2 and NPH, called **GPT2+NPH**, reaches the SOTA performance on OpenDialKG. Since this post-processing technique is agnostic to the generation model, we apply it to BART, named **BART+NPH** as our baseline. In addition, **EARL** (Zhou et al., 2021) utilizes external KGs for conversation generation without parameterizing specific entity representations. **KG-BART** (Liu et al., 2021), a KG-augmented pre-trained language generation model based on BART, introduces the information of the relations among concepts for generative commonsense reasoning. We are the first to adapt this model to the KGD generation. Furthermore, we explore **ChatGPT** on this task in Appendix A. Please refer to it for the details of baseline implementations.

4.3 Evaluation Protocols

4.3.1 Automatic Evaluation

To evaluate the generation quality, we use the classical word overlap-based NLG metrics: **BLEU** and

Model	BLEU4 ↑	ROUGE-L ↑	FeQA ↑	QuestEval ↑		Entity Coverage (%) ↑		
				RD	RF	Pre.	Recall	F1
EARL (Zhou et al., 2021)	7.97	23.61	39.93	37.88	35.59	86.61	45.17	64.44
GPT2 (Radford et al., 2019)	10.27	29.59	39.60/26.54†	46.86	42.07	91.62	33.26	52.30
GPT2+NPH (Dziri et al., 2021)	10.41	29.93	40.83/28.98†	47.45	42.45	95.61	33.39	53.96
BART (Lewis et al., 2020)	14.45	33.33	39.00	46.97	42.75	96.99	44.96	62.87
BART+NPH	15.53	34.99	42.41	47.94	43.56	96.44	44.12	65.98
KG-BART (Liu et al., 2021)	13.72	33.31	41.87	45.55	42.86	97.68	45.63	64.58
RHO (LKG)	19.89	**39.95**	43.04	48.91	44.37	97.38	45.57	67.77
RHO (GKG)	**20.77**	39.54	40.65	48.41	43.84	97.20	45.63	67.40
RHO (LKG+GKG)	20.63	39.51	45.96	50.35	46.03	98.26	50.74	71.47
RHO (Full Implementation)	19.11	38.45	**47.99**	**50.58**	**46.41**	**98.53**	**51.77**	**72.29**

Table 1: Automatic evaluation results for **RHO** and baselines, where "RD", "RF", and "Pre." refer to reference-dependent, reference-free mode, and Precision, respectively. The results of the ablation study are shown in the last four rows. "LKG", "GKG" and "RR" refers to local knowledge grounding, global knowledge grounding and response re-ranking, respectively. "Full Implementation" means that we implement all three components, i.e., LKG+GKG+RR. † The FeQA scores we calculate (former) are higher than those reported in Dziri et al. (2021) (latter).

ROUGE-L (Lin, 2004). Due to the possible presence of hallucinations in the dataset (especially extrinsic ones) the metrics based on the n-gram overlap between the golden answer and generated texts are not sufficient (Ji et al., 2022a). Therefore, we also use source-dependent metrics, i.e., **FeQA** (Durmus et al., 2020), **QuestEval** (Scialom et al., 2021), and **Entity Coverage**, to estimate the hallucination degree. **FeQA** and **QuestEval** are both question-answering (QA)-based metrics for evaluating the faithfulness of the output in generation tasks [7]. **QuestEval** has two modes: reference-dependent mode assesses a text with one or multiple ground-truth references; and reference-free mode assesses a text without any references. In addition to the metrics used in previous works (Dziri et al., 2021; Zhou et al., 2021), we assume that entities in generated responses should be covered by those in the given knowledge triples and dialogue history. The higher **Entity Coverage** is, the lower hallucination degree can be to some extent. Specifically, we utilize a named entity recognition (NER) model to extract named entities in generated responses and the dialogue history. We compute **Entity Precision, Recall and F1** scores between generated entities and entities in KG and dialogue history to evaluate the faithfulness of generated responses.

4.3.2 Human Evaluation

To further assess the quality of generated responses from different systems, we conduct human evaluations using Amazon Mechanical Turk[8]. For hallucination level assessment, we first ask annotators to identify whether each response is **Faithful**, or **Hallucinated** given the dialogue history and KG triples. The judgment criteria are as described in § 1. "Faithful" means that the response is supported by the knowledge triples and dialogue context, while "hallucinated" means that the response contradicts or cannot be verified by the source. If the response is hallucinated, we further ask annotators to identify whether the hallucination is **Extrinsic**, **Intrinsic** or **Both** (Dziri et al., 2021).

