Evaluating Full Dialogue History vs. Summarized Context for Personalized Knowledge Assistance: Findings from the TREC 2025 iKAT Track

Suveyda Yeniterzi *
GenAIus Technologies
suveyda@genaius.tech

Reyyan Yeniterzi
GenAIus Technologies
reyyan@genaius.tech

February 15, 2026

Abstract

We present GenAIus's participation in the TREC 2025 Interactive Knowledge Assistance Track (iKAT), focusing on personalized and context-aware response generation in both offline and interactive settings. We develop a multi-stage pipeline that integrates conversation summarization, Personal Textual Knowledge Base (PTKB) classification, query rewriting, passage retrieval, and grounded response generation. To study the impact of conversational context modeling, we compare two configurations: conditioning on the full dialogue history versus using an evolving conversation summary updated at each turn. Experimental results show that full-history conditioning yields slightly stronger performance in offline generation and dialogue-level interactive metrics, while summary-based conditioning achieves comparable overall results with improvements in engagement and contextual efficiency. Both approaches rank within the top tier of participating systems, demonstrating the robustness of our pipeline and the viability of structured conversational summarization as a scalable alternative to full-history conditioning.

1 Introduction

The TREC Interactive Knowledge Assistance Track (iKAT) focuses on advancing research in collaborative information-seeking conversational agents that generate personalized, context-aware responses [1]. Unlike traditional conversational search settings that rely primarily on dialogue history, iKAT incorporates knowledge learned about the user. As a result, conversations evolve not only based on prior system responses but also on user attributes, preferences, and contextual signals. This more realistic interaction paradigm means that users with the same underlying topic may follow different conversational paths and require different forms of assistance. The core goal of the track is to evaluate how effectively systems retrieve relevant knowledge and generate helpful, personalized responses grounded in both conversational context and user profiles.

To support different research focuses and system capabilities, iKAT offers multiple task formats spanning both offline and interactive settings. In the offline passage ranking and response generation task, systems are given static conversational turns from the test collection and are required to retrieve and rank relevant passages for each user utterance before generating grounded responses. Responses must be supported by at least one passage provenance from the collection, typically implemented through Retrieval-Augmented Generation (RAG) pipelines. A second offline variant, response generation only, removes the retrieval component by providing pre-ranked passages, allowing participants to focus exclusively on response construction quality, grounding, and personalization.

Beyond offline evaluation, iKAT introduces an interactive submission setting that more closely mirrors real-world deployments. In the passage ranking and interactive response generation task, systems engage in real-time conversations with simulated users. For each user turn, the system retrieves and ranks passages, generates a personalized response, and returns it to the simulator, which then produces

*Both authors contributed to this paper equally.

the next utterance. This iterative loop continues until the conversation concludes, enabling evaluation of multi-turn adaptation, personalization consistency, and long-horizon conversational effectiveness.

During our participation in this track, we studied the role of conversational context representation in response generation by experimenting with two alternative conversation history modeling approaches. In the first setting, we used the full conversation history verbatim, providing the model with the complete sequence of prior user and system turns. In the second setting, we constructed evolving conversation summaries at each turn, updating the summary based on the previous history summary together with the latest user utterance and system response. This design allowed us to investigate the trade-offs between raw dialogue preservation and structured conversational memory. While full histories retain the complete contextual record, they introduce challenges related to context window constraints, conversational noise, and retrieval diffusion. In contrast, when performed effectively, summary-based histories compress prior exchanges into a concise, intent-focused representation that may improve retrieval alignment, grounding fidelity, and generation efficiency. Evaluating these two settings enables us to assess whether structured conversational summarization can preserve response quality.

The track utilizes a curated subset of the ClueWeb22-B [3] collection to make experimentation feasible while preserving domain diversity. The collection contains over 116 million passages. Due to limited computational resources, we were unable to process the full corpus for retrieval. Instead, we relied on the BM25 Pyserini index provided by the organizers for this subset as our retrieval resource. Consequently, we participated in the offline response generation–only task and relied on the provided retrieval index for passage access in the interactive setting.

