IIUoT at TREC 2024 Interactive Knowledge Assistance Track

YATING ZHANG, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan

HAITAO YU, Institute of Library, Information and Media Science, University of Tsukuba, Japan

In conversational information-seeking (CIS), the ability to tailor responses to individual user contexts is essential for enhancing relevance and accuracy. The TREC Interactive Knowledge Assistance Track addresses this need by advancing research in personalized conversational agents that adapt dynamically to user-specific details and preferences. Our study aligns with this framework, which involves three core tasks: *personal textual knowledge base (PTKB) statement ranking*, *passage ranking*, and *response generation*. To address these tasks, we propose a comprehensive framework that incorporates user context at each stage. For PTKB statement ranking, we integrate embedding models with large language models (LLMs) to optimize relevance-based ranking precision, allowing for more nuanced alignment of user characteristics with retrieved information. In the passage ranking stage, our adaptive retrieval strategy combines BM25 with iterative contextual refinement, enhancing the relevance and accuracy of retrieved passages. Finally, our response generation module leverages a Retrieval-Augmented Generation (RAG) model that dynamically synthesizes user-specific context and external knowledge, producing responses that are both precise and contextually relevant. Experimental results demonstrate that our framework effectively addresses the complexities of personalized CIS, achieving notable improvements over traditional static retrieval methods.

1 INTRODUCTION

The TREC Interactive Knowledge Assistance Track (iKAT) builds on the success of previous tracks, particularly the Conversational Assistance Track (CAsT) [2], to advance research in conversational information-seeking (CIS). iKAT encourages the development of personalized conversational agents that adapt dynamically to the unique needs and preferences of individual users, transforming static information retrieval into adaptive, multi-turn dialogue systems. Unlike general conversational search models, iKAT emphasizes a personalized approach by equipping systems with user-specific inputs to enhance response relevance and accuracy. This track provides a structured interaction context, where each query is shaped by user details stored in a PTKB, which consists of factual statements that describe the user's specific characteristics and preferences, along with previous conversational exchanges and the user's current query. However, conventional CIS methods typically rely on static retrieval mechanisms, which, while effective for broad information retrieval tasks, often fall short in integrating user-specific contexts or adapting to the dynamic nuances of real-time queries. This limitation can impede response relevance in CIS applications, where user intents and contextual dependencies are diverse. For instance, a query on "diet alternatives" may involve distinct requirements, such as vegan preferences or medical dietary restrictions, each necessitating a tailored response strategy. Traditional systems may lack the flexibility to discern and address these subtleties, underscoring the need for an advanced approach that fully incorporates user context and customizes responses dynamically.

Authors' addresses: YATING ZHANG, s2426078@u.tsukuba.ac.jp, Graduate School of Comprehensive Human Sciences, University of Tsukuba, Japan; HAITAO YU, yuhaitao@slis.tsukuba.ac.jp, Institute of Library, Information and Media Science, University of Tsukuba, Japan.

This work introduces a structured three-stage pipeline tailored to the primary components of personalized conversational search for the TREC 2024 iKAT task. This approach is organized into the following stages: *(i) Query Rewriting*, *(ii) PTKB Statement Ranking*, and *(iii) Passage Ranking and Response Generation*. The following sections provide an in-depth description of each stage in the framework.

2 METHODOLOGY

This section describes the methodologies applied to the main components of our approach, as illustrated in Figure 1. The following subsections provide detailed explanations of the key processes within our framework.