We also conduct an A/B test of our framework against the baselines GPT2+NPH and BART+NPH to evaluate generated responses on **Fluency** (Ji et al., 2022b; Dathathri et al., 2020). The annotators are asked whether the writing is smooth and grammatically correct and given four choices: **Neither**, **Both**, **Sample A**, or **Sample B**. Please refer to Appendix B.2 for details.

5 Results and Analysis

5.1 Overall Evaluation Results

Automatic Evaluation. The first eight rows of Table 1 shows the experimental results on automatic metrics over the OpenDialKG test set. Our model outperforms all baselines on both classic overlap metrics and hallucination metrics, indicating the high quality of the generated utterances. Specifically, compared to SOTA (GPT2+NPH), **RHO** gives a significant rise of 17.54% in FeQA, 9.33% in QuestEval (RF), and 33.97% in Entity Coverage (F1). **RHO** also achieves better results compared to the stronger baseline BART+NPH. The results indicate the faithfulness of KGD systems can be improved by knowledge grounding and re-ranking

[7] Please refer to Appendix B.1 for details.
[8] https://www.mturk.com/

Model	Faith. (%) ↑	Hallucination (%) ↓		
		In.	Ex.	Both
GPT2+NPH	72.67	8.67	18.00	0.67
BART+NPH	75.00	9.33	15.33	**0.33**
RHO w/o RR	80.67	**7.67**	10.67	1.00
RHO	**81.67**	**7.67**	**10.00**	0.67

Table 2: Human evaluation results for hallucination degree, where "Faith.", "In.", and "Ex." refers to faithfulness, intrinsic, and extrinsic hallucination, respectively.

Model \ Fluency	Win (%)↑	Lose (%)↓	Tie (%)
RHO w/o RR vs. GPT2+NPH	37.33	20.67	42.00
RHO w/o RR vs. BART+NPH	24.67	18.67	56.67
RHO vs. GPT2+NPH	32.33	16.00	51.67
RHO vs. BART+NPH	17.00	12.67	70.33

Table 3: Human evaluation results for fluency.

techniques.

Human Evaluation. As in Table 2, the faithfulness of **RHO** is higher than GPT2+NPH and BART+NPH. The results are statistically significant with p-value$<$0.05. Specifically, we see a 12.38% increase in faithfulness compared to SOTA (GPT2+NPH). As shown in Table 4, the information *"Judy Davis starred in My Brilliant Career"* in golden answer, is not supported by the input, although it is factual according to the world knowledge. The baseline model hallucinates unfactual information, i.e., *"Judy Davis starred in both The Referendum and The Golden Compass"*, while our model is better grounded on the input source.

At a more granular level, the extrinsic hallucination problem is more frequent than the intrinsic one in all models. This phenomenon is also observed in other works (Dziri et al., 2021; Nan et al., 2021). Specifically, compared to SOTA (GPT2+NPH), **RHO** reduces extrinsic hallucination by 42.85%. Compared to BART+NPH, **RHO** reduces intrinsic hallucination by 13.66% and extrinsic hallucination by 46.74%. According to the A/B test results for fluency in Table 3, **RHO** is slightly more fluent than SOTA methods. Overall, human evaluation results are in line with automatic evaluation. **RHO** mitigates both intrinsic and extrinsic hallucination issues without sacrificing fluency.

5.2 Ablation Study

We conduct an ablation analysis to assess the contribution of each component of our method: Local Knowledge Grounding (LKG), Global Knowledge Grounding (GKG), and Response Re-ranking (RR). As shown in the last four rows of Table 1, fully-implemented **RHO** performs best in automatic hallucination metrics with a slight sacrifice of classical overlap metrics. Specifically, compared to models equipped with only local/global knowledge grounding, the model equipped with both two (LKG+GKG) gains higher scores in FeQA, QuestEval, and Entity Coverage. The same trend is observed when comparing the fully-implemented model with the model without re-ranking (LKG+GKG) and please refer to Appendix E.2 for an example. The trade-off between hallucination metrics and the others is due to the fact that some reference responses in the dataset diverge from the sources (Dziri et al., 2022). Hence, improving responses' consistency with the source will inevitably reduce that with the references. Overall, the above three mechanisms synergy to improve the generated responses' faithfulness without significant reduction in quality.

5.3 Error Analysis

Although **RHO** achieves better performance than the baselines in the KDG task, it still generates a few cases of failure. To gain more insight into our model, we present failure examples from **RHO** in Appendix D and conduct an error analysis.