2 Approach

Our approach is designed to generate personalized, context-aware responses by jointly leveraging conversational context, user-specific knowledge, and retrieved external evidence. To support this objective, we developed a multi-stage pipeline that progressively refines contextual signals and integrates them into response generation. The pipeline begins by modeling conversational memory through an evolving summary representation, enabling efficient context tracking across dialogue turns. We then incorporate user attributes via Personal Textual Knowledge Base (PTKB) classification to identify profile signals relevant to the current interaction. Next, the user's latest utterance is rewritten into a retrieval-optimized query that reflects both conversational and user-specific context. This query is used to retrieve supporting documents from the provided BM25 Pyserini index. Finally, the system generates a grounded and personalized response conditioned on the conversation context, selected PTKB attributes, and retrieved passages, while also providing transparent evidence attribution. Together, these components form an end-to-end framework for interactive knowledge assistance that balances contextual coherence, personalization, and evidence grounding.

2.1 Summarize Conversation History

To manage conversational context across turns, we implemented an LLM-based evolving conversation summarization mechanism, operationalized through the prompt presented in Figure 1. At each dialogue turn, the model was provided with the existing conversation summary together with the previous user utterance, the previous system response, and the current user utterance. Using this structured input, the model generated an updated summary that refined prior context while incorporating the most recent exchange. This design enables the summary to function as a rolling conversational memory rather than a static compression. The motivation for this approach is to preserve salient intent and decision context while reducing conversational noise and context length. By consolidating relevant information into a concise, coherent narrative, the summary can provide a more focused conditioning signal for downstream response generation compared to using the full dialogue history verbatim.

For the remainder of the pipeline, prompts are illustrated using the full conversation history setting for clarity and consistency. However, when operating under the conversation summary configuration, the prompts were applied in the same manner with a single modification: wherever conversation history was specified as input, it was replaced with the generated conversation summary. Aside from this substitution, all prompt instructions, contextual inputs, and generation procedures remained unchanged across the two settings.

```
You are an intelligent assistant trained to generate an evolving, concise, and
     coherent summary of a conversation for use in future turns of the same
    dialogue.

You will be provided with:
- Current conversation summary: A condensed version of the full sequence of
    prior user and system messages.
- Previous user utterance: The user's last message before the current turn.
- Previous system response: The assistant's reply to that message.
- Current user utterance: The user's most recent message in the ongoing
    conversation.

Instructions:
Your goal is to update and refine the conversation summary so it accurately
    reflects the latest exchange while retaining key context from earlier
    turns.

Requirements:
- Prioritize recency: Give greater weight to the previous turn and current
    user utterance, as they reflect the most immediate intent.
- Preserve important context: Keep any prior details that remain relevant for
    continuity in future turns.
- Be concise but complete: Avoid unnecessary detail, but ensure that important
     facts, requests, and decisions are included.
- Maintain coherence: The updated summary should read as a smooth narrative of
     the conversation so far, not a list of disjointed messages.
- No speculation: Use only information explicitly provided in the conversation
    .

Output format:
Return only the updated summary text, no bullet points, labels, or extra
    commentary.
```

Figure 1: Prompt used to generate the conversation history summary

2.2 Classify PTKB

To incorporate user-specific knowledge into response generation, we performed a PTKB classification step aligned with the personalization objectives of the iKAT track. The PTKB consists of structured textual statements describing the user's background, preferences, and prior behavioral signals. At each dialogue turn, we formulated PTKB usage as a binary relevance classification task, where the goal is to identify which user attributes are pertinent to the current interaction. As specified in Figure 2, the model is provided with the PTKB entries, the conversation history, and the current user utterance, with priority given to the most recent turn while still considering prior context. Guided by this prompt, the model selects the subset of PTKB indices that can help interpret the user's intent, personalize the response, or reflect relevant preferences. This classification step enables dynamic user modeling, allowing downstream response generation to be conditioned not only on conversational context but also on the most relevant user-specific knowledge at each turn.

