EDIT: Towards Enhancing Dialogue Response Generation for Large Language Models by Asking Questions to Detect User's Intentions

Anonymous ACL submission

Abstract

Large Language Models (LLMs), such as ChatGPT, have recently been applied to various NLP tasks due to its open-domain generation capabilities. However, during the dialogue process, users may have implicit intentions that might be overlooked by LLMs. Besides, it is unlikely for LLMs to encompass all fields comprehensively and LLMs cannot update the latest knowledge in real-time. To tackle these two issues, we propose a framework *using LLM to **E**nhance dialogue response generation by asking questions to **D**etect user's **I**mplicit in**T**entions* (**EDIT**). Firstly, we construct a *Context-Open-Question* (COQ) dataset to train a question generator (QG) and generate open questions related to the dialogue context as the potential user's intention; Then, EDIT answers those questions by interacting with LLMs and retrieving domain-specific knowledge bases respectively; Finally, EDIT generates response by integrating those answers. To evaluate generated responses, we have specifically designed two metrics, *Information Content* (IC) and *Context Coherence* (CC), respectively. The results demonstrated significant improvements after combining current mainstream LLMs with EDIT on two task-oriented dialogue dataset (Wizard of Wikipedia and Holl-E).

1 Introduction

Large language models (LLMs) like ChatGPT[1] and LLaMA (Touvron et al., 2023a) have recently demonstrated remarkable performance on various natural language processing tasks (Bang et al., 2023), such as commonsense reasoning (Bian et al., 2023), sentiment analysis (Wang et al., 2023b), recommendation system (Wang et al., 2023a), etc. The powerful generative capabilities of LLMs increasingly attract widespread attention, especially for generation tasks.

[1]https://chat.openai.com

However, in dialogue generation tasks, LLMs sometimes overly focus on the semantic coherence of the context. As a result, there are a few instances where LLMs would generate plain responses that lack sufficient information and may not fully satisfy the user's needs. As depicted in Figure 1, in response to user's utterence "*Forgetting and memory loss is one of life most painful things*", the LLM only expresses "*It can be really frustrating when we forget something important or experience memory loss*" to cater to the user, which fails to provide additional information that users might be interested in, such as the causes and possible treatments for "Memory Loss". Additionally, previous evaluation metrics, such as BLUE and ROUGE-L, face challenges in discerning whether the generated responses contain sufficient information or exhibit semantic coherence, as they mainly consider word-level overlapping between the generated response and the golden answer.

There are two main reasons for the phenomenon that LLMs generate plain response.

Firstly, during conversations, LLMs may overlook users' implicit intentions. LLMs contain a considerable amount of knowledge and instruction-tuning methods enable LLMs to follow human instructions to solve problems by utilizing their knowledge. But those human instructions typically express intention directly in the text, in contrast, users usually have many implicit intentions in the conversation, which are not explicitly present in the context. As a result, during generating response, LLMs tend to prioritize dialogue coherence rather than utilizing relevant knowledge, because they haven't encountered such implicit intentions.

Secondly, LLMs are not all-encompassing and may not cover every field comprehensively. In a few specific domains, LLMs' knowledge may also be incomplete and they cannot update the latest knowledge in real-time. These limitations may hinder the ability of LLMs to generate responses for

