Hedge Removal, Query Drift, and the Simulation Gap in Tip-of-the-Tongue Retrieval

Bruno N. Sotic
University of Amsterdam
Amsterdam, The Netherlands
nadalic.sotic@uva.nl

Jaap Kamps
University of Amsterdam
Amsterdam, The Netherlands
kamps@uva.nl

Abstract

Retrieving known items from vague, verbose queries (the “Tip-of-the-Tongue” or ToT problem) poses a unique challenge for information retrieval. In the TREC 2025 ToT Track, we investigated linguistic preprocessing strategies (such as hedge removal and negation penalties) and hybrid retrieval methods across simulated and human-generated queries. Our experiments reveal a substantial divergence between LLM-simulated development data and the official test set. Hedge removal yielded large gains on verbose synthetic queries (+25.4% NDCG on dev3), but minimal improvement on sparse human queries (+2.7% on test, not significant). Negation penalties produced no measurable effect across all conditions. Pseudo-Relevance Feedback (RM3) consistently degraded performance, amplifying query drift rather than resolving vocabulary mismatch. Analysis of the recall pool reveals a fundamental bottleneck: only 32% of target documents appeared among the top 100 BM25 results on the test set. Within this constraint, hybrid dense reranking improved early precision by 10.6% on hard queries where lexical matching failed, but degraded performance on queries with strong term overlap. We conclude that systems optimized on LLM-simulated ToT data risk overfitting to synthetic linguistic patterns that do not reflect the sparse, fragmented nature of real human memory retrieval.

1 Introduction

The Tip-of-the-Tongue (ToT) phenomenon is a cognitive state in which a subject cannot retrieve a known item despite recalling specific features, often resulting in frustration [Brown, 1991, Schwartz, 2002]. In Information Retrieval (IR), this corresponds to the Known-Item Search (KIS) task, but with a distinct challenge: users cannot formulate queries using standard identifiers (e.g., exact titles or names) [Arguello et al., 2021]. Instead, ToT queries are characterized by vivid descriptions, episodic memories, and significant uncertainty [Elsweiler et al., 2011]. Similar observations were reported in the CLEF Social Book Search Track (2011-2016) [Koolen et al., 2011, 2012, 2013, Hall et al., 2014, Koolen et al., 2015, 2016], which reused LibraryThing Forum data, including many requests from the "Name that Book" forum.

The TREC ToT track began in 2023 with known-item retrieval in the movie domain [Arguello et al., 2023] and was extended in 2024 to include movies, celebrities, and landmarks [Arguello et al., 2024]. However, the 2025 edition introduces a critical methodological shift: the inclusion of Large Language Model (LLM) simulated queries alongside human-generated ones [Arguello et al., 2025]. This hybrid dataset raises a fundamental question about the validity of synthetic data in evaluating retrieval systems for cognitive tasks. While LLMs can simulate the content of a memory, they may introduce stylistic artifacts, specifically, an overabundance of epistemic modality markers (e.g., “I believe,” “if I recall,” “maybe”) - that might not reflect the fragmented, keyword-sparse nature of real human forgetting [Liu et al., 2025].

This distinction creates a tension between lexical parsimony and semantic expansion. Standard lexical models (BM25) struggle with the “vocabulary mismatch” [Furnas et al., 1987] inherent in ToT scenarios, where the user’s description (e.g., “the movie with the blue alien”) shares no terms with the metadata (e.g., "Avatar"). While query expansion techniques (RM3) typically bridge this gap, they risk inducing query drift when the initial retrieval precision is low [Carpineto and Romano, 2012, Dehghani et al., 2016]. In addition to query noise, a more fundamental issue may arise when lexical retrieval fails to include the correct item within the initial candidate pool. In such cases, reranking strategies are bounded by a recall ceiling, suggesting a structural limitation of sparse retrieval under severe vocabulary mismatch.

In this work, we present a comparative analysis of linguistic de-noising and hybrid retrieval strategies across simulated and real ToT scenarios. We utilize the presence of "hedges" (uncertainty markers) as a probe to distinguish the retrieval dynamics of synthetic versus human queries. Our experiments show a significant contrast, i.e., strategies that succeed on verbose, simulated queries (hedge removal) yield diminishing returns on real, human-curated test sets. For queries with severe vocabulary mismatch (as expected in real ToT), we investigate whether semantic matching can recover the recall gap that lexical methods cannot address. Consequently, we implement a Hybrid Dense Reranking pipeline that, despite being bounded by the recall limitations of the initial sparse retrieval, achieves a 10.6% improvement in early precision on the official Test Set by reordering semantically similar candidates within the top-100 pool.

We guide our analysis through three Research Questions (RQs):

- **RQ1 (Linguistic Preprocessing and Simulation Gap):** To what extent does the removal of epistemic modality markers (hedges) and the application of negation penalties impact retrieval performance, and do these effects differ between LLM-simulated and human queries?
- **RQ2 (The Risk of Expansion):** Does Pseudo-Relevance Feedback (RM3) mitigate vocabulary mismatch in known-item search, or does the initial low precision of ToT queries lead to query drift?
- **RQ3 (Precision within Recall Constraints):** When lexical retrieval yields limited recall in ToT queries, can dense semantic reranking meaningfully improve early precision within the constrained candidate pool?

