Improving Factuality for Dialogue Response Generation via Graph-Based Knowledge Augmentation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) succeed in many natural language processing tasks. However, their tendency to hallucinate — generate plausible but inconsistent or factually incorrect text — can cause problems in certain tasks, including response generation in dialogue. To mitigate this issue, knowledge-augmented methods have shown promise in reducing hallucinations. Here, we introduce a novel framework designed to enhance the factuality of dialogue response generation, as well as an approach to evaluate dialogue factual accuracy. Our framework combines a knowledge triple retriever, a dialogue rewrite, and knowledge-enhanced response generation to produce more accurate and grounded dialogue responses. To further evaluate generated responses, we propose a revised fact score that addresses the limitations of existing fact-score methods in dialogue settings, providing a more reliable assessment of factual consistency. We evaluate our methods using different baselines on the OpendialKG and HybriDialogue datasets. Our methods significantly improve factuality compared to other graph knowledge-augmentation baselines, including the state-of-the-art G-retriever. The code will be released on GitHub.

1 Introduction

Large Language Models (LLMs) have been shown to perform powerfully on Natural Language Processing (NLP) tasks. Despite their general superiority, LLMs will generate some plausible but fact-inconsistent text, namely hallucination. Flawed pre-training data, model bias and randomness in inference are the factors contributing to hallucinations (Zhang et al., 2023; Huang et al., 2023). In dialogue response generation, generating incorrect responses will mislead people and have a negative impact on society.

A number of methods have been proposed to enhance the factuality of language models. Among them, knowledge-augmented inference methods aim to do this by adding knowledge triples into the prompt. This has been effective in tasks such as question-answering (Baek et al., 2023; Sen et al., 2023; Wu et al., 2023).

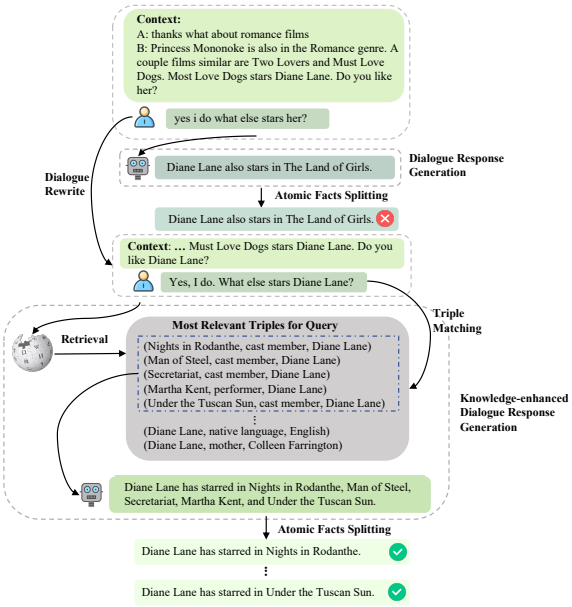

Figure 1: Examples comparing an LLM's dialogue response generation when utilising retrieved triples and a rewritten dialogue versus generating responses without external knowledge or dialogue rewriting. The knowledge triple retriever module assists in selecting the triples most relevant to the query.

However, these works (Baek et al., 2023; Wu et al., 2023) tend to use simple sentence embedders to encode knowledge and determine the match between a query and knowledge based on similarity, rather than employing a specialised model designed to assess the relevance between the query and knowledge. He et al. (2024) proposed a graph-based method called G-retriever, which employs the Prize-Collecting Steiner Tree to help select knowledge triples. However, applying these methods to dialogue response generation is challenging

due to their lack of consideration for dialogue context. Dialogues frequently involve intricate coreference structures, which complicate entity linking and hinder the LLMs' comprehension, ultimately compromising the factual consistency and quality of the generated responses.

To address above limitations, we propose a novel framework to improve dialogue factuality that consists of three key components: a knowledge triple retriever, a dialogue rewriting and knowledge-enhanced dialogue response generation. The knowledge triple retriever is fine-tuned on our well-collected samples, which helps select the valuable triples for the dialogue. The dialogue rewrite is based on Chain of Thought (CoT), which resolves the dialogue coreference. The knowledge-enhanced dialogue generation is proposed to improve the dialogue factuality, based on two kinds of approaches: prompt-based and graph-based. Figure 1 illustrates how the knowledge-augmented generation works. The query asks about what movies star "Diane Lane", and the initial response generated by the LLM contains factual inaccuracies. After applying a dialogue rewrite module that resolves coreference, a triple retriever helps select the most relevant N triples, including ("Nights in Rodanthe", "cast member", "Diane Lane"). With these corrected triples, the model generates an accurate response.

To evaluate factuality for responses, we primarily rely on the fact score (Min et al., 2023). However, the original design of the fact score does not consider dialogue context and situations where the knowledge source is unavailable, making it difficult to evaluate dialogue responses. To address this limitation, we adapt the fact score to the dialogue settings and assess model-human agreement. Annotation results show that our adapted fact score achieves substantial model-human agreement. Additionally, we propose Not Enough Information Proportion (NEIP) to evaluate dialogue factuality comprehensively. It is a metric that quantifies the proportion of atomic facts in a response that cannot be verified, such as opinions or hallucinated content that can not find any evidence to verify. We utilize the F1 score as a supplementary metric to assess factual accuracy and evaluate response quality through BLEU (Reiter, 2018), ROUGE-L (Lin, 2004), and Perplexity (PPL) (Jelinek et al., 1977). Additionally, we perform human evaluations to further assess response quality.

We compare our proposed method not only with standalone LLMs but also with knowledge-enhanced approaches such as KAPING, the BM25 algorithm, and the State-of-the-Art (SOTA) G-retriever (He et al., 2024). Experimental results show that our framework consistently outperforms both existing knowledge-enhanced methods and the current best-performing G-retriever.

Our contributions to this work can be listed as:

1. We propose a novel framework designed to enhance the factual accuracy of dialogue response generation. This framework incorporates three key components: a knowledge triple retriever that selects valuable triples, a dialogue rewrite that resolves coreference, and a knowledge-enhanced mechanism that effectively boosts overall factuality.