2.3 Rewrite Utterance

To improve retrieval quality in the interactive setting, we introduced a query rewriting step that reformulates the user's latest utterance into a retrieval-optimized search query. As specified in Figure 3, the model is provided with the PTKB, the conversation history, and the current user utterance. The prompt instructs the model to generate a precise, semantically rich query that preserves the user's original intent while incorporating relevant contextual signals. When appropriate, user-specific attributes from the PTKB and salient details from the conversation history are integrated to enhance search relevance, while unrelated or extraneous information is excluded. This rewriting process produces

```
You are an intelligent assistant trained to personalize system responses based
    on user-specific background, preferences, and past interactions.
Your task is to identify which user profile attributes are relevant for
    generating the system's response at this turn.

You will be provided with:
- Personal Textual Knowledge Base (PTKB): A list of numbered statements
    describing the user's persona, background, and interests. Each entry
    begins with an index number (e.g., 1: ...).
- Conversation History: A sequence of previous user and assistant utterances.
- Current User Utterance: The user's most recent message.
While the conversation history offers helpful context, always prioritize the
    current user utterance, as it reflects the user's most immediate intent.

Task:
For each PTKB statement, determine whether it is relevant (True) to the
    current user utterance, considering the full conversation history. This is
    a binary classification task.

Relevance Criteria:
Mark a PTKB entry as relevant only if it helps:
- Interpret or clarify the user's current message or intent,
- Personalize or enrich the system's response,
- Reflect the user's preferences, needs, or behaviors in context.
It is valid for none of the PTKB statements to be relevant. In such cases,
    return an empty list.

Output Format:
Return a list of PTKB index numbers that are relevant for this turn.
Do not include the statement text, explanations, or any additional content.
```

Figure 2: Prompt used to classify PTKBs

a focused query representation that is better aligned with the underlying information need, thereby supporting more accurate document retrieval for downstream response generation.

2.4 Retrieve Documents

Following query rewriting, we performed document retrieval using the reformulated utterance as the search query. Specifically, the rewritten query was issued against the BM25 Pyserini index provided by the track organizers, which was built over the curated ClueWeb22-B subset used in the evaluation. For each dialogue turn, we retrieved the top 20 ranked documents, which served as the evidence pool for response generation. Using the rewritten query within the standardized BM25 retrieval framework ensured consistency with the track setup while enabling context-aware access to relevant knowledge sources grounded in both conversational and user-specific context.

2.5 Generate Response

In the final stage of the pipeline, we generated personalized system responses conditioned on conversational context, user attributes, and retrieved evidence. Following the prompt in Figure 4, the model was provided with the PTKB, the conversation history (either full history or the summary), the current user utterance, and the top-20 retrieved passages. The prompt guided the model to produce a concise, coherent, and user-tailored response that directly addressed the user's intent while incorporating relevant profile signals and evidence where appropriate. Retrieved passages were used selectively, with the model instructed to disregard any content that did not add meaningful value. To ensure grounded and context-faithful generation, the response was constrained to rely solely on the provided inputs, with explicit instructions to avoid hallucination or external knowledge. In addition to producing the response text, the model also identified which retrieved passages supported the generated content by

```
You are an intelligent assistant tasked with rewriting the user's latest
    utterance into a high-quality search query for retrieving relevant
    documents.

You will be provided with:
- Personal Textual Knowledge Base (PTKB): Statements describing the user's
    background, interests, and persona.
- Conversation History: All prior user and system messages.
- Current User Utterance: The user's most recent message.

Task:
- Reframe the query so that it is precise, semantically rich, and informed by
    the provided context.
- Incorporate relevant details from the PTKB and conversation summary if they
    directly enhance search relevance.
- Exclude unrelated or extraneous information.
- Use clear and natural language that aligns with the user's intent.

Guidelines:
- Preserve the original intent of the user's utterance.
- Enhance specificity using relevant context from the PTKB and conversation
    history.
- Avoid adding speculative or fabricated details.

Output Format:
Return only the rewritten query text following 'Query:'. Do not include
    explanations, labels, or extra formatting.
```

Figure 3: Prompt used to rewrite the last utterance

returning the corresponding rank IDs. This dual-output design enabled both personalized answer synthesis and transparent evidence attribution within the interactive assistance setting.

While the response generation framework is described in the context of the interactive pipeline, the same prompting and grounding principles were applied in the offline response generation–only setting. In the offline task, the model was provided with the PTKB, the conversational context (either full history or summary), the current user utterance, and the top-ranked passages supplied by the organizers rather than passages retrieved through our own query rewriting and retrieval steps. The generation prompt and grounding constraints remained identical, ensuring that responses were personalized, context-aware, and strictly supported by the provided evidence.

3 Submissions

We evaluated the impact of using the full conversation history versus a summarized version of the conversation across both offline and interactive tasks.

We used GPT-4o [2] as the underlying large language model across all stages of the pipeline, including conversation summarization, PTKB classification, query rewriting, and response generation.

3.1 Offline - Response Generation (only)

For the offline response generation task, we submitted the following two runs.

