USIIR at TREC 2025 iKAT Track

Lili Lu
Università della Svizzera italiana (USI)
Switzerland
lili.lu@usi.ch

Abstract

This year's TREC iKAT track contains several tasks such as passage ranking, response generation and Personal Text Knowledge Base (PTKB) statement classification. We focus on response generation (only) task due to time and budget limitations. This task is to generate a response based on retrieved passages, given the additional context. We submitted two runs for this task mainly to explore the impact of user profiles on the generation quality of responses. In this short report, we describe the method that was used for generation and present the results.

1 Introduction

In the scenario of conversational information-seeking, a system needs to interact with a user through natural language in order to satisfy the user's information need [7]. In other words, the system can provide natural and coherent responses to the user during the conversation, given the context. With the advancements of Large Language Models (LLMs), we can rely on them for response generation. However, generating generic responses to users is ineffective, as each user has a different background. Hence, an effective conversational information-seeking system aims to provide personalized responses, while considering different user profiles. To facilitate the development of personalized conversational information-seeking systems, TREC iKAT track [1, 2, 3] was proposed along with the released datasets. More specifically, the data includes conversations between a system and a user, along with the Personal Text Knowledge Base (PTKB) that describes user profiles. Hence, different personalized conversational information-seeking systems can be evaluated on the datasets.
In this year's track, it includes several different tasks such as passage ranking, response generation and PTKB statement classification. Due to time and budget limitations, we focus on response generation (only) task. More specifically, every response should be generated by retrieved passages, given the user query, conversation history and user profiles. Additionally, this year's data includes 17 topics and 188 turns of conversations. In the following sections, we present the response generation approach and show the results that we achieved.

2 Method

For the task of response generation, we treated it as a summarization task conditioned on the given passages. Moreover, we instructed open-source LLMs with zero-shot learning to generate the response for each turn. More specifically, we used Qwen3-8B [9] for generation and HuggingFace[1] for implementation, following the official examples from the documentation.
Prior studies [5, 6] explored the impact of user profiles on query reformulation, indicating that the utilization of statements from PTKB is not always beneficial for retrieval performance, when all the queries are rewritten based on PTKB. Here, we aim to explore the impact of user profiles

[1] https://huggingface.co/Qwen/Qwen3-8B

Instruction: Your task is to summarize the following passages in no more than 50 words.
- Read the passages thoroughly.
- Generate a natural sentence.
Passages:
{pas}

Figure 1: The prompt for usiir_run1.

Instruction: Your task is to summarize the following passages in no more than 50 words, incorporating the provided additional information to enhance context.
- Read the passages thoroughly.
- Generate a natural sentence by incorporating the provided additional information.
Passages:
{pas}
Additional information:
{ad}

Figure 2: The prompt for usiir_run2.

on response generation after achieving relevant passages. Hence, we submitted the following two runs in this year's TREC iKAT track:

1. usiir_run1. For this run, we only considered the top three retrieved passages for generation and Figure 1 illustrates the prompt used for generation.
2. usiir_run2. For this run, we further considered user profiles and the top three retrieved passages for generation. Figure 2 shows the prompt for this run.

3 Results

We reported metrics of Rouge [4] and F1 on the official evaluation data. Table 1 illustrates the results.

Table 1: Results for our submitted two runs.

Runs	$F1$	ROUGE-1	ROUGE-2	ROUGE-L
usiir_run1	0.1916	0.1740	0.0235	0.1564
usiir_run2	0.1877	0.1708	0.0218	0.1534

As we can see from Table 1, usiir_run2 did not perform better than usiir_run1, implying that using all statements from PTKB introduces noises into response generation, resulting in slightly poor generation quality. Moreover, it may also indicate that not every response needs information from the PTKB, reflecting the similar finding on query reformulation [6].

4 Conclusions and Future Work

In this short report, we describe the method for response generation. Although our results are not strong based on the organizer's evaluation, our method is budget-friendly since we only relied on open-source models. We believe that the performance can be further improved by using state-of-the-art LLMs, such as GPT-5.4[2]. Moreover, we consider several possible directions to improve performance in future work. Here are just a few:

[2] https://developers.openai.com/api/docs/models/gpt-5.4

In-context learning. In the task, we only utilized LLMs with zero-shot learning. In the future, it is potential to leverage in-context learning to generate high-quality responses since LLMs may implicitly learn the generation patterns from the selected examples. Moreover, chain-of-thought [8] prompting can also be considered to leverage the reasoning capabilities of LLMs, thereby generating more coherent and logical responses.
Fine-tuning a generator. Although LLMs have strong generative abilities, they are still less controllable. Hence, it is potential to fine-tune a generator for responses to achieve more controllability. More specifically, we can leverage previous years' iKAT data as training data to fine-tune a generator. If the training data is not enough, the combination of synthetic data with previous years' data is a possible solution.
PTKB usage strategy for response generation. Based on our results, we found that using all the statements in PTKB for response generation achieves inferior performance. A possible solution for improving performance is to use the reasoning abilities of LLMs to automatically decide whether and how to use PTKB.
Although TREC tracks are typically performed by teams, this was a *solo* attempt. Consequently, the outcome reflected the limited effort that could be devoted to it. However, it was a very valuable experience that contributed to the author's doctoral work. In that sense, it was certainly beneficial.

References

[1] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffery Dalton, and Leif Azzopardi. Trec ikat 2023: The interactive knowledge assistance track overview. *arXiv preprint arXiv:2401.01330*, 2024.

[2] Mohammad Aliannejadi, Zahra Abbasiantaeb, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. Trec ikat 2023: A test collection for evaluating conversational and interactive knowledge assistants. In *SIGIR*, pages 819–829, 2024.

[3] Mohammad Aliannejadi, Zahra Abbasiantaeb, Simon Lupart, Shubham Chatterjee, Jeffrey Dalton, and Leif Azzopardi. Trec ikat 2024: The interactive knowledge assistance track overview. In *Text Retrieval Conference*, 2025.

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

[5] Fengran Mo, Yuchen Hui, Yuxing Tian, Zhaoxuan Tan, Chuan Meng, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. Adaptive personalized conversational information retrieval. *arXiv preprint arXiv:2508.08634*, 2025.

[6] Fengran Mo, Longxiang Zhao, Kaiyu Huang, Yue Dong, Degen Huang, and Jian-Yun Nie. How to leverage personal textual knowledge for personalized conversational information retrieval. In *CIKM*, pages 3954–3958, 2024.

[7] Paul Owoicho, Ivan Sekulić, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *SIGIR*, pages 632–642, 2023.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.