The rest of this paper is organized as follows. Section 2 discusses our approach and experimental methodology. We detail our results on the synthetic development data in Section 3 and the results on the human data in Section 4. Section 5 conducts further analysis of our results. Section 6 ends the paper with discussion and conclusions.

2 Methodology

All experiments were implemented in the PyTerrier framework using a standard inverted index over the provided corpus (6.5M JSONL documents; approximately 20GB). Due to hardware constraints (16GB RAM, single consumer GPU), we implemented a lightweight byte-offset lookup structure over the raw JSONL files. This enables constant-time access to document text during reranking without requiring the full corpus to be loaded in memory. The indexer stores byte pointers to document start positions, allowing efficient retrieval of document content after initial ranking.

2.1 Lexical Query Processing

We evaluate two query-level interventions prior to retrieval: hedge removal and negation-based penalties. These transformations are deterministic and applied identically across development and test splits.

Hedge Removal. We compiled a fixed lexicon of approximately 120 epistemic modality markers (e.g., *apparently, maybe, roughly, I think, I believe, it seemed*). The lexicon was constructed manually by inspecting a sample of development queries and prior linguistic descriptions of uncertainty markers.

During preprocessing, all hedge terms were removed from the query string before tokenization. No stemming or additional normalization beyond the PyTerrier default pipeline was modified. This intervention affects only query terms and does not modify the index.

Negation Constraints. We implemented a deterministic post-retrieval penalty for explicit negation patterns. First, we identify negation cues (e.g., *no, not, without*) in the query. These are matched against a predefined list of "negative heads" (e.g., *sequel, remake, horror, animation*), selected based on common contrastive ToT descriptions in prior editions of the track.

If a query contains a negation phrase targeting one of these heads (e.g., "not a sequel"), then any document whose title or first paragraph contains the negated concept receives a fixed additive penalty of $w = -3.0$ to its BM25 score. This penalty is applied after retrieval but before reranking.

No machine-learned weighting or dynamic calibration was performed. The weight was selected conservatively to avoid removing documents entirely from the candidate pool.

2.2 Design Choices

We restrict our interventions to rule-based preprocessing and deterministic modifications to scoring. In contrast to generative query rewriting approaches (e.g., LLM-based reformulation), this design ensures reproducibility and avoids introducing stochastic effects that could confound ablation analysis.

We do not perform manual query annotation beyond lexicon construction, as prior ToT track reports have documented characteristic linguistic patterns in this task setting. Our goal is to isolate the effect of minimal, interpretable modifications to the retrieval pipeline.

2.3 Retrieval Architecture

We evaluate three retrieval configurations: a sparse lexical baseline (BM25), a relevance-based query expansion variant (RM3), and a hybrid sparse–dense reranking strategy. All parameters were fixed across development and test splits.

BM25 (Baseline). As a standard lexical baseline, we use BM25 with $k_1 = 1.2$ and $b = 0.75$. No additional field weighting or query normalization beyond the preprocessing described in Section 3.1 is applied. Scores are computed over the full index, and the top 1,000 documents are retained for evaluation.

RM3 (Pseudo-Relevance Feedback). To assess whether query expansion mitigates vocabulary mismatch, we implement RM3 using PyTerrier's default formulation with $fb\_docs = 10$ and $fb\_terms = 20$. Expansion terms are derived from the top-ranked documents of the initial BM25 run. The expanded query is then reissued against the same index. No interpolation parameter tuning was performed beyond framework defaults. This configuration allows us to measure the effect of pseudo-relevance feedback under known-item retrieval conditions.

Hybrid Dense Reranking. To incorporate semantic similarity without constructing a full dense index, we adopt a two-stage reranking approach.

First, we retrieve the top $k = 100$ documents using the best-performing sparse configuration (BM25 with hedge removal). This defines the lexical recall pool.

Second, we encode the preprocessed query and each candidate document using the Sentence-BERT bi-encoder.[1] For computational consistency, only the first 1,000 characters of each document are encoded. Embeddings are generated independently (bi-encoder setting) without cross-attention.

Third, we compute cosine similarity between the query embedding and each candidate document embedding. The top 100 documents are reordered according to this similarity score. Documents ranked below 100 retain their original BM25 ranking.

This hybrid strategy preserves the sparse retrieval stage for initial candidate generation while allowing semantic similarity to influence early precision within the lexical recall pool.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Table 1: **Development Set Ablation Results.** Metrics reported are Reciprocal Rank (RR), NDCG@10, Success@1, and Success@10. Significance (p-val) is calculated via paired t-test against the respective backbone baseline (BM25 or RM3). Note the massive improvement on Dev3 compared to Dev1/2.