As shown in Table 2, the extrinsic hallucination problem is more frequent. Because the corpus itself has more extrinsic hallucinations and the model is more likely to produce unverified rather than contradictory information with the limitation of the source. To solve the problem, we suggest that the quality of the corpus used for model training, such as OpenDialKG, can be further improved by filtering out irrelevant and contradictory information. Fact-checking can also be potentially used to find evidence from world knowledge and verify the generated responses (Ji et al., 2022a).

In addition to intrinsic and extrinsic hallucination issues, we discover two other types of negative cases, namely **No Full Coverage**, and **Unnatural Connection**. **No Full Coverage** refers to the situation where the output does not cover the full answer as expressed by all triples mentioned in the given sub-graph. We believe that further research on the reasoning ability and interpretability of the model can help address this issue. **Unnatural Connection** denotes that a response is not connected to the dialogue history especially the last utterance naturally. We find that in a few cases, the dialogue system delivers an irrelevant answer or poorly replies to the dialogue history. This issue occurs because

KG	History	Golden	GPT2+NPH	RHO
<The Ref, starred actors, Judy Davis>	User: Who else was in that movie? Assistant: Aside from Woody Allen, Judy Davis and Greta Gerwig also have star credits. User: What else do I know Judy Davis from?	Assistant: She also starred in "The Ref" and "**My Brilliant Career**."	Assistant: She starred in **The Referendum** and **The Golden Compass**.	Assistant: Judy Davis also starred in The Ref. Have you seen this movie?

Table 4: An example response generated by GPT2+NPH and our model. "`My Brilliant Career`" in the golden answer, "`The Referendum`" and "`The Golden Compass`" from NPH are extrinsic hallucinations since there is no clue in external knowledge and dialogue history.

our model focuses more on hallucination reduction and sometimes pays less attention to flexibility and diversity. In the future, we can explore the controllability of the grounding degree on knowledge and strike a dynamic balance between faithfulness and flexibility in response generation.

6 Conclusion and Future Work

In this work, we investigate the hallucination in KGD systems and propose the **RHO** framework to tackle this issue with three mechanisms: Local Knowledge Grounding, Global Knowledge Grounding, and Response Re-ranking. Our method is empirically proven effective in reducing hallucinations with automatic and human evaluation. We also conduct deep error analysis on the generated responses. For future work, the re-ranking module can be combined with foundation models such as ChatGPT to reduce manual labor in the human feedback process. We also call for research to strike a better balance between response faithfulness and diversity.

Limitations

The deep neural networks in **RHO** uses feature extraction and vectorization to represent the texts. The model only detects the statistical regularities and quantitative relationships among the variables but can not see qualitative relationships, such as causality, hierarchy, and other abstractions (Tsimenidis, 2020). Although we leverage the response re-ranking technique, which improves the explainability of **RHO**, the neural networks are undoubtedly still "black boxes" to humans. Therefore, the faithfulness of generated responses can not be fully guaranteed.

Ethical Considerations

In addition to the hallucination problem, another critical challenge, the offensive language, is also introduced with the evolutionary progress toward building reliable dialogue systems. The data-driven models are susceptible to delivering offensive responses while mimicking human conversations (Xu et al., 2020b). It has been shown that racial and gender biases are encoded in the PLMs (Blodgett et al., 2020), and these biases are present in the training corpus. Since **RHO** leverages PLMs and the training corpus, it is possible to generate offensive languages. We suggest that in real-world dialogue systems, it is necessary to employ some post-processing steps to alleviate this problem when it is deployed online.

Acknowledgement

This work has been supported by the China NSFC Project (No. NSFC21EG14), SAAIR Project (No. Z1286), and HKJCCT21EG01 (RG192).

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, and Jens Lehmann. 2021. Grounding dialogue systems via knowledge graph aware decoding with pre-trained transformers. In *European Semantic Web Conference*, pages 323–339. Springer.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. *ICLR*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Xu Han, Shulin Cao, Lv Xin, Yankai Lin, Zhiyuan Liu, Maosong Sun, and Juanzi Li. 2018. Openke: An open toolkit for knowledge embedding. In *Proceedings of EMNLP*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022a. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Ziwei Ji, Yan Xu, I-Tsun Cheng, Samuel Cahyawijaya, Rita Frieske, Etsuko Ishii, Min Zeng, Andrea Madotto, and Pascale Fung. 2022b. Vscript: Controllable script generation with visual presentation. *AACL Demo*.

Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. *Advances in neural information processing systems*, 31.

Gaurav Kumar, Rishabh Joshi, Jaspreet Singh, and Promod Yenigalla. 2020. AMUSED: A multi-stream vector representation method for use in natural dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 750–758, Marseille, France. European Language Resources Association.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-relational question answering from narratives: Machine reading and reasoning in simulated worlds. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 833–844, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL 2020*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2020. Knowledge bridging for empathetic dialogue generation. *AAAI*.

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, Seattle, United States. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Ye Liu, Yao Wan, Lifang He, Hao Peng, and S Yu Philip. 2021. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6418–6425.

Zhibin Liu, Zheng-Yu Niu, Hua Wu, and Haifeng Wang. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical*

Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 1782–1792.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy. Association for Computational Linguistics.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

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

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. *NNACL*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2022. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, pages 1–101.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. 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, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Md Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure aware task-oriented dialogue generation. *NAACL*.

Rajdeep Sarkar, Mihael Arcan, and John Philip McCrae. 2022. Kg-cruse: Recurrent walks over knowledge graph for explainable conversation reasoning using semantic embeddings. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 98–107.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

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.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *ICLR*.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*, pages 2071–2080. PMLR.

Stefanos Tsimenidis. 2020. Limitations of deep neural networks: a discussion of g. marcus' critical appraisal of deep learning. *arXiv preprint arXiv:2012.15754*.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14085–14093.

Feifei Xu, Shanlin Zhou, Yunpu Ma, Xinpeng Wang, Wenkai Zhang, and Zhisong Li. 2022. Open-domain dialogue generation grounded with dynamic multiform knowledge fusion. In *Database Systems for Advanced Applications*, pages 101–116, Cham. Springer International Publishing.

Hongcai Xu, Junpeng Bao, and Junqing Wang. 2020a. Knowledge-graph based proactive dialogue generation with improved meta-learning. In *2020 2nd International Conference on Image Processing and Machine Vision*, pages 40–46.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020b. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys (CSUR)*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. 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, Minlie Huang, Yong Liu, Wei Chen, and Xiaoyan Zhu. 2021. Earl: Informative knowledge-grounded conversation generation with entity-agnostic representation learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2395.

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 Implementation Details

RHO The maximum dialogue history length is set to 3 utterances. This setting is also held constant in baselines. Our method is implemented using the Huggingface Transformers library[9]. We load the pre-trained BART-base model and train **RHO** with the following settings and hyper-parameters: the batch size 16, the learning rate 3e-5, and the AdamW optimizer with a linear scheduler. We generate multiple candidate responses using beam search (with the number of beams B=4). Our model is trained on one NVIDIA Geforce RTX 3090 GPU. It takes approximately 3 hours to train.

BART We fine-tune a BART-base model with the following settings and hyper-parameters: the batch size 16, the learning rate 3e-5, and the AdamW optimizer with a linear scheduler. We also generate responses using beam search (B=4).

Modified KG-BART We modify the code from the official library[10] to fit our KGD task. We load the pre-trained BART-base model and train the modified KG-BART with the default hyper-parameters: the batch size 16, the learning rate 1e-5, and the AdamW optimizer with a linear scheduler. We also generate responses using beam search (B=4).

GPT2 We fine-tune GPT2-small with the following settings and hyperparameters: the batch size 16, the learning rate 6.25e-5, and the AdamW optimizer with a linear decay scheduler. We also generate responses using beam search (B=4). More details of hyper-parameters can be found in Dziri et al. (2021) where GPT2 is regarded as a strong baseline for the same task.

NPH We implement NPH using the code from the official library[11] with their default hyper-parameters. We also utilize the provided graph embeddings and the hallucination critic model.

EARL We obtain the best-generated responses from the authors of Zhou et al. (2021) and evaluate the quality of the responses via our metrics.