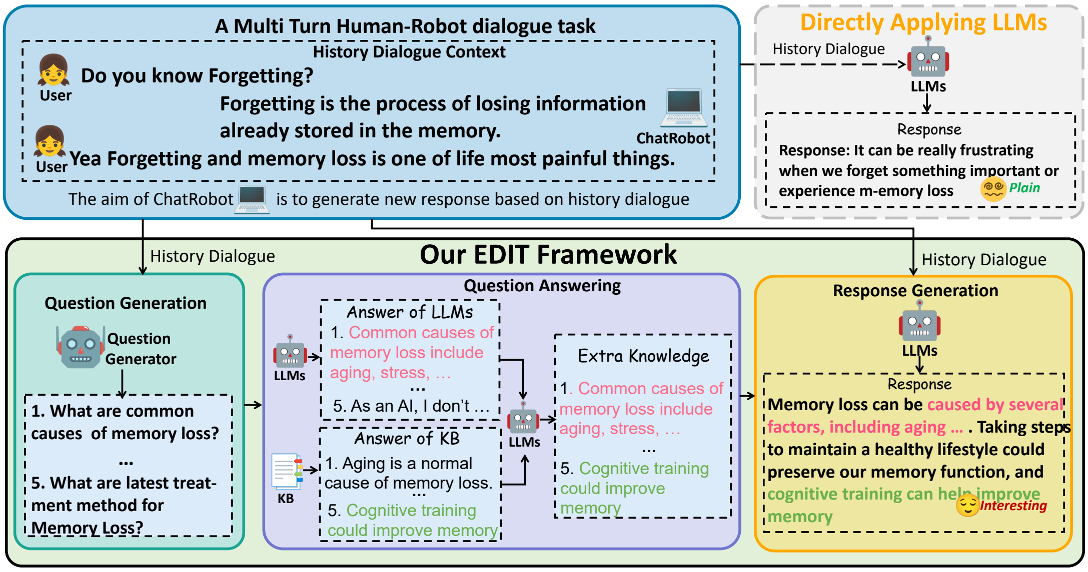

Figure 1: The framework of EDIT. The upper left part is an example of multi turn human-robot dialogue task. The upper right part represents the method of directly applying LLMs to generate response. And the lower part is the illusion of our EDIT framework.

the sake of aligning with the implicit intentions that users truly want to be understood. As demonstrated in Figure 1, “What are latest treatment method for Memory Loss?” would be the users' implicit intention, while LLM is unable to answer such questions and, more fundamentally, cannot integrate the relevant knowledge in the process of generating responses.

To address these issues, in this work we propose a framework to Enhance dialogue response generation for LLMs by asking questions to Detect user's Implicit inTentions (**EDIT**). This framework consists of three main modules: Question Generation, Question Answering, and Response Generation. Firstly, we create a *Context-Open-Question* (COQ) dataset to train a question generator and generate questions to capture users' implicit intention; Secondly, we utilize a LLM and a domain-specific knowledge base to answer the generated questions and integrate their answers as a kind of extra knowledge; Finally, we input the extra knowledge as prompts, together with the dialogue context, to generate new responses. Furthermore, in terms of comprehensively evaluating the generated response, we design two metrics *Information Content* (IC) and *Context Coherence* (CC).

We utilize representative LLMs as the backbones of EDIT, and evaluate their performance on two task-oriented dialogue datasets, Wizard of Wikipedia and Holl-E. EDIT achieves significant improvements in both human evaluation and two automatic metrics, which demonstrates the flexibility and effectiveness of EDIT, in helping LLMs generate responses with more information catering to user's intention. Furthermore, the ablation experiments verify that integrating answers of integration of LLM answers and domain-specific knowledge base can indeed enhance overall performance. Besides, we also compare our question generator with LLMs, demonstrating our model could generate more appropriate questions regarded as users' intentions.

2 Approach

In this work, we focus on the usage of LLMs in multi-turn human-robot dialogue. Specifically, we denote the dialogue history of a conversation at time step t as $C = \{u_1, r_1, ..., u_t\}$, where u and r represent the utterances from the user and the chatbot respectively. The K denotes extra context-related knowledge. The purpose of our task is to generate the chatbot's responses r_{t+1} by using extra knowledge K and dialogue history context C.

As illustrated in upper right part of Figure 1, previous work directly applies LLMs to dialogue tasks by inputting the dialogue context. In the lower part of Figure 1, our EDIT framework consists of three modules: Question Generation Module, Question Answering Module, and Response Generation Module. Firstly, the Question Generation Module generates a series of open questions based on the dialogue context regarded these questions as the user's implicit intentions. Secondly, the Question Answering Module obtains the answers to the questions. Finally, the Response Generation Module inputs the extra knowledge and historical dialogue context to LLMs to generate a new response.

2.1 Question Generation

We create a Context-Open-Question (COQ) dataset and train a question generator (QG) based on it with the aim that generating context-related questions. Then, we regard these questions as user's implicit intentions.

2.1.1 Context-Open-Question Dataset Construction

As shown in Figure 2, to enable the question generator of EDIT framework to ask open questions for dialogue context in various scenarios in life, we develop the Contex-Open-Question dataset. We train a question generator (QG) using the COQ datasets. When constructing the dataset, we employ a method combining ChatGPT and manual annotation. This approach allows us to create high-quality data consuming less labor cost to meet the requirements of our EDIT framework.