2. We adapt the fact score, originally designed for biography generation, to evaluate the factuality of dialogue systems, ensuring its validity within our task. The evaluation is fine-grained, enabling a more reliable assessment of factual consistency in dialogue systems.

3. We validate our methods against various baselines using two public dialogue datasets. Experimental results demonstrate that our approach significantly outperforms existing methods in factuality while keeping good quality at the generated response.

2 Related Work

2.1 LLM Hallucinations and Mitigation Methods

LLMs are trending in NLP, whose architecture can mainly be classified as encoder-only (Devlin, 2018), decoder-only (Brown, 2020; Dubey et al., 2024), and encoder-decoder (Chung et al., 2024). However, hallucination exists in these LLMs widely, which is the phenomenon that language models generate coherent but fact-inconsistent text. Many reasons contribute to hallucination. Huang et al. (2023) declared one potential reason is that the data in the pre-training of LLMs is incomplete, incorrect or outdated. Other reasons for hallucinations are the bias of LLMs and randomness in inference. Zhang et al. (2023) categorise hallucinations as input-conflicting, context-conflicting and fact-conflicting, with recent focus on the latter.

Several knowledge-based methods have been proposed to mitigate fact-conflicting hallucinations (Agrawal et al., 2023), which generally fall

into three categories: knowledge-aware inference, knowledge-aware training, and knowledge-aware validation. For instance, Baek et al. (2023) introduced KAPING, a prompt-based framework for question answering that retrieves knowledge triples based on embedding similarity. Building upon retrieval techniques, Sen et al. (2023) combined Knowledge Graph (KG) retrieval with language model reasoning to enhance performance on complex questions. Similarly, CRAG (Yan et al., 2024) employs an evaluator to assess generation quality and, if necessary, refines outputs via web search, showing significant improvements across various generation tasks. Another notable approach is Self-RAG (Asai et al., 2023), which retrieves relevant knowledge, generates an initial answer, and then refines it through self-critique, achieving strong results in multiple question answering benchmarks. In a related development, He et al. (2024) proposed G-Retriever, which integrates a graph-based encoder with the Prize-Collecting Steiner Tree algorithm to encode retrieved graph knowledge effectively, standing out among graph-enhanced knowledge retrieval methods.

2.2 Evaluation of Factuality

Dialogue response generation differs from question answering in that it must not only take the dialogue context into account when generating responses, but also must be evaluated based on the factuality and quality of the responses, rather than just the accuracy of answers. Previous dialogue factuality evaluation mainly focuses on humans (Ni et al., 2023; Li et al., 2022; Yu et al., 2022), and it is inefficient.

Automatic evaluating factuality is challenging, however; current methods are either *reference-based* or *reference-free*.

Several question-answering datasets provide factual references in the form of entities; reference-based evaluation can then be based directly on entity matching. Other NLP datasets offer text references, so an alternative reference-based method involves matching the extracted entities between the generated text and these references (Nan et al., 2021). Instead, dialogue datasets provide reference responses. To evaluate the factual accuracy of the generated text, we can therefore compare the entities extracted from the generated response with those in the reference response, measured as F1 score. However, relying solely on this can be inadequate, as the reference may not always provide explicit answers. It is crucial to consider additional metrics to ensure a comprehensive evaluation.

Reference-free methods can be categorised into uncertainty estimation (Farquhar et al., 2024) and external knowledge-based approaches, with the former often limited by its reliance on the generation model's own confidence. The fact score (Min et al., 2023), which falls into the latter category, is a metric designed to measure fact consistency in long-form text. The process begins with an LLM breaking down the text into fine-grained sentences called atomic facts. These atomic facts are then verified using both the LLM and external knowledge sources (Min et al. (2023) use Wikipedia titles). The fact score is calculated based on the precision.

However, the fact score is not directly suitable for our work for two key reasons: First, it evaluates only the generated text without taking the dialogue history into account, which is crucial when assessing the factuality of dialogue responses. Second, it does not account for the possibility that generated responses may lack external knowledge sources for verification.

3 Methodology

The diagram in Figure 2 illustrates the workflow of dialogue generation (left side) and factuality evaluation (right side).

Our proposed framework consists of three key components: (1) A well-designed knowledge triple retriever and see Section 3.2 for details, (2) A CoT-based dialogue rewriter, described in Section 3.3, and (3) the knowledge-enhanced dialogue response generation (Section 3.2).

For factuality evaluation (Section 3.5), we adopt the fact score as a primary metric. However, as the original fact score does not fully capture certain aspects of dialogue evaluation, we introduce modifications to better align with our objectives.

3.1 Task Formulation

Our task is to generate a response s given a dialogue context c, query x and set of triples $f = \{(h_1, r_1, t_1), (h_2, r_2, t_2), ..., (h_n, r_n, t_n)\}$ from an encyclopedia-based KG, where h and t represent the head and tail entities, and r denotes the relation between them.

3.2 Knowledge Retriever

The KG is a collection of triples, and these triples are the form of head entity, relation, and tail entity.