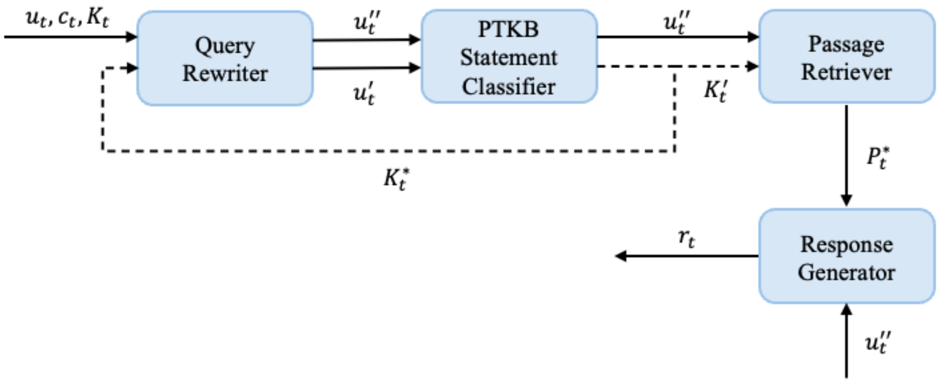

Fig. 1. Our iKAT submission overall framework.

2.1 Query Rewriting

The query rewriter module is responsible for refining the initial user query u_t by integrating contextual information and relevant entries from the PTKB. This refinement process aims to improve the accuracy of subsequent retrieval and classification by aligning the query in closer conformity with relevant information within the PTKB. The module takes three inputs: the user's current utterance u_t, the contextual information c_t, and the PTKB statements K_t. The primary function of the query rewriter is to transform these inputs into an enriched query representation that reflects both the user's intent and the surrounding conversational context. This transformation is particularly beneficial when the original query lacks clarity or specificity, as it helps reduce ambiguity and better aligns the query with the content in the knowledge base. To achieve this, the Query Rewriter employs a fine-tuned Llama3 model [3], specifically trained to rephrase queries by using information from up to three preceding conversational turns, as well as selected PTKB entries. The initial rewriting phase generates u_t'„ a refined query that more explicitly incorporates contextual and relevant PTKB statements. This refined query u_t' is then passed to the PTKB Statement Classifier module, which performs an initial selection of PTKB entries by identifying those classified as relevant. This produces a subset K_t^* that represents PTKB statements preliminarily aligned with the refined query. These selected entries are then reintroduced into the Query Rewriter in an iterative feedback loop. The second rewriting phase, using K_t^*, produces a further refined query representation, denoted u_t'', which integrates both the initial contextual information and the filtered PTKB content for greater precision.

2.2 PTKB Statement Ranking

In our PTKB statement ranking process, we employ a multi-stage approach to progressively refine the relevance of selected statements. The process begins with the SBERT model [6], which uses the refined query u_t'' as input to conduct an initial classification of PTKB entries. SBERT identifies a preliminary subset, denoted K_t^* , by evaluating the semantic alignment of statements with u_t''. This initial step functions as a coarse filtering mechanism, eliminating less relevant statements while retaining those that demonstrate a general alignment with the query context.

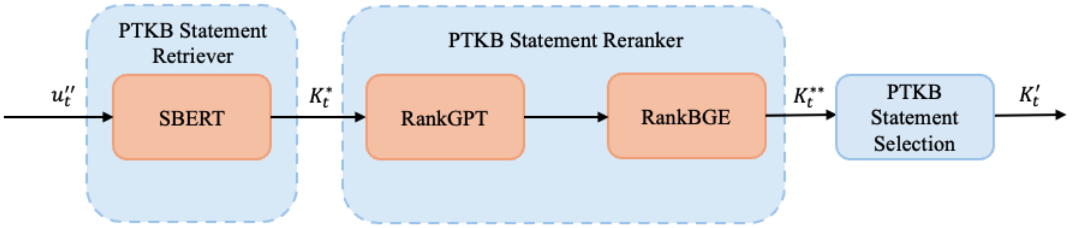

Fig. 2. PTKB Statement Ranking Pipeline.