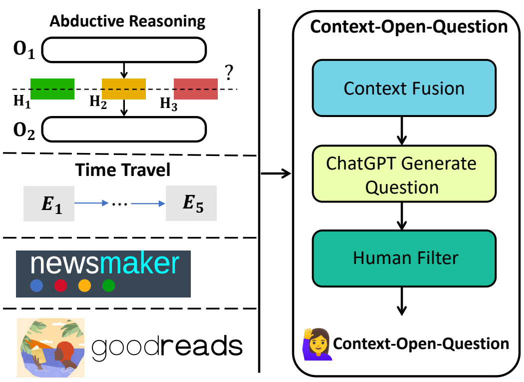

Figure 2: The process of COQ dataset curated.

Context Datasets Selection. We carefully select four datasets that represent different daily life scenarios, such as book revews, news, abductive reasoning, and story. These datasets are merged to create Context part of COQ dataset. **Goodreads Reviews**[2] (GR) is a comprehensive book review dataset to address real-life large-scale application scenarios. **News Category** (Misra, 2022) (NC) extracted news headlines from HuffPost in 2012-2022 to address the issue of the overwhelming amount of fake news in today's world. **Abductive Commonsense Reasoning** (Bhagavatula et al., 2020) (ACR) is inference to the most plausible intermediate process events based on two causally related events. **Time Travel** (Qin et al., 2019) (TT) offers a narrative experience encompassed within five sentences, seamlessly unfolding in continuous time.

Open Question Annotation. To enable the question generator (QG) to generate context-related questions about diverse extra knowledge, we initially have ChatGPT generate questions based on context through the designed prompt "Give you a context: {Context}. Help me ask questions, which is unrelated to the person in the context ... ". Afterward, we manually filter the generated Context-Open-Question pairs to ensure that the questions are relevant to the context. Additionally, we also ensure that the questions can be answered by LLMs.

		Train	Test	Valid
COQ	ACR	1564	108	115
	TT	1334	78	83
	NC	2291	197	195
	GR	689	34	42

Table 1: The COQ dataset statistics.

After manual filtering, we construct a total of 5878 Context-Open-Quesion (COQ) training data.

2.1.2 Question Generator

To enable the EDIT framework to generate various context-related open questions, we train a question generator (QG) based on the Context-Open-Question data.

The input context is denoted as C. For each context C, there are multiple corresponding questions $q_1, ..., q_n$. We concatenate all the questions to form a sequence denoted as:

$$Q = [q_1, ..., q_n],$$

To generate the questions, we employ a generative model that generates a sequence $Q = [y_1, ..., y_T]$ of length T given the input context C. The probability of decoding each token y_t is conditioned on the previously generated tokens $y_{<t}$ and the input context C:

$$P(y_t|y_{<t}, C) = \text{GeneraiveModel}(y_{<t}, C),$$

we use the standard negative log-likelihood (NLL) loss on the target question sequence Q to train the

[2] https://mengtingwan.github.io/data/goodreads.html

model. The loss is computed as:

$$Loss = -\sum_{t=1}^{T} \log P(y_t|y_{<t}, C).$$

2.2 Question Answering

As shown in right part of Figure 1, Question Answering part of EDIT consists of three components: Generating Answer with LLMs, Retrieving Answers in Knowledge Base, and Answer Integration.

2.2.1 Generating Answer with LLMs

Although LLMs possess numerous commonsense knowledge, they may ignore the users' implicit intention and do not use relevant knowledge. Therefore, we use LLM to answer generated questions using the designed prompt "Give you a question: question, Please answer it as briefly as possible.":

$$A_{LLM} = \text{LLM}(q),$$

where q is the questions generated by QG.

2.2.2 Retrieving Answer in Knowledge Base

LLMs unlikely fully understand the knowledge in all fields and cannot update it in real-time. As a result, there are certain questions that they are unable to answer. To overcome this limitation, for each downstream task, we retrieve answers in the corresponding Domain-Specific Knowledge Base.

For each question q generated by the QG, we send it to SentenceBERT (Reimers and Gurevych, 2019), and pass the final layer's output through mean pooling to obtain question's representation e_q. We also use SentenceBERT to obtain representation e_s of each knowledge s in the KB.

We calculate the semantic similarity (cosine similarity) between question q and knowledge s:

$$sim = \text{CosSim}(e_q, e_s),$$

We select the knowledge $s_1, s_2, ...s_L$ with the top L semantic similarity as the knowledge related to the question q. In our experiments, we set $L = 10$. In addition, this knowledge is discrete. Thereby, we use LLM to organize that knowledge as the answer:

$$A_{KB} = \text{LLM}(s_1, s_2, ...s_L).$$

2.2.3 Answer Integration

LLMs would refuse to answer some questions and the answers obtained through retrieving in knowledge base are often one-sided. Thereby, we use the designed prompt "Give you a question: q, and two answers to it, AnswerA: answerLLM, AnswerB:answerKB, please tell me which is better?" to make LLMs integrate those two kinds of answers:

$$A = \text{LLM}(q, A_{LLM}, A_{KB}).$$

We concatenate all questions' answers as the extra knowledge:

$$K = [A_1, ..., A_n],$$

where A_i is the answer of question q_i, n is the number of questions, and $[\cdot]$ represents the text concatenation.