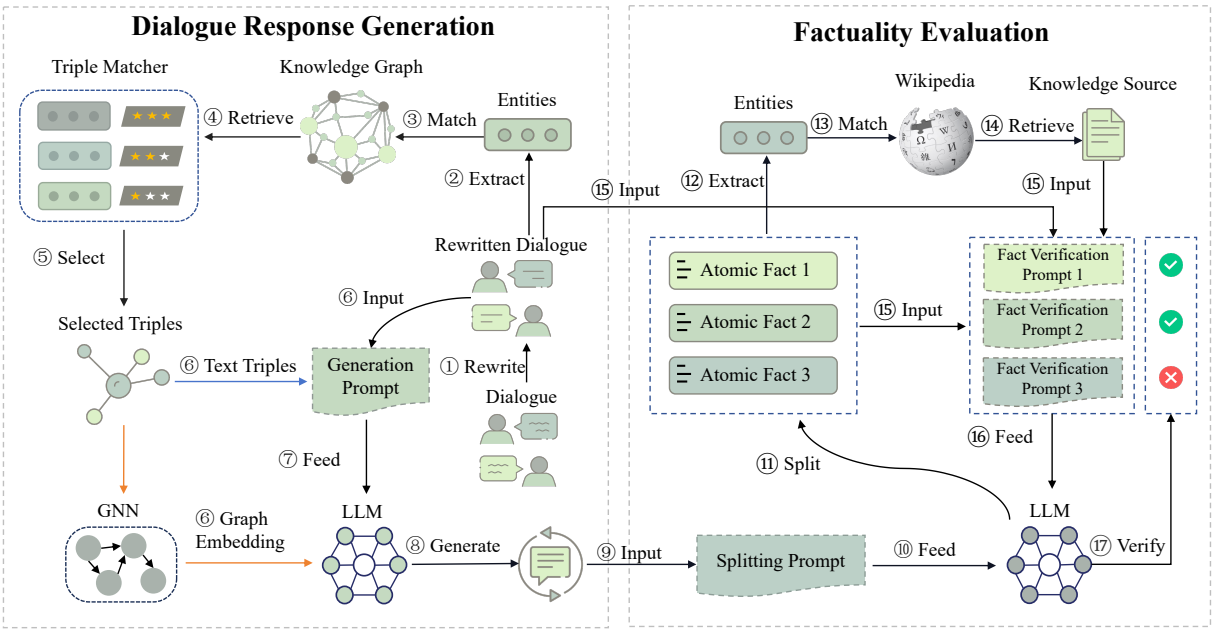

Figure 2: The workflow of dialogue response generation and factuality evaluation. The left part of the figure is our framework. It starts with rewriting the dialogue, and then the knowledge triples will be retrieved from KG. The triple retriever helps select valuable triples. The colour-coded arrows marked as Serial Number 6 represent different knowledge-enhanced methods: the orange arrow indicates the graph-based approach, while the blue arrow corresponds to the prompt-based approach. The right part is factuality evaluation, in which the response will be broken down into individual and verifiable facts, checked using both the LLM and external knowledge sources.

In encyclopedia-based KGs, like Wikidata (Vrandečić and Krötzsch, 2014), a triple usually indicates a fact. We aim to retrieve triples from the KG for the query in dialogue and incorporate triples into prompts as a supplement.

We first extract entities from the query using the entity linking approach, and then, we match all entities with the triples from the KG when the entity equals the head or tail entity. Since our entities are sourced from Wikidata, and ReFinED (Ayoola et al., 2022) is an entity linking method based on corresponding Wikidata entities, we adopt it for our retrieval process.

Knowledge retriever gathers all triples associated with the entities extracted from the query, but not all of these are relevant to the query. Irrelevant triples introduce textual noise, which has a negative impact on text generation. Thus, it is crucial to select the most relevant triples for the query. Traditional methods rank triples based on word frequency using BM25 (Robertson et al., 2009), which primarily captures word overlap. Alternatively, sentence embedding models like MPNet (Song et al., 2020), as employed by KAPING (Baek et al., 2023), encode semantic information but in a relatively straightforward manner.

To retrieve delicate knowledge, we collect query-triple pairs to fine-tune LLM as a knowledge matcher.

Data Collection Manual data annotation is both time-consuming and labour-intensive. Given GPT's strong performance in NLP tasks, we adopt it as an annotation tool. Our approach extracts query-triple pairs from the training dataset and prompts GPT-4o to determine their relevance. The model classifies each pair as either relevant or irrelevant. Based on our manual assessment, GPT-4o demonstrates high accuracy in determining query-triple relevance, described in Appendix F.

Fine-Tuning After collecting these data, we fine-tune an LLM as the binary classifier. In our work, we select the Llama3 8B model (Dubey et al., 2024). We combined LoRA (Hu et al., 2021) with fine-tuning, which is an efficient fine-tuning method that adopts low-rank modules into transformer layers. Cross-entropy loss is used to optimize the model during fine-tuning, and it is formulated as follows:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_i) \quad (1)$$

where y_i denotes the true label, p_i is defined as the predicted probability and N is the number of label classifications.

With the fine-tuned LLM $\mathcal{M}_t$, the query x and the triple f_i, the probability of relevance can be defined as:

$$P(y = \text{relevant}|x, f_i) = \mathcal{M}_{\text{t}}(x, f_i) \quad (2)$$

Then, the selected triples are formulated as:

$$f_{\text{sorted}} = \text{sort}_{\text{desc}}(P, f, x) \quad (3)$$

where $\text{sort}_{\text{desc}}$ denotes the sort function descending by P, and P is the relevance probability distribution.

3.3 Dialogue Rewrite

Unlike in question answering tasks, dialogue response generation involves coreferences that pose challenges for both entity linking and response generation. An intuitive approach to improve entity linking is to expand the scope for extracting entities from the dialogue context. However, this approach may also introduce irrelevant entities, negatively impacting performance.

To resolve coreference and improve the performance of entity linking and dialogue response generation, we propose a CoT-based coreference resolution.

Data Collection We prompt GPT-4o with dialogue to resolve the conversation in a zero-shot CoT setting, as described by Kojima et al. (2022). Step-by-step reasoning in LLMs enhances the generation of accurate results and aids in fine-tuning smaller language models.

Fine-Tuning We fine-tune the Llama3 8B model using LoRA on collected coreference resolution samples. The reason for selecting LoRA for this task is that it will not destroy the LLMs reasoning capability obtained from pre-training, making it perform better on this task. Cross-entropy loss is also selected for this task.

After fine-tuning, we prompt LLM to resolve coreference, and the dialogue is rewritten. The rewrite and corresponding sorted triples are formulated as follows:

$$c', x' = \mathcal{M}_{\text{f}}(p_{\text{rewrite}}, c, x) \quad (4)$$

$$f'_{\text{sorted}} = \text{sort}_{\text{desc}}(P, f, x') \quad (5)$$

where p_{rewrite}, described in Table 6 of Appendix A, is the prompt for rewriting dialogue. $\mathcal{M}_{\text{f}}$ is the LLM for dialogue rewriting.

3.4 Knowledge-Based Dialogue Response Generation

To mitigate hallucinations of LLM and have a factual response, we offer two patterns of generating dialogue responses given selected knowledge triples: prompt-based and graph-based methods.