Subsequently, the system transitions into a reranking phase, which involves two sequential models, RankGPT and RankBGE, each providing a more granular assessment of relevance within the subset K_t^*. Initially, RankGPT, which leverages the GPT-4o model [4], performs a detailed semantic analysis of each entry within K_t^* to assess its contextual relevance to u_t'' and assigns individual relevance scores. This stage refines the preliminary ranking by capturing subtler aspects of semantic alignment. Following this, the RankBGE model further processes the outputs from RankGPT. RankBGE, a fine-tuned version of the BGE (Gemma-2b) model [7], applies a weighted scoring function that integrates the scores from RankGPT along with its own assessments. The combined relevance score S_{rel} is calculated as follows:

$$S_{\text{rel}} = \alpha \cdot P(\text{MiniLM}) + (1 - \alpha) \cdot P(\text{BGE}) \tag{1}$$

where α is a hyperparameter optimized through cross-validation to balance the influence of each model's relevance score. This composite score, S_{rel}, synthesizes insights from both models to provide a robust estimation of relevance. In the final decision-making stage, an indicator function L(s) classifies each PTKB statement based on S_{rel}, a dynamically adjusted threshold θ_{dyn}, and a voting mechanism. Specifically, a PTKB statement is classified as relevant if $S_{\text{rel}} > \theta_{\text{dyn}}$ and at least two of the models classify the entry as relevant. The indicator function is defined as:

$$\mathrm{L(s)} = \begin{cases} 1 & \text{if } S_{\text{rel}} > \theta_{\text{dyn}} \text{ and } \sum_{i=1}^{n} v_i \geq 2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\sum_{i=1}^{n} v_i$ represents the number of positive votes from the models. By combining score-based filtering with model consensus, this approach applies rigorous selection criteria. The outcome of this multi-stage process is a final refined set of PTKB statements, denoted K_t', which consists of entries classified as relevant and arranged in descending order of their relevance scores. This ranking ensures that statements most aligned with the refined intent of the query are prioritized, thereby facilitating more effective downstream retrieval and response generation by providing a highly relevant and organized set of statements for further processing.

2.3 Passage Ranking

This section describes our methodology for passage ranking and response generation, as shown in Figure 3, which leverages a retrieval-augmented generation (RAG) framework configured with Llama RAG. This configuration utilizes the dynamic retrieval capabilities of RAG models, making it appropriate for complex information retrieval tasks that require iterative refinement and context-aware response generation.

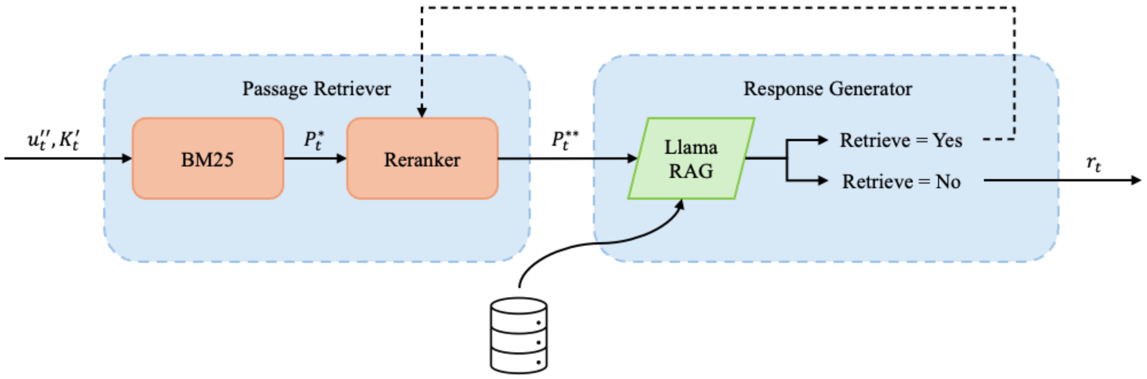

Fig. 3. Our Response Generation Pipeline.