2.3 Response Generation

While LLMs excel at generating conversational responses, in some instances, they may ignore the users' implicit intention. To address this limitation, we utilize context-related knowledge generated by Question Answering Module as supplementary input for LLMs. By incorporating historical dialogue context C and extra context-related knowledge K, we use the prompt "Give you a context: {context} and some knowledge {knowledge}. Please use those knowledge to just generate next response of {next person}." to make LLMs generate responses:

$$R = \text{LLM}(C, K),$$

The R is the response generated by our EDIT.

3 Experiment

In this work, we conduct experiments on two chosen dowmstream tasks, Wizard of Wikipedia and Holl-E, which could reflect real-life scenarios and require domain-specific knowledge.

3.1 Downstream Tasks

Wizard of Wikipedia (Dinan et al., 2019) provides a large-scale dialogue dataset, where each sample provides required commonsense knowledge for each response in the dialogue.

Holl-E (Moghe et al., 2018) provides a dataset of conversations about specific movie topics. Each conversation in it includes a document that contains information about the corresponding movie.

For each sample in our chosen downstream task-oriented tasks, it provides the associated knowledge document. We split all those knowledge documents into sentences and build a Domain Specific Knowledge Base for each downstream task.

3.2 Evaluation Metrics

Previous automatic evaluation metrics, such as BLUE and Rouge-L, mainly focus on whether the generated text is consistent with the golden response. In actual usage, there is no standard answer for the response generation task, and users would prefer the model to provide more content that they may be interested in. To address this issue, we have designed two evaluation metrics, Information Content(IC) and Context Coherence(CC). Besides, we also conduct human evaluation to more comprehensively evaluate the model's performance.

Information Content (IC). We use the information entropy to measure the information content of the generated response R. We calculate the frequency of each word w appearing in this response as f_w. Then we calculate the information entropy as Information Content metric:

$$\mathrm{IC}(R) = -\sum_{w \in R} -f_w * \log(f_w),$$

Context Coherence (CC). In addition to considering the amount of information contained in the generated response, we also use BERT to calculate the similarity between the generated response and the context to evaluate their relevance. However, we find that sometimes the model generates very short responses with limited relevance to the context, resulting in high semantic similarity scores. We assume that in a context, there usually needs to be a certain amount of information content for dialogue response. Therefore, regarding the amount of information in the context as the standard, we combine SentenceBERT (Reimers and Gurevych, 2019) and Information Content (IC) to measure the degree of Context Coherence between the response and the context:

$$\mathrm{CC}(R, C) = \mathrm{sim}(R, C) \times scale,$$

where T represents the golden response, R represents the generated response, $\mathrm{sim}(R, C)$ represents the semantic similarity calculated by SentenceBERT, and $scale = \mathrm{IC}(R)/\mathrm{IC}(C)$ represents the scaling factor. We reduce the scores of overly short responses by comparing the information content between generated response and context.

Human Evaluation. There is still a certain gap between these automated assessments and the real feedback from users. Therefore, we conduct a human evaluation to assess the generated results of EDIT and other LLMs. The human-evaluated Reasonable metric is mainly to evaluate whether the generated response is consistent with the context semantics and to use additional useful knowledge.

3.3 Compared Methods

We compare mainstream LLMs and our EDIT framework based on different LLMs as backbones.

Applying LLMs Directly. We also use some OpenAI-series LLMs in our compared system, including OpenAI's Text-Davinci-001, Text-Davinci-002, Text-Davinci-003 and Gpt-3.5-Turbo (ChatGPT). To systematically evaluate the model's performance, we conduct experiments on several mainstream open-source LLMs including LLaMA (Touvron et al., 2023b), LLaMa2 [3], ChatGLM (Du et al., 2022), and FLAN-T5-XL (Chung et al., 2022).

Our EDIT Framework. We use the various LLMs mentioned above as the backbone of the EDIT framework. As for the OpenAI-series LLMs, we just select the Gpt-3.5-Turbo (ChatGPT) as backbone due to its significant performance. Specifically, when using different LLMs in EDIT, the Question Answering Module in EDIT remains using ChatGPT for the sake of obtaining effective knowledge related to the context. In addition, we regard the EDIT(Gpt-3.5-Turbo) as our main model.