Prompt Based Generation Given dialogue context c, query x, rewritten dialogue c' and x', as well as selected triples f_{sorted} and f'_{sorted}, the generation of knowledge-augmented dialogue responses and their counterparts with the rewritten dialogue are formulated as follows:

$$s = \mathcal{M}_{\text{g}}(p_{\text{gen}}, c, x, f_{\text{sorted}}) \quad (6)$$

$$s' = \mathcal{M}_{\text{g}}(p_{\text{gen}}, c', x', f'_{\text{sorted}}) \quad (7)$$

where p_{gen} denotes the prompt for generating the response. $\mathcal{M}_{\text{g}}$ is the LLM for response generation. Table 6 in Appendix A describes the details of the prompt template.

Graph-Based Generation Prompt-based generation can effectively improve factuality with training, while it is unable to capture the connectivity information within a Knowledge Graph. To address this limitation, we encode the knowledge graph with GNN (Graph Netural Network). We first convert the sorted triples f'_{sorted} into a graph $\mathcal{G}$. Then the graph and dialogue encoding are defined as follows:

$$v = E_{graph}(\mathcal{G}) \quad (8)$$

$$z = \mathcal{M}(c', x') \quad (9)$$

where E_{graph} is GNN. $\mathcal{M}$ represents an arbitrary LLM. v and z denote the graph and text embedding, respectively.

We generate the dialogue response by decoding the concatenation of the graph and text embedding, formulated as follows:

$$s_g = Decoder([v; z]) \quad (10)$$

where ; denotes concatenation and s_g is the response generated by graph-based method. To preserve the capability of the LLM, we combine LoRA with cross-entropy to train our graph model.

3.5 Evaluation Metrics on Fact

Fact score (Min et al., 2023) measures the fact consistency for long-form text. As we mentioned in Section 2.2, it does not consider the dialogue history and the possibility that the generated response may lack the knowledge source to support it.

The atomic facts and fact verification prompts of fact score are modified in this paper. In the revised atomic fact-splitting prompt, only complete sentences are permitted to be split, and the model is prohibited from adding any additional information to the atomic facts. We incorporate dialogue history and introduce the "Not Enough Information" label for the revised fact verification prompt. If an atomic fact is not a factual claim or lacks direct support from any known sources, the model will output 'Not Enough Information". The detailed prompts for atomic fact-splitting and fact verification are provided in Table 8 of Appendix A.

Since the original fact score framework does not specify how to get knowledge sources for dialogue, we extract entities from atomic facts and query as Wikipedia titles to search related passages as knowledge sources.

4 Experiment

4.1 Dataset

Two public dialogue datasets, OpendialKG (Moon et al., 2019) and HybriDialogue (Nakamura et al., 2022), are used to evaluate our methods. The two datasets both provide response references but not factual ones. See Appendix B for the details and examples of datasets.

4.2 Conventional Evaluation Metrics

We aim to evaluate factuality and quality for dialogue response generation. BLEU (Reiter, 2018), ROUGE-L (Lin, 2004), Perplexity (Jelinek et al., 1977) and F1 score (Nan et al., 2021) are employed to evaluate the quality and fact consistency of the generated responses. For more details, please refer to Appendix E.

4.3 Modified Evaluation Metrics

We adopt both the adapted fact score and the NEIP (Not Enough Info Proportion) to comprehensively evaluate factuality. We also manually assess fact score with humans, as shown in Table 1. See the Appendix F for more annotation details.

Datasets	Agreement	Cohen's Kappa
HybriDialogue	0.78	0.65
OpendialKG	0.78	0.67

Table 1: The agreement and Cohen's Kappa between the ground truth and LLM outputs regarding the fact score.

4.4 Baseline Methods

Several widely recognised LLMs, such as ChatGLM (Zeng et al., 2022) and Flan-T5 (Chung et al., 2022), have been selected for this work. Furthermore, we compare our proposed Triton, Triton-C and Triton-X with BM25 (Robertson et al., 2009), KAPING (Baek et al., 2023), G-Retriever (He et al., 2024), and a basic dialogue generation model. Additional details of LLMs can be found in the Appendix G.

4.5 Experimental Setup

We illustrate the details of the experimental setup in Appendix D.

4.6 Experimental Result and Analysis

Tables 2 and 3 show the experimental results of different methods on two public datasets. We analyze the generated responses from the perspectives of text quality and factual consistency.

Text Quality With the KG, the BLEU-4 and ROUGE-L scores fluctuate across two datasets. Our proposed Triton, the knowledge retriever, generally perform better than other knowledge-enhanced baselines, while we see a drop with Triton-C, the combination of dialogue rewrite and knowledge retriever. Our proposed framework, Triton-X, which consists of three components, achieves the highest BLEU and ROUGE-L scores, surpassing the SOTA G-retriever and demonstrating the superiority of our method.

There is a slight growth in PPL of smaller language models like ChatGLM and Flan-T5-Large: incorporating the KG decreases text fluency. After analysis, we found that the LLMs sometimes can not understand the KG well and thus generate unpolished responses. We also utilise Case 1 from Section 4.9 to illustrate the factors contributing to low PPL.

Overall, knowledge-augmented methods show variable performance across different LLMs in terms of BLEU-4 and ROUGE-L scores, and they tend to increase PPL slightly for smaller language

Methods	BLEU-4	ROUGE-L	PPL↓	Fact Score*	NEIP↓*	F1 Score
ChatGLM-6B	**1.42**	**14.24**	**12.98**	70.91	50.13	15.14
+BM25	1.38	14.02	13.15	70.99	49.07	14.92
+KAPING	1.31	13.93	13.15	72.37	48.74	15.34
+Triton((Ours)	1.37	14.07	13.31	74.64	48.73	**15.40**
+Triton-C(Ours)	1.34	13.47	14.41	**76.46**	**37.93**	15.33
Flan-T5-Large	**2.57**	**13.00**	18.37	67.76	54.94	**16.35**
+BM25	2.65	12.90	18.92	69.32	56.09	16.25
+KAPING	2.56	12.72	**18.18**	68.34	55.79	16.03
+Triton(Ours)	2.60	12.66	18.89	71.56	54.28	16.09
+Triton-C(Ours)	2.19	11.27	20.22	**74.25**	**28.53**	15.00
Flan-T5-XXL	**3.44**	**13.63**	23.37	69.29	36.62	14.07
+BM25	2.68	12.71	23.18	72.06	31.98	13.69
+KAPING	2.91	12.82	23.22	73.00	32.07	13.85
+Triton(Ours)	3.16	13.06	23.07	76.20	31.42	**14.50**
+Triton-C(Ours)	2.33	10.69	**20.03**	**77.77**	**20.35**	13.78
G-Retriever	3.79	**26.00**	14.98	84.79	62.65	11.70
Triton-X(Ours)	**3.99**	25.86	**14.43**	**87.73**	**59.20**	**13.75**