- **genaius-genonly-full-gpt4o**: Generate responses using GPT-4o. The prompt includes: (1) the provided top 5 passages as received, (2) the PTKB information, and (3) the full conversation history.
- **genaius-genonly-summary-gpt4o**: Generate responses using GPT-4o. The prompt includes: (1) the provided top 5 passages as they are, (2) the PTKB information, and (3) an LLM-generated summary of the conversation history.

```
You are an intelligent assistant trained to generate concise, relevant, and
    personalized responses tailored to each user.

You will be provided with:
- Personal Textual Knowledge Base (PTKB): Statements describing the user's
    background, interests, and persona.
- Conversation History: All prior user and system messages.
- Current User Utterance: The user's most recent message.
- Retrieved Passages: Passages retrieved by a retrieval model, each with a
    rank ID.

Task 1 - Generate the Response
Produce a coherent and engaging answer to the current user utterance by:
- Addressing the user's intent directly.
- Incorporating relevant PTKB details where appropriate.
- Using insights from retrieved passages only if clearly relevant.
- Disregard any retrieved passage that does not add meaningful value.
- Make sure the response is focused, informative, and personalized.
- Base your response solely on the provided PTKB, conversation history, and
    retrieved passages
- Do not hallucinate, do not use any outside knowledge. Use only the provided
    context.
- Response should be maximum 220 words.

Task 2 - Identify References
List the rank IDs of all retrieved passages that were actually used in Task 1'
    s response.
- Include only the rank IDs.
- Exclude any IDs for unused passages.
- Always ensure that at least one rank ID is returned - if no passages were
    directly used, return the rank ID of the passage most semantically similar
     to the generated response content.

Output Format:
- Return a JSON object with exactly two keys: 'Task 1' and 'Task 2'.
- 'Task 1': Contains only the assistant's response text as a string (no labels
    , metadata, explanations, or extra formatting).
- 'Task 2': Contains only the list of rank IDs used, as an array of integers.
- Do not include any additional text, headings, or phrases. Return only the
    JSON object.
```

Figure 4: Prompt used to generate a response

The official performance of these two runs is presented in Table 1. Overall, the full conversation history outperformed the conversation summary across most evaluation metrics, with the exception of ROUGE-1. Notably, under the LLM GPT-4.1 evaluation, both approaches achieved identical scores. These competitive results indicate that summary-based context representations hold promise as a viable alternative to using the full conversation history, though further investigation is required. One potential limitation of the summary-based approach is the loss of long-range contextual information: if a response depends on details from much earlier turns that are no longer preserved in the summary due to topic or context shifts, performance may degrade. This limitation likely contributes to the slightly lower scores observed on certain metrics.

In the official leaderboard, our runs ranked 3rd in terms of BEM, 4th on LLMeval (SOLAR), and 5th across all remaining evaluation metrics. These rankings demonstrate the overall competitiveness and robustness of our approach.

Run	LLMeval		Nugget	BEM	F1	ROUGE-1
	SOLAR	GPT-4.1	Recall			
genaius-genonly-full-gpt4o	0.8889	0.7556	0.0999	0.1672	0.2827	0.2485
genaius-genonly-summary-gpt4o	0.8667	0.7556	0.0811	0.1407	0.2750	0.2500

Table 1: Scores from the Offline Task

Run	Rubric Level					Dialogue Level							Score
	Eng	Rel	Qual	Conf	Score	Mix	Pers	Flow	Trust	Sat	Conf	Score	
full	0.71	0.80	0.82	0.94	0.78	0.46	0.59	0.93	0.72	0.82	0.94	0.78	0.77
summary	0.72	0.80	0.85	0.95	0.81	0.48	0.59	0.91	0.76	0.78	0.90	0.72	0.75

Table 2: Scores from the Interactive Task

3.2 Interactive - Passage Ranking and Response Generation

For the interactive response generation task, we again submitted two runs. Both runs share the following steps:

1. **Classify PTKBs**: Use GPT-4o to classify the PTKB statements. The prompt includes: (1) the PTKB information, (2) the full conversation history or the summary, and (3) the current user utterance.
2. **Rewrite utterance**: Use GPT-4o to rewrite the current user utterance. The prompt includes: (1) the selected PTKB information, (2) the full conversation history or the summary, and (3) the current user utterance.
3. **Retrieve documents**: Search the rewritten utterance against the provided BM25 Pyserini index, and retrieve the top 20 documents.
4. **Generate response**: Use GPT-4o to generate the final response. The prompt includes: (1) the selected PTKB information, (2) the full conversation history or the summary, (3) the current user utterance, and (4) the retrieved documents.

