Webis at TREC 2025: Tip-of-the-Tongue Track and AutoJudge

Maik Fröbe*
Friedrich-Schiller-Universität Jena
Jena, Germany

Jan Heinrich Merker*
Friedrich-Schiller-Universität Jena
Jena, Germany

Eric Oliver Schmidt*
Martin-Luther-Universität Halle
Halle, Germany

Martin Potthast
Universität Kassel
Kassel, Germany

Matthias Hagen
Friedrich-Schiller-Universität Jena
Jena, Germany

Abstract

This paper describes the Webis Group's participation in the 2025 edition of TREC. We participated in the Tip-of-the-Tongue track and the pilot round of the AutoJudge track. For our participation in the Tip-of-the-Tongue track, we re-executed our query relaxation strategies that we developed in our previous years submissions (removing terms that likely reduce retrieval effectiveness). For the pilot round of the AutoJudge track we apply axiomatic thinking by using preferences and features from all 29 axiomatic constraints for retrieval augmented generation that are implemented in the `ir_axioms` package (evaluation is in progress).

Keywords

Tip-of-the-Tongue, Retrieval-Augmented Generation

1 Introduction

We describe our submissions to the Tip-of-the-Tongue Track [1–3, 8] and the pilot round of AutoJudge. For the Tip-of-the-Tongue track, we submitted 2 runs that focused to re-execute our submission from the last years [5, 6]. Our last-years approaches applied long-query reduction approaches. The idea is to remove terms from the query that confuse the retrieval model, i.e., improving the recall by making the query smaller. We re-executed our previous years long query reduction, but switched from the closed-source GPT backbone to open source language models. For AutoJudge, we derive leaderboards from axiomatic constraints that were developed for retrieval augmented generation [10] that are implemented in the `ir_axioms` package [4]. We will experiment with extending the existing axiom implementations to not only expose pairwise preferences but also pointwise explanations. By combining axiom preferences for different objectives (e.g., consistency or correctness of a generated text), we plan to compose more comprehensive judgments.

2 AutoJudge

Our submissions to the pilot round of the AutoJudge shared task are still under development (as we had the idea to submit them only very shortly before the deadline). We aim to exploit 11 axiomatic constraints designed for retrieval augmented generation [10] and 18 classical axiomatic retrieval constraints as implemented in the `ir_axioms` package [4] to derive judgment qrels for the AutoJudge task. While the `ir_axioms` package currently only exposes pairwise preferences between documents for a given query, we are working on extending the axioms to also provide pointwise explanations that can be used to derive absolute judgments for a pool of generated responses with optional citations. Ultimately, our goal is to combine the preferences and explanations of all 29 axioms to derive comprehensive judgments that consider generated response quality from five different perspectives [7]: coverage, consistency, correctness, coherence, and clarity. Further task-specific axioms may enrich the derived judgments.

Table 1: Effectiveness of our 2 runs in the Tip-of-the-Tongue track.

Approach	nDCG@10	Recall@100
webis-bm25-gpt-oss	0.3078	0.7637
webis-bm25-llama	0.1968	0.6785

3 Tip-of-the-Tongue Track

We submit two runs to the Tip-of-the-Tongue track. We use the official baseline [9] PyTerrier index. We use query relaxation with large language models with our best approach from our 2023/2024 submission by instructing a large language model to leave out terms that likely reduce the retrieval effectiveness. We use the prompt that worked best in our 2023 submission [1].

3.1 Submitted Approaches

Our two runs are:

webis-bm25-gpt-oss. This run uses the PyTerrier index to retrieve with BM25 for LLM-rewritten queries. The LLM is instructed to remove terms that are unlikely to help retrieval. We use the OSS variant of GPT as LLM.

webis-bm25-llama. This run is identical to our other run but we use Llama as underyling large language model.

3.2 Results

Table 1 shows the effectiveness in terms of Recall at 1000 and at 100 and nDCG@10 on the Tip-of-the-Tongue track.