Table 2: The experimental result with various methods on the OpendialKG dataset. The column heads with * indicate our primary metrics for evaluating factual accuracy. **Ours** means generating responses with our proposed methods. The bold number is the best result within each model. PPL indicates Perplexity and a lower PPL means more fluency. NEIP denotes the "not enough information" proportion, representing the percentage of atomic facts that either lack direct knowledge support or fail to qualify as factual claims.

models. Our proposed Triton-X gains the best performance in BLEU-4 on both datasets and ROUGE-L on HybriDialogue, indicating a higher quality in generating responses.

Factuality With retrieved methods, the fact score rises in all LLMs, and Triton has a higher increase than KAPING and BM25. From Triton-C, we can see a continuous improvement in fact score, which means the rewritten dialogue also contributes a slight improvement to the fact score.

The NEIP is the proportion of atomic facts that can not be directly found in external knowledge to support or which are not a factual claim, and the lower one indicates the responses contain more verifiable facts. Compared to baseline LLMs, knowledge-augmented methods generally help reduce this proportion and produce more verifiable facts. From the tables, we observe a consistent decrease in NEIP for the Triton compared to KAPING and BM25. Furthermore, we noticed a significant improvement in NEIP with the rewritten dialogue. This suggests that rewriting the dialogue can substantially enhance factuality.

It is worth noting that the F1 score is not always consistent with the fact score. For instance, the trend of the F1 score shown in Flan-T5-XXL is almost opposite to the fact score on the OpendialKG dataset. One reason is that the reference responses lack explicit answers. Another reason is that the F1 score ignores the semantic aspects, and two sentences with the same entities can express contrary meanings.

In short, knowledge-augmented methods improve the factuality of LLMs, including the fact score and NEIP. Our proposed Triton series methods perform better than G-Retriever, KAPING and BM25 in the above metrics.

Model Scale We conduct experiments with different LLMs. The Flan-T5 series LLMs (Large and XXL) are beneficial for us in analysing the correlation between size and performance.

As the model scale grows, fact score and NEIP exhibit differing trends across the two datasets—factuality declines on HybriDialogue, whereas the OpendialKG dataset shows the opposite pattern. The factuality is not always consistent with the model size, which contradicts our intuition.

4.7 Ablation Study

We conduct an ablation study with our proposed framework, shown in Table 4. We remove our components one by one from Triton-X to observe the performance. We only report factually related metrics as they are our aim. The decline in performance after removing either the rewrite or the retriever component highlights their importance.

4.8 Human Evaluation Results

We report the human evaluation with the perspective: coherence, fluency and informativeness in

Methods	BLEU-4	ROUGE-L	PPL↓	Fact Score*	NEIP↓*	F1 Score
ChatGLM-6B	4.62	24.03	23.66	46.58	58.19	39.57
+BM25	4.68	24.19	23.77	53.32	55.18	40.26
+KAPING	4.56	24.07	**23.58**	55.84	54.51	**40.99**
+Triton(Ours)	**4.71**	**24.43**	23.95	58.15	53.07	40.96
+Triton-C(Ours)	4.53	24.23	26.03	**59.10**	**50.99**	40.92
Flan-T5-Large	**9.61**	**27.12**	30.89	60.04	50.48	40.04
+BM25	8.71	25.81	32.79	67.42	47.43	38.50
+KAPING	8.92	26.43	31.96	68.92	46.56	40.41
+Triton(Ours)	8.90	26.14	**30.81**	70.79	43.52	39.70
+Triton-C(Ours)	8.20	24.63	34.01	**74.41**	**33.68**	**40.63**
Flan-T5-XXL	**11.34**	29.62	37.47	53.06	50.62	44.07
+BM25	11.08	**29.75**	38.89	60.40	48.79	**44.65**
+KAPING	10.95	29.44	37.88	63.84	48.94	44.13
+Triton(Ours)	10.93	29.30	41.35	66.62	46.49	43.91
+Triton-C(Ours)	9.89	28.19	**32.48**	**69.34**	**39.23**	43.65
G-Retriever	20.36	39.53	49.66	70.29	55.77	42.05
Triton-X(Ours)	**20.85**	**40.77**	**39.12**	**73.00**	**54.87**	**43.98**

Table 3: The experimental result on the HybriDialogue dataset with different baselines. The best results within each model are displayed in bold.

	HybriDialogue		OpendialKG	
Method	Fact	NEIP↓	Fact	NEIP↓
w/o Retriever	-1.1	-0.11	-1.1	+1.2
w/o Rewrite	-2.3	+0.48	-1.46	+0.44
w/o Graph	-3.41	-0.03	-1.00	-1.52

Table 4: Ablation study of Triton-X, and w/o means removing corresponding components.

	HybriDialogue			OpendialKG		
Methods	Cohe.	Flu.	Info.	Cohe.	Flu.	Info.
ChatGLM-6B	1.90	**1.99**	1.71	1.99	1.91	1.11
+bm25	1.85	1.97	1.75	1.99	1.89	1.09
+SenEmb	1.84	1.98	1.71	1.98	1.95	1.18
+Triton	1.93	1.98	1.76	**2.00**	1.94	1.11
+Triton-C	**1.94**	1.96	**1.85**	1.98	**1.97**	**1.28**
Flan-T5-XXL	1.62	1.74	1.74	**1.56**	**2.00**	1.36
+bm25	**1.67**	1.72	**1.80**	1.38	1.94	1.49
+SenEmb	1.59	1.80	1.77	1.46	1.99	1.53
+Triton	1.59	1.81	1.77	1.52	1.98	1.39
+Triton-C	1.51	**1.83**	1.76	1.47	1.97	**1.65**
G-retriever	1.92	**1.82**	**1.84**	1.96	1.96	**0.71**
TritonX	**1.96**	1.81	1.82	**2.00**	**2.00**	0.60

Table 5: Human evaluation results for coherence (Cohe.), fluency (Flu.), and informativeness (Info.). The best results are marked within each category.