3.4 Main Results

Overall Analysis. On the two downstream tasks of Holl-E and Wizard of Wikipedia, we use the current mainstream LLMs and EDIT to generate responses.

Table 2 illustrates the performance of the models on downstream tasks. Combining EDIT with most LLMs can bring improvements to almost all mainstream LLMs. On the Holl-E dataset, some models show less significant improvements and even a decrease in Context Coherence. This is mainly because the topics of the Holl-E dataset are relatively simple and adding too much additional information may make the responses not semantically aligned with the context. Besides, LLaMa and LLaMa2 do not have excellent instruction-following abilities, which results in them having relatively worse performance on response generation tasks.

The results of human evaluation also effectively demonstrate that the responses generated by our

[3] https://ai.meta.com/resources/models-and-libraries/llama/

Types	LLMs	Wizard of Wikipedia			Holl-E		
		HUMAN	IC	CC	HUMAN	IC	CC
Open-Source LLMs	LLaMa	6.67	0.61	1.30	1.67	0.96	2.25
	EDIT(LLaMa)	8.33	0.65	1.45	1.00	0.54	3.22
	LLaMa2	1.67	1.24	5.07	1.67	1.77	14.13
	EDIT(LLaMa2)	3.33	0.84	5.56	6.67	1.20	11.96
	ChatGLM	50.00	3.46	50.38	53.33	3.43	60.91
	EDIT(ChatGLM)	70.00	3.85	57.36	61.67	3.93	60.19
	FLAN-T5-XL	12.00	1.96	15.15	23.33	2.09	19.17
	EDIT(FLAN-T5-XL)	24.00	1.98	16.59	38.33	2.34	17.56
OpenAI-Series LLMs	Text-Davinci-001	33.00	1.74	21.06	67.00	1.85	30.37
	Text-Davinci-002	43.00	2.20	25.20	56.50	2.18	32.01
	Text-Davinci-003	45.00	2.31	23.36	55.50	2.21	32.85
	Gpt-3.5-Turbo	72.50	3.72	53.62	86.19	3.51	63.02
	EDIT(Gpt-3.5-Turbo)	**93.50**	**4.25**	**62.57**	**87.62**	**4.11**	**64.66**

Table 2: The response results on downstream task. We report the Human Evaluation metric (Human), Information Entropy (Information Content) and the CC (Context Coherence).

EDIT framework can provide more information content, meeting the needs of users.

Improvement of Information in Response. Notably, for the IC metric, all LLMs have shown improvements after incorporating EDIT, indicating that the additional knowledge generated by EDIT can effectively enhance the information content in the responses. However, the FLAN-T5 model, which is not specifically designed for dialogue, does not show significant improvement in the IC metric, suggesting that it cannot directly apply the additional knowledge to response generation effectively without fine-tuning.

Semantic coherence with Dialogue Context. Almost all LLMs show a certain increase in Context Coherence on both datasets when combined with EDIT. This indicates that the extra knowledge integrated through EDIT not only enhances the information content but also ensures the coherence between the generated responses and the context. It demonstrates that our EDIT framework not only ensures the semantic coherence between the generated responses and the context, but also provides users with more useful information, rather than simply using contextually irrelevant information.

The Performance of ChatGPT. Furthermore, ChatGPT and EDIT(ChatGPT) exhibit the best performance. On the Wizard of Wikipedia dataset, EDIT(ChatGPT) achieves an IC score of 4.25 and a CC of 62.57%. On the Holl-E dataset, it achieves an IC score of 4.11 and a CC of 64.66%. The combination of ChatGPT and the EDIT framework shows the most significant performance improvement on both datasets, with IC scores increasing by more than 0.5 and CC increasing by over 8% on the Wizard of Wikipedia dataset. The results demonstrate that on LLMs with stronger generation capabilities, the additional knowledge provided by EDIT can better improve model performance.

3.5 Ablation Study

Dataset	LLMs	HUMAN	IC	CC
WoW	ChatGPT	72.50	3.72	53.62
	EDIT(ChatGPT)	**93.50**	**4.25**	**62.57**
	-w/o KB's Ans	91.50	4.23	62.48
	-w/o CG's Ans	82.38	4.17	61.21
Holl-E	ChatGPT	86.19	3.51	63.02
	EDIT(ChatGPT)	**87.62**	**4.11**	**64.66**
	-w/o KB's Ans	84.00	4.08	64.26
	-w/o CG's Ans	74.29	4.05	62.91

Table 3: The human evaluation result of ablation study for EDIT(ChatGPT).