The process begins by inputting a user query u_t' along with relevant contextual information K_t'. In the initial retrieval phase, the BM25 algorithm, implemented through Pyserini, retrieves an initial set of 1,000 candidate passages (P_t^*) from a comprehensive document corpus. This broad retrieval step establishes a diverse set of semantically relevant passages, forming a foundational evidence base for subsequent refinement. In the next stage, the BGE-base-en-v1.5 model [5] is applied to re-rank these candidate passages based on embedding-based semantic similarity, refining the selection to identify the top 5 passages (P_t^{**}) most closely aligned with the intent of the query. Additionally, the re-ranking process incorporates the top 3 contextual statements from the PTKB, thereby augmenting the selected passages with user-specific contextual information.

Upon completing the re-ranking phase, the Self-RAG model [1], configured as *"selfrag/selfrag-llama2-7b,"* progresses to the final stages of retrieval validation and response generation. Unlike traditional retrieval models, Self-RAG integrates an adaptive retrieval mechanism that continuously assesses the sufficiency of information within the selected passages (P_t^{**}). During response generation, the model uses "reflection tokens" to evaluate the need for additional information, adjusting the retrieval scope dynamically in response to the complexity and specificity of the query. When further detail is required, the model accesses an expanded set of resources, which includes the initial collection of 1,000 candidate passages as well as comprehensive external corpora, such as an extensive Wikipedia dataset. This adaptive retrieval mechanism enables flexible adjustments to the information scope, accommodating diverse informational needs and enhancing the relevance of responses. The integration of Wikipedia [8] as an additional knowledge base provides access to contextual information that may complement the initial retrieval set. Through this structured process of retrieval, re-ranking, and adaptive response generation, as illustrated in Figure 3, the approach supports the handling of advanced information retrieval tasks that require both contextual sensitivity and information depth.

2.4 Implementation Details

Query rewriting was conducted using a Llama3 model fine-tuned over 10 epochs, employing a learning rate of 4.0e-4 and a warmup ratio of 0.05. In the PTKB statement ranking task, the performance of the BGE reranker was enhanced through fine-tuning and subsequent integration with the Gemma-2b model. This integration leveraged the *merge_llm* function, which combined outputs from both the bge-reranker-v2-gemma and the pre-trained gemma-2b models. Hyperparameters for fine-tuning the BGE reranker included a learning rate of 2.0e-4, 5 training epochs, and a batch size of 4 per device. To optimize computational efficiency, LoRA (Low-Rank Adaptation) was applied with a rank of 32 and an alpha of 64, and flash attention was used to accelerate the training process. The maximum input sequence length for both queries and passages was set to 512 tokens, and gradient checkpointing was activated to minimize memory usage. Model weights were dynamically adjusted through five-fold cross-validation to balance the contributions of each model in the scoring mechanism. The Self-RAG model was configured for passage ranking and response generation, with a temperature set to 0 and a top-p value of 1 to ensure deterministic outputs. The retrieval depth parameter, max_depth, was set to 2 to achieve a balance between iterative retrieval and response relevance. This configuration was designed to support precise and contextually relevant responses, effectively managing the trade-off between retrieval scope and output accuracy.

3 RESULTS EVALUATION

In this section, we provide a comprehensive analysis of the proposed methods by evaluating their performance across multiple key metrics. The tables included summarize the results, presenting comparative statistics and benchmarks across a range of topics to assess the model's ranking accuracy and relevance in handling complex information retrieval tasks.

3.1 Passage Ranking Evaluation Results

The Passage Ranking task presented notable challenges, as the majority of the results for metrics such as *ndcg_cut_10*, *P_5*, and *recip_rank* fell below the median benchmarks, as summarized in Table 1. This indicates that while the retrieval and re-ranking components capture some relevant passages, they struggle to consistently distinguish between highly relevant and marginally relevant content, particularly under ambiguous or implicit query contexts.