Table 5. Each aspect is rated as one of three levels: positive, neutral, or negative (e.g., coherent, neutral, incoherent).

Our proposed Triton series methods demonstrate notable improvements in coherence across various settings, with the exception of the Flan-T5-XXL model. In terms of fluency, our methods achieve the highest scores for Flan-T5-XXL on HybriDialogue, as well as for ChatGLM and graph-based methods on OpendialKG. Overall, the proposed methods maintain strong performance in fluency and coherence. Additionally, the Triton series generally enhances informativeness.

4.9 Case Study

We use two cases that are beneficial for us in understanding our methods' limitations. See more descriptions in Appendix H.

5 Conclusion

Given the limitations of previous methods in accurately retrieving knowledge triples and their challenges in handling dialogue tasks due to coreference issues, we propose a novel framework. This framework features a carefully designed knowledge retriever to identify valuable knowledge triples, an advanced dialogue rewriter to resolve coreference effectively, and a knowledge-enhanced dialogue generation component to improve factuality.

Building on this, we also adapt the fact score to make it suitable for evaluating the factuality of the dialogue response and verifying its validity in the dialogue task.

The results demonstrate that our method significantly enhances the factual accuracy of dialogue responses compared to other baselines, including the SOTA G-retriever.

We analyse some cases and see some limitations in our current work. We plan to explore the mechanism of LLM reactions to triples in the future.

Limitations

Although we propose a well-designed framework for dialogue-related tasks, some limitations remain.

Our current knowledge retrieval method relies on the exact entity overlap between the query and Wikidata. This approach lacks semantic understanding, making it inefficient in retrieving semantically related knowledge triples. Constructing a graph database from Wikidata could be an effective way to address this problem, but building such a large-scale database presents significant challenges.

Additionally, our approach relies on knowledge triples. However, the validity of certain triples, such as ("Montevideo", "population", "309331") shown in the key knowledge of Case 1 in Section 4.9, will change over time. Retrieving such triples for LLMs could lead to outdated responses.

Ethical Statement

In this work, we follow the ethical standards in data use and research. The datasets used in this study, OpendialKG and HybriDialogue, are publicly available and have been widely used in research. We have not added any extra personal information to these publicly available datasets.

The data from human annotators was limited to some specific labels, and the results reported in this paper are based on statistical analysis. No personal or sensitive information was accessed during the research process.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghami, and Huan Liu. 2023. Can knowledge graphs reduce hallucinations in llms?: A survey. *arXiv preprint arXiv:2311.07914*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking. *arXiv preprint arXiv:2207.04108*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*.

Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haußmann. 2014. Easy access to the freebase dataset. In *Proceedings of the 23rd international conference on World Wide Web*, pages 95–98.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yanyang Li, Jianqiao Zhao, Michael R Lyu, and Liwei Wang. 2022. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. *arXiv preprint arXiv:2211.01587*.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

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.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. *arXiv preprint arXiv:2204.13243*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-source multi-type knowledge exploration and exploitation for dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12522–12537.

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.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Priyanka Sen, Sandeep Mavadia, and Amir Saffari. 2023. Knowledge graph-augmented language models for complex question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 1–8.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yike Wu, Nan Hu, Guilin Qi, Sheng Bi, Jie Ren, Anhuan Xie, and Wei Song. 2023. Retrieve-rewrite-answer: A kg-to-text enhanced llms framework for knowledge graph question answering. *arXiv preprint arXiv:2309.11206*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. Xdai: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4422–4432.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Prompts

The prompts of the knowledge-augmented approach are shown in Table 6. The factuality-related prompts are described in Table 8.

B Datasets Details

This section provides a detailed description of the datasets, listed as follows:

OpenDialKG It is an English knowledge-driven dialogue dataset that consists of a recommendation task related to movies and books and a chit-chat task related to sports and music. The dataset contains 13,802 samples. After processing, some samples are not valid in the task, so we remove them. Finally, we randomly selected 15% (1,962 samples) for the validation set, another 15% (1,973 samples) for the test set, and the remaining 70% (9,120 samples) for the training set. The dataset provides a number of triples extracted from Freebase (Bast et al., 2014). These triples were collected several years ago. To address any potential issues with outdated information, we re-extract the triples from Wikidata. OpendialKG is released under CC-BY-NC-4.0 license[1], which permits non-commercial research use.

HybriDialogue It is an open-domain and information-seeking dialogue dataset in English. It is constructed by splitting complex questions into multi-turn dialogue. The original dataset does not offer triples, so we collect them by matching the entities with triples from Wikidata. The training set contains 4,359 samples, while the validation set includes 242 samples, and the test set consists of 243 samples. HybriDial follows the MIT license (Nakamura et al., 2022), which allows both commercial and non-commercial use.

Table 9 presents the two examples of OpendialKG and HybridDial datasets with selected triples from different methods.

C Examples from GPT-4o

We employ GPT-4o to annotate query-triple pairs and resolve dialogue coreference from the training dataset. In the end, we collected 29K samples of query-triple pairs and 35.4K samples of coreference resolution with reasoning steps. The examples for fine-tuning are listed in Tables 10 and 11.

[1] https://github.com/facebookresearch/opendialkg

D Experimental Setup

We provide detailed information regarding the experimental setup of our framework and the evaluation metrics used.

For fine-tuning the Triton, we only collect the query-triple pairs from the HybriDial dataset. We fine-tuned 3 epochs for either the LoRA method or the full-tuning method. The rank and alpha parameters for LoRA are both set to 8.

We utilise the `multi-qa-mpnet-base-dot-v1` version for the baseline KAPING. As G-retriever is implemented using Llama2 7B (Touvron et al., 2023), we adopt the same model in our Triton-X under identical settings to ensure fairness. For knowledge-augmented methods, the maximum number of selected triples is set to 5. Additionally, all experiments were run a single time.

Regarding the evaluation metrics, we utilised the `evaluate` package [2] to compute BLEU-4 and ROUGE-L scores. Besides, the PPL is calculated using GPT-2 (Radford et al., 2019).

