GUIDANCE@TREC iKAT 2025

Ahmed Rayane Kebir* Victor Morand†
Pierre-Antoine Lequeu‡ Zineddine Tighidet§
Mitodru Niyogi¶ Jonah Turner ‖ Rishu Kumar **
Benjamin Piwowarski††

1 Introduction

The report describes the work conducted by several teams involved in the ANR GUIDANCE project for the iKAT evaluation campaign. The ANR GUIDANCE aims to advance research in General Purpose Dialogue-assisted Digital Information Access. This involves tackling challenges such as the design and adaptation of large language models (LLMs) for better information access, enhancing LLMs' generalization capabilities to new domains and languages, ensuring the truthfulness of outputs, and addressing the lack of open-access state-of-the-art models. The GUIDANCE project also seeks to unite the French Information Retrieval community and produce open-access resources for model evaluation and development.

This report first details the retrieval module Section 2, then the generation methodology in Section 3. We finally describe the runs that were finally submitted in Section 4.

2 Conversational IR

Document retrieval is done using a first stage sparse retriever tailored for conversation. We grounded our approach in CoSPLADE [2] and explored ways to increase the length of the context that can be considered by using more modern architectures than DistilBERT.

*IRIT, Université de Toulouse, Work completed while the author was at ISIR, Sorbonne Université
†ISIR, Sorbonne Université
‡ISIR, Sorbonne Université
§ISIR, Sorbonne Université
¶CNRS, LIG, Université Grenoble Alpes
‖Mixedbread AI, Work completed while the author was at IRIT, Université de Toulouse
**IRIT, Université de Toulouse
††CNRS, ISIR, Sorbonne Université

2.1 CoSPLADE

CoSPLADE extends SPLADE to conversational contexts by aligning the sparse representation of the full conversational history Q_{conv} with that of a gold rewritten query q^* (e.g "when is she born" rewritten into "when is Queen Elizabeth II born" according to the conversation). Different from CoSPLADE, we minimize the cosine between the student's predicted term weights and the frozen teacher's weights:

$$\mathcal{L}_{\text{MSE}} = \sum_{v \in \mathcal{V}} \left\| w_\theta(v \mid Q_{\text{conv}}) - w_{\text{SPLADE}}(v \mid q^*) \right\|^2 \tag{1}$$

where $w(v \mid \cdot)$ denotes the log-saturation weight for token v in the vocabulary $\mathcal{V}$. The model is trained on the QReCC and OrConvQA datasets.

2.2 Long Context CoSPLADE - CoSPLADEv2

The main limitation of SPLADE++ is its context length, which is limited to 512 tokens. It is not enough to fit the whole conversation history in many cases. We therefore decided to use another backbone, namely NeoBERT [1]. NeoBERT is a modern BERT variant that shares the original BERT tokenizer but allows analyzing texts of up to 4096 tokens.

NeoBERT Backbone The first phase of retrieval training was the distillation of SPLADEv3 into NeoBERT. Queries and documents from the MSMARCO document corpus were embedded by both models, and a simple Distillation MSE loss was computed between both embeddings to bring the student's representations closer to those of the teacher, SPLADEv3. A maximum sequence length of 512 tokens was set for training.

Distillation for conversation Using this backbone, we then trained a CoSPLADE model, which we name CoSPLADEv2. Its longer context length allows us to input the whole conversation in the encoder. We then followed the same training methodology than in Section 2.1.

3 Answer generation

For response generation, we employ a modified Retrieval-Augmented Generation (RAG) pipeline [3], designed to reduce hallucinations and ensure that outputs remain grounded in retrieved evidence. In all runs, we select the 10 best documents of the passage ranking phase as documents for the generation. We use GPT-4.1 as a generative model for all steps in the generation phase.

PTKBs Selection. The relevant PTKBs are selected at each user utterance. To do so, GPT-4.1 is given the PTKB selection task explanation as a system instruction, and then prompted with the full set of PTKBs and their indices, the

conversation history, and the latest utterance of the user. It is asked to output the list of PTKBs indices, and allowed to output an empty list if no PTKB is relevant.