One key factor influencing this performance could be the reliance on BM25 for initial retrieval. Although it effectively retrieves a broad set of passages, its lexical matching approach may fail to capture deeper semantic relationships between the query and the passages. While the BGE-base-en-v1.5 model provides re-ranking based on embeddings, its impact may be limited by the quality of the initial retrieval set. Additionally, integrating PTKB contextual statements into the ranking process introduces complexity, which may inadvertently amplify noise in some scenarios. However, isolated successes, such as higher P_5 scores for certain topics, suggest that the re-ranking mechanism is effective when the initial retrieval set contains highly relevant candidates. This points to opportunities for improvement, such as refining the initial retrieval process to better align with downstream re-ranking and leveraging external context more dynamically. Furthermore, incorporating a broader and more diverse training dataset could enhance the model's generalization, enabling it to handle a wider variety of queries and topics.

Pn	**ndcg_cut_10**	**recip_rank**	**P_5**
better_than_median	1	3	3
worse_than_median	107	101	101
equal_best	1	13	1
equal_worst	80	80	80

Table 1. Passage Ranking Statistics Results of Our Submitted Runs' Evaluation Results

3.2 PTKB Statement Ranking Evaluation Results

PTKB	**ndcg_cut_10**	**recip_rank**	**P_5**
better_equal_median	81	95	90
worse_equal_median	67	83	84
only_worse	33	19	24
equal_best	35	75	40
equal_worst	37	37	37

Table 2. Statistics Results of Our PTKB Statement Ranking Evaluation Results

Table 2 summarizes the performance of our PTKB statement ranking evaluation runs across various metrics. Our approach demonstrated strong results, with 81, 95, and 90 topics achieving *ndcg_cut_10*, *recip_rank*, and *P_5* scores better than or equal to the median, respectively. Furthermore, 75 topics reached the best score for *recip_rank*, highlighting the method's effectiveness in certain scenarios. However, areas for improvement are evident, as 33, 19, and 24 topics fell below the median exclusively in *ndcg_cut_10*, *recip_rank*, and *P_5*. Additionally, 37 topics across all metrics matched the worst scores, signaling challenges with specific query contexts. These results indicate that while our approach is robust overall, targeted refinements are necessary to address these weaker cases.

4 CONCLUSION

In this study, we introduced a structured and adaptive framework for personalized conversational information-seeking (CIS) within the TREC 2024 iKAT framework, enhancing response relevance and accuracy through user-specific contextual integration. Our methodology spanned query rewriting, PTKB statement ranking, and passage ranking with response generation, each dynamically tailored to meet individual user needs and preferences. By integrating advanced models—Llama3 for query rewriting, a multi-model approach for PTKB ranking, and the Self-RAG model for passage ranking—this framework effectively leverages both user context and external information. The evaluation results reveal strong performance on precision-focused tasks, particularly for the *P_5* and *recip_rank* metrics, where our framework consistently ranks relevant content among the top results for select topics. These findings highlight the framework's efficiency in prioritizing relevant information and maintaining responsiveness to user-specific nuances. However, the variation in *ndcg_cut_10* scores across topics suggests potential areas for optimization. While the approach effectively captures relevance in certain contexts, further fine-tuning could enhance its adaptability to a broader range of topic complexities and content variations. Our findings emphasize the importance of a tailored, context-aware approach in advancing CIS and the need for refined retrieval and ranking mechanisms to balance precision and relevance. Future work will focus on strengthening robustness through cross-domain adaptability, model parameter refinement, and additional methods to optimize user-specific context integration. We anticipate these improvements will expand the framework's applicability to real-world, multi-turn dialogue systems, advancing personalized CIS to meet diverse and dynamic user needs.

REFERENCES

[1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511* (2023).

[2] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[3] 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* (2024).

[4] Raisa Islam and Owana Marzia Moushi. 2024. GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. *Authorea Preprints* (2024).

[5] Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503* (2023).

[6] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[7] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).

[8] Wikipedia. [n. d.]. English Wikipedia dump, December 20, 2018. Internet Archive. Available: https://archive.org/download/enwiki-20181220/enwiki-20181220-pages-articles.xml.bz2. [Accessed: Oct. 31, 2024].