We fine-tune our relevance classifier using a single A100 GPU and our dialogue generation model with three A100 GPUs. We utilized two A100 GPUs for inference tasks. All A100 GPUs we used in this work have 80GB of memory. In total, our GPU runtime amounted to approximately 575 hours.

E Conventional Evaluation Metrics in Detail

We list all the details of conventional evaluation metrics as follows:

- **BLEU** (Reiter, 2018), used in the machine translation initially, calculates the n-gram overlap between generated text and references and reflects the correlation between generated text and human writing.
- **ROUGE-L** (Lin, 2004) is calculated based on the length of the longest common subsequence, which captures the word order and sentence-level structure from the ground truth.
- **PPL** (Jelinek et al., 1977) is widely used in measuring text fluency of generated output, in which calculation is based on the language model.
- **F1 Score** (Nan et al., 2021) is based on the precision and recall of the extracted entities

[2] https://huggingface.co/docs/evaluate

Dialogue Rewrite Prompt
You are tasked with resolving all pronouns and references in the given dialogue to their explicit entities. Use CoT (Chain of Thought) reasoning to identify what each pronoun or reference corresponds to. Do not answer any questions; your only goal is to perform co-reference resolution. Instructions: 1. Analyze the dialogue and process each turn in the conversation. 2. For every pronoun, ambiguous term, or reference, trace back in the conversation to determine its explicit entity or subject. 3. Clearly document your CoT reasoning for each resolution. 4. Provide the explicit reference for each pronoun or ambiguous term. Output Format: **Chain of Thought**: [Your reasoning process for resolving the references] **Resolved Dialogue**: [The dialogue with all pronouns and references resolved] **Dialogue**: {Dialogue}

Table 6: Prompt for rewriting the dialogue and generating a response with dialogue and knowledge triples.

Dialogue Response Generation Prompt	**Atomic Fact Splitting Prompt**
Knowledge: {Selected Triples} **Dialogue**: {Dialogue Context} Given the above knowledge and dialogue, please respond to the input below and ensure the response is fluent and fact-consistent in English. **Input**: {Query} **Response**: ...	**{Examples..}** If the following input is an incomplete sentence or a phrase, please output it exactly as it is. Otherwise, if it is a complete sentence, split it into atomic sentences based only on the given information, without adding any additional information or making inferences: **Input**: {Response} **Output** ...:

Table 7: The prompts for getting atomic facts and fact score.

matching between the generated response and ground truth.

F Details of Manual Annotation

We invited some students from Asian countries to annotate samples. All of them are well-educated and possess good English proficiency. The annotators were told that the annotated data would be used for research, and we got consent for data use.

To assess the accuracy of query-triple pair annotated from GPT-4o, we invite tree annotators to this task. We randomly selected 100 samples and instructed participants to determine whether the query and triple were relevant. The output is either *relevant* or *irrelevant*. In the beginning, two annotators have 0.83 in agreement and 0.66 in Cohen's Kappa. We introduced another annotator to mediate the disagreement. We utilized the final annotated result to assess the accuracy of query-triple pairs, which is 0.81 in agreement and 0.62 in Cohen's Kappa. These results indicate a high level of accuracy.

Three annotators were involved in the assessment of the fact score task. We used the fact score prompt, described in Table 8, as an instruction. Two annotators independently assessed the fact score for 100 samples from the HybriDial and OpendialKG datasets respectively. Initially, the raw agreement scores were 0.76 for HybriDial and 0.81 for OpenDialKG. Their inter-annotator agreement, measured by Cohen's Kappa, was 0.613 for OpendialKG and 0.705 for HybriDial. To resolve discrepancies, we introduced a third annotator to mediate disagreements. Finally, we report the final agreement and Cohen's Kappa score with the evaluation model, as shown in Table 1.

Two annotators were involved in the dialogue human evaluation. The accuracy of human annotations is consistently high across all aspects, with HybriDial achieving 0.81 for coherence, 0.88 for

Fact Score Prompt
Instruction: The statement is part of a response in a dialogue. Evaluate the statement strictly based on the provided knowledge source and dialogue history only. If the statement is not a factual claim (e.g., opinion, question, or unclear assertion), output: "no enough information." If it is a factual claim: Output true if the statement is directly supported by evidence in the knowledge source or dialogue history. Output false if the statement is directly contradicted by the knowledge source or dialogue history. Output no enough information if there is no direct evidence for or against the statement. Important: Do not use your intern knowledge or make inferences. Please only output your final answer and do not output any explanations. **Evidence**: {Wikipedia Passages} **Dialogue history**: {Dialogue} **Speaker A**: {Speaker} **Statement**: {Atomic Fact}

Table 8: The prompts for getting atomic facts and fact score.

fluency, and 0.89 for informativeness, while OpendialKG reaches 0.83, 0.96, and 0.80 respectively, indicating the overall reliability of manual evaluations. We finally report the average score in our dialogue human evaluation.

G Baseline Descriptions

LLMs used in this work are presented as follows:

- **ChatGLM** (Zeng et al., 2022) is an open-source and bilingual language model, based on GLM (Du et al., 2021) architecture. The technique of ChatGLM-6B is similar to ChatGPT, optimized in dialogue tasks.
- **Flan-T5** (Chung et al., 2022) proposed several instruction-based LLMs, scaling the number of tasks and model size and fine-tuning in the chain-of-thought data. We adopt Flan-T5-XXl (11B), Flan-T5-Large (780M) and Flan-T5-Small (80M) as baselines in our work.