To assess the performance of our Question Answering Module of EDIT(ChatGPT), we deleted the answers to questions provided by ChatGPT (referred to as CG's Ans) and the answers obtained from a Domain-Specific Knowledge Base(referred to as KB's Ans) on both datasets.

The ablation experiments are shown in Table 3. On Wizard of Wikipedia dataset and Holl-E dataset, after removing KB's Ans and CG's Ans, EDIT(ChatGPT) performance experienced a dropping movement which demonstrates the effectiveness of these two kinds of answers. Besides, the experimental results prove that our two Question Answer methods can improve the performance of EDIT. In particular, adding any type of answer alone can significantly improve the performance of the model compared with ChatGPT.

3.6 Discussion on Question Generation

It is important to identify the users' implicit intentions and use relevant knowledge when generating dialogue responses. In our approach, we train a model to generate context-related open questions as uers' intention. To ensure that the generated questions are useful, we evaluated these generated questions.

3.6.1 The Effect of Different Backbones in Our Question Generator

We use some generative models, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2019), as the base model for training the question generator (QG). Then we use BLUE and ROUGE-L to assess the generated questions. Table 4 presents the results, indicating that the QG based on different generative models all perform well.

	BLEU-1	BLEU-2	ROUGE-L
BART-Base	42.24	16.98	13.17
BART-Large	37.08	13.05	11.94
T5-Base	60.34	35.31	13.44
T5-Large	**60.72**	**35.51**	**14.00**

Table 4: The result of different backbones in question generator (QG).

Specifically, the T5 series models exhibit a BLEU-1 score of over 60%, which is twice as high as the BART series models. Furthermore, the ROUGE-L of T5-Large surpasses that of BART-Large. Finally, we apply the question generator trained by T5-Large to our EDIT framework.

3.6.2 Comparison with LLMs as Question Generator

Human Evaluation on Question Quality. To evaluate the performance of our question generator more reasonably, we conduct a human evaluation and compare our question generator with other LLMs such as Davinci, Text-Davinci-003, and Gpt-3.5-Turbo. Specifically, we design a prompt "Please generate 5 questions for this context, ensuring that the questions do not involve subjective judgments and focus on well-known objective facts." to make LLMs generate questions based on context.

Model	Human Evaluation
Davinci	0.20
Text-Davinci-003	0.35
Gpt-3.5-Turbo	0.52
Ours Question Generator	**0.94**

Table 5: The Human Evaluation result of the Question Generation on Context-Open-Question dataset.

We select 20 samples and ask each model to generate 100 questions for human evaluation. The results of the evaluation are presented in Table5. It is observed that the question generator achieves an impressive accuracy rate of 94% for the generated questions, while the other LLMs generally scored lower than 52%.

These findings demonstrate that our question generator possesses the ability to design effective questions and assist users in obtaining valuable information.

Performance on Downstream Tasks. On the downstream tasks, we separately have both the question generator and ChatGPT generate questions based on context, applying these within the EDIT. For a more direct comparison, we don't incorporate the answers obtained from a Domain-Specific Knowledge Base.

	Model	IC	CC
WoW	QG	4.06	54.32
	ChatGPT	4.23	62.48
Holl-E	QG	3.82	65.40
	ChatGPT	4.08	64.26

Table 6: The performance of EDIT(ChtaGPT) using questions generated by our QG and ChatGPT.

As shown in the table 6, applying questions generated by our question generator makes the responses contain more information and more contextually semantically consistent. Since the subject matter in the Holl-E dataset is relatively straightforward and confined to movies, the responses produced by both methods do not exhibit significant differences in terms of semantic coherence.

3.7 Discussion on Our Designed Metrics

To validate the ability of the IC and CC metrics, we randomly change or delete varying proportions of knowledge generated by the Question Answer Module and test EDIT on Wizard of Wikipedia task.