ChatGPT We randomly select 50 samples from the test set of OpenDialKG. Prompt engineering is needed when applying ChatGPT to our task. At first, ChatGPT refuses to generate a response because it is too cautious to express opinions or feelings. For example, we input "`Given the knowledge: We Rode in Trucks, Composer, Luke Bryan\n \n User: I like Luke Bryan's music. What do you think about him?\n \n Please generate the next turn.`" The output is "`As an AI, I do not have personal opinions or feelings. However, I can provide information about Luke Bryan's music and career...`"

Therefore, we try different prompts, and the successful one we adopt is "`Can we try dialogue generation? I will give you turns, and you can generate the next turn, but only one.\n \n You can also consider the knowledge of "We Rode in Trucks, Composer, Luke Bryan" for your reference in the dialogue.\n \n User: I like Luke Bryan's music. What do you think about him?`" The output is "`Assistant: I think Luke Bryan is a talented musician and songwriter. His hit songs like` **`"Country Girl (Shake It for Me)"`** `and "We Rode in Trucks" showcase his ability to connect with audiences through his music.`"

After human observation and analysis, we find the following phenomena: The generated responses tend to be long, sometimes even generating multiple turns, even if we ask it to generate only one. The length distinction from the golden answer. In addition, ChatGPT refers to both the given knowledge and the parametrized background knowledge injected during pre-training. There are lots of extrinsic hallucinations that cannot be verified and supported by the input source. For example, "`Country Girl (Shake It for Me)`" in the previous paragraph. More exploration is needed on detecting and mitigating hallucination in ChatGPT on KGD.

B Evaluation

B.1 Automatic Metrics

FeQA FeQA, a QA-based metric for evaluating the faithfulness of the generated output, has been

[9] https://github.com/huggingface

[10] https://github.com/yeliu918/KG-BART

[11] https://github.com/nouhadziri/Neural-Path-Hunter

applied in summary (Chen et al., 2021) and dialogue (Dziri et al., 2021) tasks. As a reference-free metric, it takes the source (such as a document) and the corresponding output to be evaluated (such as a summary) as input. Given the source, a question generation (QG) model generates a question based on the source and then a QA model generates an answer A. The QA model generates another answer B based on the question and output to be evaluated. The average F1 score between answers A and B reflects the hallucination level of the output. Following the setting in (Dziri et al., 2021), we concatenate all the knowledge triples in $\mathcal{G}_{\mathcal{H}}$ with the dialogue history H as the source, and the generated response is the output. We calculate the FeQA score using the code and models from the official library[12] with their default hyper-parameters.

QuestEval QuestEval (Scialom et al., 2021) is also a QA-based metric that has reference-dependent and reference-free modes. The reference-dependent mode depends both on the input source and golden answers, while the reference-free mode does not require any ground-truth references. The input source is built in the same way as in FeQA. We calculate the QuestEval score using the code and models from the official library[13] with their default hyper-parameters.

B.2 Human Evaluation

We conduct the human evaluation to assess **RHO**'s performance in response generation, especially the ability to reduce hallucination. In detail, we randomly select 100 samples generated by each model. Each sample is then evaluated by three different annotators to rule out potential bias. We specify that annotators must meet the following qualifications: Their Human Intelligence Task (HIT) approval rates are greater than or equal to 95%, and the numbers of HITs approved are greater than or equal to 5000. The annotators are located in Australia, the United Kingdom, and the United States. Figure 4 and 5 are the user interfaces (UIs) on Amazon Mechanical Turk for human evaluation of Hallucination and Fluency, respectively. The instructions, questions, and examples for annotators are shown.

Model	MR ↓		Hits@10 (%) ↑	
	raw	filter	raw	filter
TransE	**950.31**	**499.34**	**59.66**	**71.05**
TransH	1776.69	946.60	51.71	67.76
RotaE	1733.45	973.26	55.46	66.50
SimplE	2255.05	1455.30	45.98	53.21
DistMult	2288.26	1492.64	45.15	51.75
ComplEx	2381.29	1557.51	47.31	57.68

Table 5: Link prediction results on OpenDialKG.

C KG Representation Learning

For the proposed knowledge grounding techniques in § 3.3 and § 3.4, we employ several KG representation learning algorithms including TransE (Bordes et al., 2013), TransH (Wang et al., 2014), RotaE (Sun et al., 2019), SimplE (Kazemi and Poole, 2018), DistMult (Yang et al., 2015), and ComplEx (Trouillon et al., 2016) via OpenKE [14]. The link prediction results are shown in Table 5, where MR is the mean rank and Hits@10 is the proportion of correct entities ranked in the top 10. "Filter" means removing all corrupted triplets in the dataset (Han et al., 2018). Due to the effectiveness and simplicity, we finally choose TransE to gain the KG embeddings of all entities and relation predicates in $\mathcal{G}$.

D Error Analysis

As discussed, we characterize the negative cases of KDG systems into four types: Extrinsic Hallucination, Intrinsic Hallucination, No Full Coverage, and Unnatural Connection. To gain more insights into our model, we present more failure examples from **RHO** in Table 6.