The baseline methods are described as follows:

- **No Knowledge**: We feed the prompt into LLMs to generate dialogue responses without external knowledge.
- **BM25** (Robertson et al., 2009): It is the ranking function between a query and several documents widely used in information retrieval. In our experiments, we also apply RefinED to retrieve triples for BM25 and replace documents with triples for retrieval.
- **KAPING** (Baek et al., 2023): It is a popular knowledge-augmented method for question answering. Since the authors have not released the official code, we follow the experimental setup of the KAPING method: we first extract entities from the query using RefinED, rank the query with triples using MPNet, select the top-N most relevant triples, and supplement them as prompts to generate dialogue responses.
- **G-Retriever** (He et al., 2024): It is a retrieval-augmented generation method for textual graphs that selects informative subgraphs via a Prize-Collecting Steiner Tree, encodes them with GNNs, and generates responses with high accuracy, outperforming strong baselines and reducing hallucinations.
- **Triton(Ours)**: We use GPT-4o to generate high-quality query-triple pairs based on the

OpendialKG Example
Query: Could you recommend some movies staring Chiaki Kuriyama? **Retrieved triples from BM25**: (Chiaki Kuriyama, given name, Chiaki) (Chiaki Kuriyama, family name, Kuriyama) (Chiaki Kuriyama, place of birth, Tsuchiura) (Chiaki Kuriyama, occupation, film actor) (Chiaki Kuriyama, instrument, voice) ... **Retrieved triples from KAPING**: (Chiaki Kuriyama, given name, Chiaki) (Chiaki Kuriyama, occupation, model) (Chiaki Kuriyama, occupation, singer) (Chiaki Kuriyama, occupation, actor) (Chiaki Kuriyama, occupation, fashion model) ... **Retrieved triples from Triton**: (Kill Bill Volume 1, cast member, Chiaki Kuriyama), (Into the Sun, cast member, Chiaki Kuriyama), (Kagen no Tsuki, cast member, Chiaki Kuriyama), (Kamogawa Horumo, cast member, Chiaki Kuriyama), (Gonin, cast member, Chiaki Kuriyama) ...
HybriDialogue Example
Query: Hi. Can you tell me who Judd Trump is? **Retrieved triples from BM25**: (Judd Trump, given name, Judd)(Judd Trump, family name, Trump)(2021 Champion of Champions, winner, Judd Trump)(Judd Trump, victory, 2016 China Open)(Judd Trump, victory, 2011 China Open) ... **Retrieved triples from KAPING**: (Judd Trump, occupation, snooker player) (Judd Trump, nickname, The Ace) (Judd Trump, given name, Judd) (Judd Trump, family name, Trump) (Judd Trump, place of birth, Whitchurch) ... **Retrieved triples from Triton**: (2019 World Snooker Championship, winner, Judd Trump), (Judd Trump, award received, Snooker Hall of Fame), (Judd Trump, country of citizenship, United Kingdom), (Judd Trump, sport, snooker), (Judd Trump, award received, player of the year award) ...

Table 9: The examples extracted from OpendialKG and HybriDialogue dataset. The triples are retrieved from Wikidata and selected using different methods.

training dataset and fine-tune Llama3 8B base model with collected samples. The model acts as a triple retriever to help select valuable triples for generating dialogue responses.

- **Triton-C(Ours)**: We fine-tune a coreference resolution model based on Llama 8B base model using reasoning distillation. Our approach integrates dialogue rewriting with entity linking for enhanced performance.
- **Triton-X(Ours)**: It is the framework that consists of our well-designed knowledge retriever, dialogue rewrite and graph-based module.

H Details of the Case Study

The examples of case studies are shown in Table 12.

The topic in Case 1 is about the population of Montevideo, and our methods do help select the relevant triples. However, the LLM then generates one of triples directly, which is not fluent and has a higher PPL i.e. lower *quality*.

Case 2 is a dialogue that asks about the release date of "Team Fortress 2", and the original response responds to the dialogue and is close to the writing of reference. After rewriting the dialogue, the LLM generates a response that differs from the reference, causing a lower BLEU-4 score.

Query	Triple	Relevance
Which teams has the Colorado Avalanche played against in the playoffs?	(Colorado Avalanche, sport, ice hockey)	Irrelevant
Hi! Do you know anything about Lutheran churches?	(Lutheranism, described by source, Brockhaus and Efron Encyclopedic Dictionary)	Irrelevant
Tell me more about Upper Canada District School Board?	(Upper Canada District School Board, headquarters location, Brockville)	Relevant
Hi, do you by any chance know anything about Vladimir Horowitz?	(Stefanos Korkolis, student of, Vladimir Horowitz)	Relevant

Table 10: The query and triple are sourced from the training dataset and Wikidata. The relevance is annotated by GPT-4o.

Section	Content
Original Dialogue	**Speaker A:** I like Lev Grossman books can you recommend some of his book **Speaker B:** Sure! He wrote The Magician's Land and The Magicians. **Speaker A:** Oh, I think I've read The Magicians it a High Fantasy book.
CoT	Speaker A mentions "Lev Grossman" as the author they like. This establishes that "his" in their request for book recommendations refers to Lev Grossman. Speaker B mentions "He wrote The Magician's Land and The Magicians." The pronoun "He" refers to Lev Grossman, as he is the author being discussed. Speaker A uses "it" in "I think I've read The Magicians it a High Fantasy book." The pronoun "it" refers to "The Magicians," as it is the last explicitly mentioned book in the conversation.
Resolved Dialogue	**Speaker A:** I like Lev Grossman books can you recommend some of Lev Grossman's books **Speaker B:** Sure! Lev Grossman wrote The Magician's Land and The Magicians. **Speaker A:** Oh, I think I've read The Magicians. The Magicians is a High Fantasy book.

Table 11: An example generated by GPT-4o illustrates how CoT reasoning works in coreference resolution.

Case 1	Case 2
Context: A: What is the capital and largest city in Uruguay? B: Montevideo is the capital and largest city in Uruguay. **Query**: What was the population of Montevideo, Uruguay, in 2011? **Response**: The population of Montevideo, Uruguay, in 2011 was 59,027. **Key Knowledge**: (Montevideo, population, 309331) **New Response**: Montevideo, population, 309331	**Dialogue**: A: What kind of game was "Team Fortress 2"? B: It's a multiplayer first-person shooter. A: And what was it's release date? **Response**: It was released on October 26, 2007. **Rewritten Dialogue**: B: What kind of game was "Team Fortress 2"? A: "Team Fortress 2" is a multiplayer first-person shooter. A: And what was "Team Fortress 2"'s release date? **New Response**: "Team Fortress 2" was released on November 20, 2007.

Table 12: The responses of cases were both generated by Flan-T5-Large. The new response was created using key knowledge or a rewritten dialogue. Triton selected the key knowledge.