During the process of randomly changing knowledge, while LLMs receive extra knowledge which augments the information content of generated responses, the semantic coherence between generated response and the context sharply decreases due to the introduction of irrelevant knowledge. In contrast, in the course of randomly deleting knowledge, LLMs are provided with less supplementary input, thus leading to a decrease in the informational content of the generated response. But, the inherent ability of the LLM makes it generate contextually

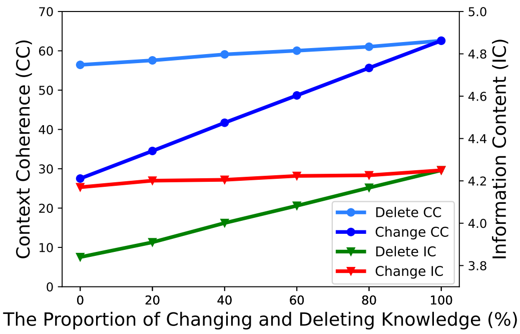

Figure 3: Variation of CC and IC with changing and deleting knowledge under different ratio.

coherent responses. As shown in Figure 3, our designed metrics, IC and CC, effectively capture the variations in the amount of information within the responses and the degree of semantic coherence between response and context. This demonstrates the rationality of our designed metrics.

4 Related Work

LLMs Application. As LLMs, such as ChatGPT, have demonstrated exceptional performance in various natural language processing (NLP) tasks, there has been a growing interest in exploring their applications. These models have found direct applications in several domains, including Healthcare(Tang et al., 2023; Nov et al., 2023; Yang et al., 2023), Education(Malinka et al., 2023; Tan et al., 2023; Kamalov and Gurrib, 2023), Finance(Wu et al., 2023) and Scientific research(Zhao et al., 2023). Besides, there are also some works to realize the application of LLM on specific tasks through Agent. The study by (Ruan et al., 2023) implements the assessment of LLMs' ability to perform reasoning processes through an Agent. Meanwhile, (Yao et al., 2023) focuses on optimizing the reasoning and planning capabilities of LLMs using reinforcement learning techniques. Additionally, (Chen et al., 2023) harnesses the story generation capabilities of LLMs to the realm of imaginative play. Furthermore, various practical applications leverage LLMs in real-life scenarios, including a series of plug-ins (Zhao et al., 2023) designed for Copilot and OpenAI. While these applications can successfully address specific tasks through agents, they still grapple with the challenge that LLMs may ignore users' implicit intentions and are unlikely to cover all fields. In light of these considerations, our work introduces the EDIT framework, aimed at asking questions to detect users' potential intentions and use related knowledge to generate responses with more information.

Question Generation. Before the advent of LLMs, context information was primarily used as a supplementary source for task-related questions. In most cases, answers to task-related questions can either be directly or indirectly inferred through context reasoning. As a result, question generation task usually generate questions based on a given context. Based on the type of input context, question generation tasks can be classified into several categories. There are document-level context input problems (Pan et al., 2020), paragraph-level context input (Zhao et al., 2018), and sentence-level context input (Gao et al., 2019) question generation tasks. Answering questions can be helpful in the question generation process. In the question generation task, the answer to the question is usually a fragment of the contextual text (Rajpurkar et al., 2016). Most question generation (QG) tasks are used as a supplement to the question answering (QA) task. The generated questions usually rely on information present in the given context and do not involve mining information from external sources. However, users' implicit intentions may be some questions beyond context, and it is important to consider mining information beyond the given context. In our work, we have developed a Context-Open-Question dataset that focuses on generating questions about out-of-context information. This dataset is used to train the question generator for generating questions as users' implicit intention.

5 Conclusion

In this study, we present an EDIT framework for the dialogue generation task to address the issue that LLMs may generate plain responses. Firstly, EDIT uses a question generator to generate a variety of questions related to dialogue context as the users' implicit intentions. Then EDIT interacts with LLMs and retrieves in Domain-Specific Knowledge Base to get answers respectively. Finally, EDIT uses the extra knowledge to enhance the response generated by LLMs. To evaluate the effectiveness of our approach, we design Information Content (IC) and Context Coherence (CC) to access the generated response. Our experiments show that combining EDIT with various mainstream LLMs can enhance their capabilities of generating responses with more useful information. This demonstrates the ability of EDIT.

Limitations

Due to limitations in computational resources, we did not test the performance of our proposed EDIT framework on a broader range of larger LLMs. Moreover, the metrics we designed do not account for cases where the model directly copies context content as a new generated response.

Ethics Statement

In the stage of human evaluation, we employed three graduate students experienced in natural language processing for human evaluation. We paid the graduate students about $8 per hour, well above the local average wage, and engaged in constructive discussions if they had concerns about the process.

Despite our control over the quality of the dataset, there remains a risk that our question generator might produce offensive questions, which could in turn prompt LLMs to generate potentially aggressive responses. Therefore, in future work, we need to devise methods to prevent the question generator from producing biased questions.