The two runs share the same retrieval, personalization, and response generation framework; the only difference lies in how conversational context is represented. Specifically, one configuration conditions generation on the full conversation history, while the other replaces the raw dialogue with an LLM-generated summary. This summary is produced at each turn using the previous conversation summary together with the previous user utterance, the previous system response, and the current user utterance. As a result, the comparison isolates the impact of conversational context representation on response quality while holding all other pipeline components constant.

Our submissions are:

- **genaius-full-rewrite**: The generation pipeline run with the full conversation history
- **genaius-summary-rewrite**: The generation pipeline run with the LLM-generated conversation summary

As shown in Table 2, across rubric-level evaluation, the two runs exhibit closely aligned performance. The summary-based run achieves a slightly higher rubric score (0.81) compared to the full-history run (0.78), driven primarily by marginal gains in engagement (0.72 vs. 0.71) and overall response quality (0.85 vs. 0.82). Relevance scores remain comparable across both configurations (0.80 for summary-based vs. 0.80 for full-history). Overall, these results indicate that compressing conversational context into an evolving summary can maintain, and in some dimensions slightly improve, response engagement and perceived quality without sacrificing relevance.

At the dialogue level, the performance pattern partially reverses. The full-history configuration attains a higher dialogue score (0.78) than the summary-based variant (0.72). This gap is primarily driven by modest differences in information flow (0.93 vs. 0.91) and user satisfaction (0.82 vs. 0.78), suggesting that access to the complete conversational record slightly improves end-to-end conversational continuity and perceived overall success of the interaction. At the same time, the summary-based

run achieves a slightly higher trustworthiness score (0.76 vs. 0.72), indicating that compressing the dialogue into a more focused representation can help maintain a consistent and credible response tone. Overall, both runs remain strong on dialogue-level criteria, with summarization largely preserving conversational coherence while incurring a small degradation in flow and satisfaction relative to the full-history setting.

When aggregated into the system-level score, computed as the harmonic mean of rubric and dialogue scores, the two runs remain highly competitive and closely ranked. The full-history run achieves an overall score of 0.77, while the summary-based run attains 0.75. The narrow gap between the two confirms that evolving conversation summaries can retain most of the response quality benefits of full dialogue conditioning while offering a more compact context representation.

Relative to the full set of participating systems, both GenAIus submissions rank within the top tier of the leaderboard. The full-history run places 2nd overall, while the summary-based run ranks 3rd, directly behind the leading system. This positioning demonstrates that GenAIus approaches perform competitively across both rubric-level response quality and dialogue-level interaction metrics. Notably, the strong placement of the summary-based run indicates that structured conversational memory can support high-quality, personalized response generation even under compressed context representations.

Overall, the results indicate that both conversational context modeling strategies are effective. Full-history conditioning provides modest advantages at the dialogue level, particularly in information flow and user satisfaction, reflecting the benefits of preserving the complete conversational record for end-to-end interaction quality. In contrast, summary-based conditioning achieves comparable overall performance, while demonstrating slight gains at the rubric level in engagement and perceived response quality. Together, these findings validate the robustness of the GenAIus interactive response generation pipeline across alternative context representation designs, showing that structured conversational summarization can preserve response effectiveness while improving contextual efficiency.

4 Conclusion

In this work, we introduced a multi-stage, personalization-aware response generation framework for the TREC 2025 iKAT Track. Our pipeline combines conversational memory modeling, user profile conditioning through PTKB classification, retrieval-optimized query rewriting, and evidence-grounded response synthesis within a unified LLM architecture. By evaluating both full conversation histories and evolving summary representations, we systematically examined how conversational context modeling impacts grounding, personalization, and dialogue quality.

Experimental results across offline and interactive tasks demonstrate that both context strategies are highly effective. Full-history conditioning provides modest advantages in long-horizon dialogue coherence and user satisfaction, while summary-based conditioning achieves comparable overall performance with improved contextual efficiency and engagement. The strong leaderboard placements of both runs confirm the robustness of our pipeline design and highlight the viability of structured conversational memory for scalable interactive systems.

References

[1] Mohammad Aliannejadi, Simon Lupart, Marcel Gohsen, Nailia Abbasiantaeb, Zahra andMirzakhmedova, Johannes Kiesel, and Jeffrey Dalton. TREC iKAT 2025: The interactive knowledge assistance track overview. In *Proceedings of the 34th Text REtrieval Conference (TREC 2025)*, 2025.

[2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[3] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3360–3362, 2022.