In the example of Extrinsic Hallucination, we find that in some cases, both the golden answers and our responses suffer from this issue. Based only on the given dialogue context, we cannot know or infer the genre of Windtalkers. The referenced answer is "`thriller`", while **RHO** guesses it is "`sci-fi`". They are both extrinsic hallucinations since the source cannot confirm them without other information. For Intrinsic Hallucination, the opinion in the reference answer is "`Todd Walker drags the team down`", while **RHO** generates "`Minnesota`

[12] https://github.com/esdurmus/feqa

[13] https://github.com/ThomasScialom/QuestEval

[14] https://github.com/thunlp/OpenKE

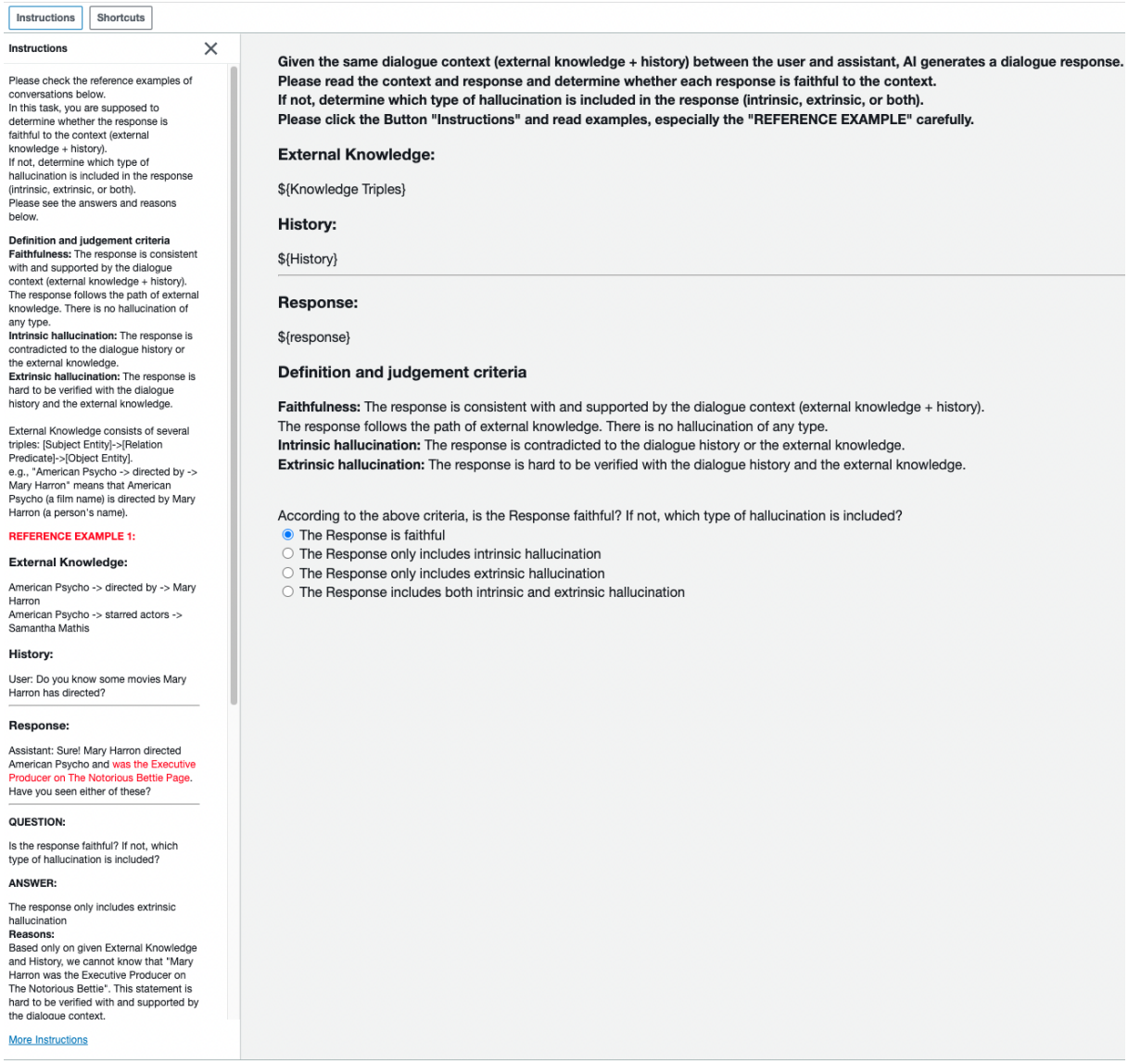

Figure 4: The UI for human evaluation on hallucination.