References

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

Zexin Chen, Eric Zhou, Kenneth Eaton, Xiangyu Peng, and Mark O. Riedl. 2023. Ambient adventures: Teaching chatgpt on developing complex stories. *CoRR*, abs/2308.01734.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Yifan Gao, Lidong Bing, Wang Chen, Michael R. Lyu, and Irwin King. 2019. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4968–4974. ijcai.org.

Firuz Kamalov and Ikhlaas Gurrib. 2023. A new era of artificial intelligence in education: A multifaceted revolution. *CoRR*, abs/2305.18303.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Kamil Malinka, Martin Peresíni, Anton Firc, Ondrej Hujnak, and Filip Janus. 2023. On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree? In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2023, Turku, Finland, July 7-12, 2023*, pages 47–53. ACM.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2322–2332. Association for Computational Linguistics.

Oded Nov, Nina Singh, and Devin M. Mann. 2023. Putting chatgpt's medical advice to the (turing) test. *CoRR*, abs/2301.10035.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 1463–1475. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 5042–5052. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Guoqing Du, Shiwei Shi, Hangyu Mao, Xingyu Zeng, and Rui Zhao. 2023. TPTU: task planning and tool usage of large language model-based AI agents. *CoRR*, abs/2308.03427.

Kehui Tan, Tianqi Pang, and Chenyou Fan. 2023. Towards applying powerful large AI models in classroom teaching: Opportunities, challenges and prospects. *CoRR*, abs/2305.03433.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *CoRR*, abs/2303.04360.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023a. Rethinking the evaluation for conversational recommendation in the era of large language models. *CoRR*, abs/2305.13112.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023b. Is chatgpt a good sentiment analyzer? A preliminary study. *CoRR*, abs/2304.04339.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *CoRR*, abs/2304.03347.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. *CoRR*, abs/2308.02151.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3901–3910. Association for Computational Linguistics.

A Case Study

We observe that using the EDIT framework can help LLMs respond better. As shown in Table 4, depending on the context, the Question Generation Module of the EDIT can provide questions to cater users' intentions. Compared with ChatGPT, the responses generated by EDIT(ChatGPT) are not limited to the improvement of writing skills. Specifically, EDIT(ChatGPT) could provide more options about the benefits of reading, such as 'gain insight' (blue), 'enhances critical thinking and analytical

Context

PersonA: I love reading! It's a means of sharing information and ideas

PersonB: Reading is one of my favorite ways to spend my time. My favorite book series is Harry Potter by J.K. Rowling.

PersonA: Many people love that series! Reading requires continuous practice development and refinement

PersonB: So reading can help widen your vocabulary? Are there any other benefits to reading?

Questions Generated by Question Generation Model

Q1: What are some other ways to spend time besides reading?

Q2: How does reading help develop specific skills?

Q3: Are there any specific genres of books that are particularly popular among readers?

Q4: What are some benefits of reading as a means of sharing information and ideas?

Q5: How does reading differ from other forms of entertainment, such as movies or television?

Responese of ChatGPT

PersonA: Absolutely! Reading not only expands your vocabulary, but it also enhances your communication skills. It allows you to explore different writing styles and perspectives, which in turn improves your own writing abilities. Additionally, reading can boost your empathy and understanding of others, as you delve into different characters' experiences and emotions. It's truly a gateway to knowledge, imagination, and personal growth.

Responese of EDIT(ChatGPT)

PersonA: Absolutely! Reading is a wonderful way to expand our knowledge and understanding of the world. It allows us to explore different perspectives and gain insights into various subjects. Additionally, reading enhances critical thinking and analytical skills, as it requires us to ana lyze and interpret information. It also stimulates our creativity and imagination, allowing us to envision and create vivid worlds and characters in our minds. Moreover, reading promotes empathy and understanding by exposing us to different cultures, experiences, and viewpoints. Lastly, reading helps improve our communication and language skills, as it exposes us to a wide range of vocabulary and writing styles. Overall, reading is a valuable means of sharing information and ideas that offers numerous benefits to individuals.

Figure 4: The examples from ChatGPT and EDIT(ChatGPT) in Wizard of Wikipedia dataset. The different highlight words represent options related to the benefits of reading.

skills' (green), 'stimulates creativity and imagination' (yellow) and 'promote empathy' (grey), which may cater to potential users' implicit intentions. While the response produced by ChatGPT only focuses on communication skills and ignores those potential users' intentions.