References

[1] Jaime Arguello, Samarth Bhargav, Fernando Diaz, Evangelos Kanoulas, and Bhaskar Mitra. 2023. Overview of the TREC 2023 Tip-of-the-Tongue Track. In *Proceedings of TREC 2023 (NIST Special Publication)*, Ian Soboroff and Angela

*These authors contributed to the paper equally and are listed alphabetically.

The Thirty-Fourth Text REtrieval Conference, TREC 2025, December 11–12, 2025, Rockville, Maryland
2025.

Ellis (Eds.). NIST, Gaithersburg, 13 pages. https://trec.nist.gov/pubs/trec32/papers/Overview_tot.pdf

[2] Jaime Arguello, Samarth Bhargav, Fernando Diaz, To Eun Kim, Yifan He, Evangelos Kanoulas, and Bhaskar Mitra. 2024. Overview of the TREC 2024 Tip-of-the-Tongue track. In *Proceedings of the Thirty-Third Text REtrieval Conference, TREC 2024, Gaithersburg, MD, USA, November 18-22, 2024 (NIST Special Publication, Vol. 1329)*, Ian Soboroff, Hoa Dang, and George Awad (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec33/papers/Overview_tot.pdf

[3] Jaime Arguello, Fernando Diaz, Maik Fröbe, To Eun Kim, and Bhaskar Mitra. 2026. Overview of the TREC 2025 Tip-of-the-Tongue track. *CoRR* abs/2601.20671 (2026). https://doi.org/10.48550/ARXIV.2601.20671 arXiv:2601.20671

[4] Alexander Bondarenko, Maik Fröbe, Jan Heinrich Reimer, Benno Stein, Michael Völske, and Matthias Hagen. 2022. Axiomatic Retrieval Experimentation with ir_axioms. In *45th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2022)*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 3131–3140. https://doi.org/10.1145/3477495.3531743

[5] Maik Fröbe, Christine Brychcy, Elisa Kluge, Eric Oliver Schmidt, and Matthias Hagen. 2023. Webis at TREC 2023: Tip-of-the-Tongue Track. In *Proceedings of TREC 2023 (NIST Special Publication)*, Ian Soboroff and Angela Ellis (Eds.). NIST, Gaithersburg, 3 pages. https://trec.nist.gov/pubs/trec32/papers/Webis.T.pdf

[6] Maik Fröbe, Lukas Gienapp, Jan Heinrich Merker, Harrisen Scells, Eric Oliver Schmidt, Matti Wiegmann, Martin Potthast, and Matthias Hagen. 2024. Webis at TREC 2024: Biomedical Generative Retrieval, Retrieval-Augmented Generation, and Tip-of-the-Tongue Tracks. In *33th International Text Retrieval Conference (TREC 2024) (NIST Special Publication)*, Ellen M. Voorhees and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST), 9 pages. https://trec.nist.gov/pubs/trec33/index.html

[7] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Evaluating Generative Ad Hoc Information Retrieval. In *Proceedings of SIGIR 2024*. ACM, New York, 14 pages. https://doi.org/10.1145/3626772.3657849

[8] Yifan He, To Eun Kim, Fernando Diaz, Jaime Arguello, and Bhaskar Mitra. 2025. Tip of the Tongue Query Elicitation for Simulated Evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (Eds.). ACM, 3398–3407. https://doi.org/10.1145/3726302.3730335

[9] Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval. In *Proceedings of CIKM 2021*. ACM, New York, 4526–4533. https://doi.org/10.1145/3459637.3482013

[10] Jan Heinrich Merker, Maik Fröbe, Benno Stein, Martin Potthast, and Matthias Hagen. 2025. Axioms for Retrieval-Augmented Generation. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025*, Hamed Zamani, Laura Dietz, Benjamin Piwowarski, and Sebastian Bruch (Eds.). ACM, 67–77. https://doi.org/10.1145/3731120.3744601