`Twins drafted Todd Walker`". However, according to the conversation, the relationship "`Drafted`" in the KG is wrong and should be "`Dragged`". In the example of No Full Coverage, the first triple "`<Batman, Romantically involved with, Catwoman>`" is not mentioned in the conversation. In the Unnatural Connection's example, our response grounds the response on the KG, but "`The Rise of Theodore Roosevelt has genre Autobiography`" is expressed mechanically and rigidly.

E Response Re-ranking

E.1 Performance of Re-rankers

For entities and relations in the KG, the original settings in Sarkar et al. (2022) only utilize the semantic sentence embeddings via a contextual sentence encoder, i.e., Sentence-BERT. In contrast, we sum their KG embeddings obtained by TransE with the above sentence embeddings. The evaluation results are shown in Table 8, where Hits@k is the proportion of the ground-truth path ranked in the top-k highest probability paths searched by the model. Since our settings achieve higher recall, we then use the model leveraging both sentence and KG embeddings as our re-ranker.

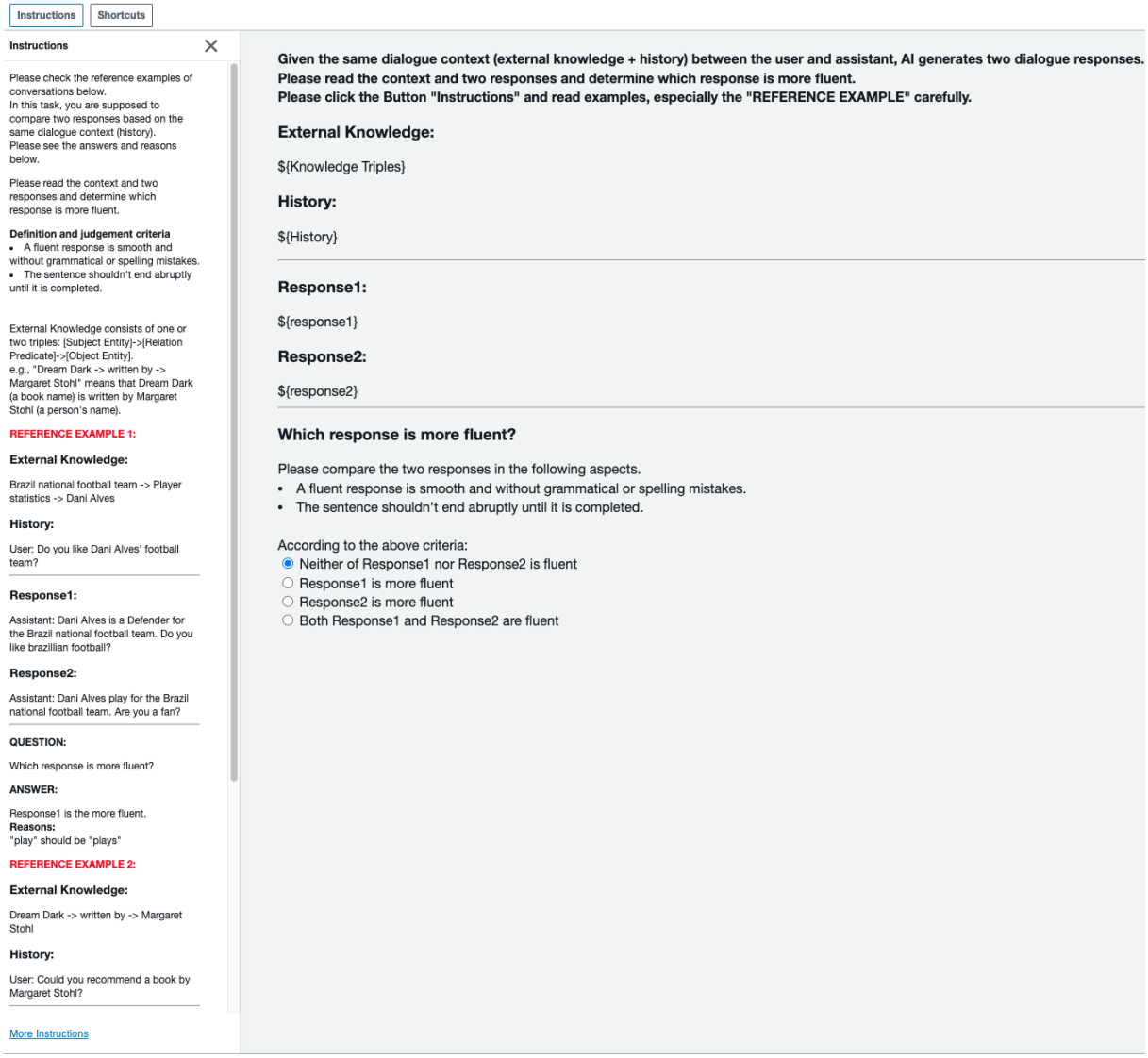

Figure 5: The UI for human evaluation on fluency via A/B testing.

E.2 Case Study of Re-ranking

As shown in Table 7, before re-ranking, there is a No Full Coverage issue mentioned in § 5.3. The response does not include the information from the given triple "<No Strings Attached, has genre, Romantic comedy>", which deviates from the established flow. After re-ranking, the response includes the missing information and becomes more similar to the golden answer.

Category	KG	History	Golden	RHO
Extrinsic Hallucination	<Windtalkers, starred actors, Nicolas Cage>	User: I like American Sniper, can you recommend something similar? Assistant: Sure. Windtalkers, Passchendaele, The Lying Game and The Mentalist are all in the same genre. User: Windtalkers is a Nicolas Cage flick, what genre is that exactly?	Assistant: Windtalkers is a **Thriller** and a similar thriller starring him is Knowing.	Assistant: Windtalkers is a **Sci-Fi** film starring Nicolas Cage.
Intrinsic Hallucination	<Todd Walker, **Drafted**, Minnesota Twins>	Assistant: Who's your favorite player in Utah Jazz? I only know of Kris Humphries. User: Kris Humphries is one of my favorites! I'm just glad he doesn't play for the Minnesota Twins, although he is from there.	Assistant: I'm not also fond of the Minnesota Twins. I think Todd Walker **drags** the team down.	Assistant: Did you know that Minnesota Twins **drafted** Todd Walker?