Ambiguity and Clarification Check. Rather than generating responses directly, we adopt a Chain-of-Thought (CoT) prompting strategy inspired by Tang et al. [5], who propose an ambiguity type taxonomy for conversational clarification. We extend this approach by instructing the LLM to first predict potential ambiguity types, which serve as intermediate reasoning signals. The model then determines whether the query is ambiguous and, if so, generates clarification requests to enable mixed-initiative interaction with the user.

Document Summarization. If no ambiguity is detected, the model produces a direct response summarizing the retrieved passages and grounded in the conversational context (relevant PTKBs, the conversation history and the last user utterance) avoiding hallucinations. We summarize retrieved passages to retain essential information and reduce input length, optimizing response generation efficiency.

GEval. We integrated the reference-free GEval [4] evaluation pipeline to evaluate the outputs of our modified RAG-based answer generation system, using metrics-Relevance, Completeness, Naturalness, and Groundedness to systematically assess how well responses leverage selected PTKBs, resolve ambiguities, and summarize retrieved evidence. This alignment ensures that generated answers are not only fluent and contextually appropriate but also demonstrably tied to the underlying retrieval and clarification stages, providing a consistent measure of end-to-end response quality.

4 Submitted runs

4.1 Response Generation Only

`gpt-clarif-sum-top10` gpt-4.1 pipeline that detects the need for clarification question, ask for a clarification or answer the user's query. use top-10 documents from retrieval and summarize each of them before querying.

4.2 Passage Ranking and Response Generation

`agg_false-qrec-mse-sum-top10` retrieval: CoSPLADEv2 model (NeoBERT backbone) aligned with SPLADEv3, trained on QReCC with MSE loss. The query representation is aggregated over the current query only. generation: gpt-4.1 pipeline to extract relevant PTKBs, to detect the need for clarification question, and to answer the user's query. use top-10 documents from retrieval and summarize each of them before querying.

`agg_true-qrec-mse-sum-top10` retrieval: CoSPLADEv2 model (NeoBERT backbone) aligned with SPLADEv3, trained on QReCC with MSE loss. The query representation is aggregated over all the conversation. generation: gpt-4.1 pipeline to extract relevant PTKBs, to detect the need for clarification question, and to answer the user's query. We use the top 10 documents from retrieval and summarize each of them before querying.

`cosine-orconvqa-sum-top10` retrieval: CoSPLADE model with a SPLADEv3 backbone, trained on OrConvQA with cosine loss, history size of 1. generation: gpt-4.1 pipeline to extract relevant PTKBs, to detect the need for clarification question, and to answer the user's query. use top-10 documents from retrieval and summarize each of them before querying.

5 Conclusion

In this report, we presented the GUIDANCE project's submission to the TREC iKAT 2025 campaign, focusing on robust Conversational Information Retrieval and grounded answer generation. To overcome the context limitations of previous sparse retrieval models, we introduced CoSPLADEv2, a NeoBERT-backed architecture capable of processing extended conversation histories, which we aligned with SPLADEv3 via distillation.

For the generation phase, we implemented a modified RAG pipeline that leverages Chain-of-Thought prompting to actively detect ambiguities and generate clarification requests when necessary. By integrating these retrieval advancements with a hallucination-resistant generation strategy—validated through GEval metrics—we aim to provide more accurate and context-aware responses. Future work will focus on analyzing the official iKAT evaluation results to further refine the interplay between sparse conversational embeddings and Large Language Model reasoning. Furthermore, we aim to achieve a better integration of the PTKB in both the retrieval and generation phases

Acknowledgments

The authors acknowledge ANR – FRANCE (French National Research Agency) for its financial support of the GUIDANCE project n°ANR-23-IAS1-0003. This work was carried out using HPC resources from GENCI-IDRIS (Grant 2025-A0191016944)

References

[1] Lola Le Breton, Quentin Fournier, Mariam El Mezouar, and Sarath Chandar. Neobert: A next-generation bert, 2025.

[2] Nam Le Hai, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. CoSPLADE: Contextualizing SPLADE for Conversational Information Retrieval.

[3] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[4] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[5] Anfu Tang, Laure Soulier, and Vincent Guigue. Clarifying ambiguities: on the role of ambiguity types in prompting methods for clarification generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 20–30, 2025.