No Full Coverage	**<Batman, Romantically involved with, Catwoman>** <Batman, starred actors, Jack Palance>	Assistant: Have you seen the latest Batman movie. The Dark Knight Yet? User: Not yet. Who is in it? Assistant: Christopher Heyerdahl stars in The Dark Knight. Another good one is Catwoman. Have you seen it?	Assistant: Who is the star in Catwoman?	Assistant: I have not seen Catwoman. Who stars in it?
Unnatural Connection	<The Rise of Theodore Roosevelt, has genre, Autobiography>	Assistant: Have you read he Rise of Theodore Roosevelt or the Bell Jar? User: No, those both sound interesting though. Can you tell me more? Assistant: Yes, The Rise of Theodore Roosevelt won a Pulitzer Prize for Biography or Autobiography. Would you like more modern autobiographies, as well?	User: No, I think I will go with that, sounds like a winner.	User: The Rise of Theodore Roosevelt has genre Autobiography. Do you know of any other autobiographies?

Table 6: More cases of error analysis.

KG	History	Golden	w/o. Re-ranking	w/ Re-ranking
<No Strings Attached, starred actors, Natalie Portman> <No Strings Attached, has genre, Romantic comedy>	User: Do you like Natalie Portman?	Assistant: I do, Did you know she was in No Strings Attached is starring Ophelia Lovibond. It's a Romantic Comedy.	Assistant: Yes, she starred in No Strings Attached. Have you seen it?	Assistant: Yes, she starred in No Strings Attached, a romantic comedy. Do you like romantic comedies?

Table 7: Case study of Re-ranking.

Model	Hits@k (%) ↑			
	k=1	k=5	k=10	k=25
w/ Sent	19.86	44.43	56.42	70.38
w/ Sent+KG	**20.49**	**44.77**	**56.63**	**71.47**

Table 8: Performance of re-rankers on OpenDialKG. "Sent" denotes using sentence embeddings, while "Sent+KG" denotes using both sentence and KG embeddings.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3, Section 4.2, 4.3

☑ B1. Did you cite the creators of artifacts you used?
Section 3, Section 4, Appendix A

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We only use the publicly available pretrained models and datasets.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3, Section 4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We only use the publicly available corpus (OpenDialKG).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 5, Appendix D, E

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.3, Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, 4, Appendix A, B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 4.3.2, Appendix B.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B.2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 4.3.2

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B.2