Split	**Run ID**	**RR**	**NDCG@10**	**Succ@1**	**Succ@10**	**P-Value**
Dev1 (Hard)	BM25 (Baseline)	0.0811	0.0836	0.0634	0.1056	-
	BM25 + Hedges	**0.0875**	**0.0935**	0.0634	**0.1268**	0.091
	BM25 + Negation	0.0811	0.0836	0.0634	0.1056	-
	RM3 (Baseline)	0.0396	0.0451	0.0211	0.0704	-
	RM3 + Hedges	0.0463	0.0531	0.0282	0.0845	0.185
Dev2 (Hard)	BM25 (Baseline)	0.0875	0.0994	0.0490	0.1469	-
	BM25 + Hedges	**0.0976**	**0.1070**	**0.0629**	0.1469	0.204
	BM25 + Negation	0.0875	0.0994	0.0490	0.1469	-
	RM3 (Baseline)	0.0474	0.0583	0.0210	0.1049	-
	RM3 + Hedges	0.0509	0.0645	0.0280	0.1189	0.153
Dev3 (Easy)	BM25 (Baseline)	0.3129	0.3366	0.2444	0.4347	-
	BM25 + Hedges	**0.3964**	**0.4221**	**0.3284**	**0.5224**	**< 0.001**
	BM25 + Negation	0.3129	0.3366	0.2444	0.4347	-
	RM3 (Baseline)	0.1569	0.1868	0.0877	0.3022	-
	RM3 + Hedges	0.2006	0.2434	0.1175	0.4011	< 0.001

3 Results: Development Phase (Ablation Study)

We first evaluated our approaches on three development splits (‘dev1‘, ‘dev2‘, ‘dev3‘) provided by the organizers. A pattern emerged distinguishing ‘dev3‘ from the others.

The ‘dev3‘ split (which is a mix-domain set, whereas dev1 and dev2 are movie queries) yielded baseline scores (0.33) nearly $4\times$ higher than ‘dev1‘ or ‘dev2‘. Our “Hedge Removal” strategy achieved a statistically significant improvement ($p < 0.001$) of **+25.4%** on ‘dev3‘.

We interpret ‘dev3‘ as a dataset dominated by LLM-simulated queries that explains the results. Qualitative inspection shows a difference in query composition. Dev1 and Dev2 appear to be human-generated, characterized by misspellings (e.g., ”dengeros trip”,”alot”), informal grammar, and highly specific but fragmented episodic memories (e.g., “I saw it on Netflix in 2010”). These represent a "Hard" retrieval scenario with significant vocabulary mismatch.

In contrast, dev3 queries exhibit the hallmarks of Large Language Model simulation. They are grammatically pristine, structurally verbose, and descriptively exhaustive (e.g., ”I remember visiting this incredible place... it had this ancient, almost mythical vibe”). While they contain "hedge" phrases (”I think,” “maybe”), the underlying descriptions act as high-quality semantic summaries. The Hedge Removal strategy achieved a massive +25.4% gain on dev3 by removing the synthetic "roleplay" markers, which revealed a near-perfect lexical match with the target documents. Hence, the high performance on ‘dev3‘ is largely an artifact of the simulation process, whereas ‘dev1/2‘ better represents the messy reality of the ToT task.

3.1 The Failure of Expansion (RM3)

Across all splits, Pseudo-Relevance Feedback (RM3) degraded performance (e.g., -45% on ‘dev3‘). In known-item retrieval, if the top-ranked documents are incorrect, RM3 learns “distractor” terms from the wrong movies and expands the query with noise, causing query drift.

Table 2: **Official Test Set Results.** Note that BM25+Negation was omitted as it produced identical rankings to the baseline in all previous runs.

Run ID	**RR**	**NDCG@10**	**Succ@10**	**P-Val**
BM25 (Base)	0.1114	0.1223	0.1736	-
BM25 + Hedges	**0.1136**	**0.1257**	**0.1817**	0.239
RM3 (Base)	0.0725	0.0846	0.1399	-
RM3 + Hedges	0.0821	0.0954	0.1559	**< 0.001**
Hybrid Dense	-	0.1190	**0.2010**	-

4 Official Test Set Results

The Official Test Set results diverged significantly from the simulated ‘dev3‘ environment, aligning more closely with the “Hard” development splits.

4.1 Divergence from Simulation

Table 2 shows the official test set results. On the Official Test Set, the massive gains seen in ‘dev3‘ disappeared. Hedge removal resulted in a modest +2.7% gain in NDCG, which was not statistically significant ($p = 0.239$). This indicates that the Official Test Set queries are qualitatively different – likely shorter, human-curated, or less reliant on synthetic “hedge” patterns.

However, Hedge Removal played a critical role in stabilizing the RM3 pipeline, providing a highly significant improvement ($p < 0.001$) over the RM3 baseline. Removing noise from the “seed query” helped purify the expansion terms, though not enough to beat BM25.

5 Analysis

To explain the discrepancy between ‘dev3‘ and ‘test‘ and to address the limitations of our pipeline, we perform an additional diagnostic analysis that correlates query characteristics with performance.

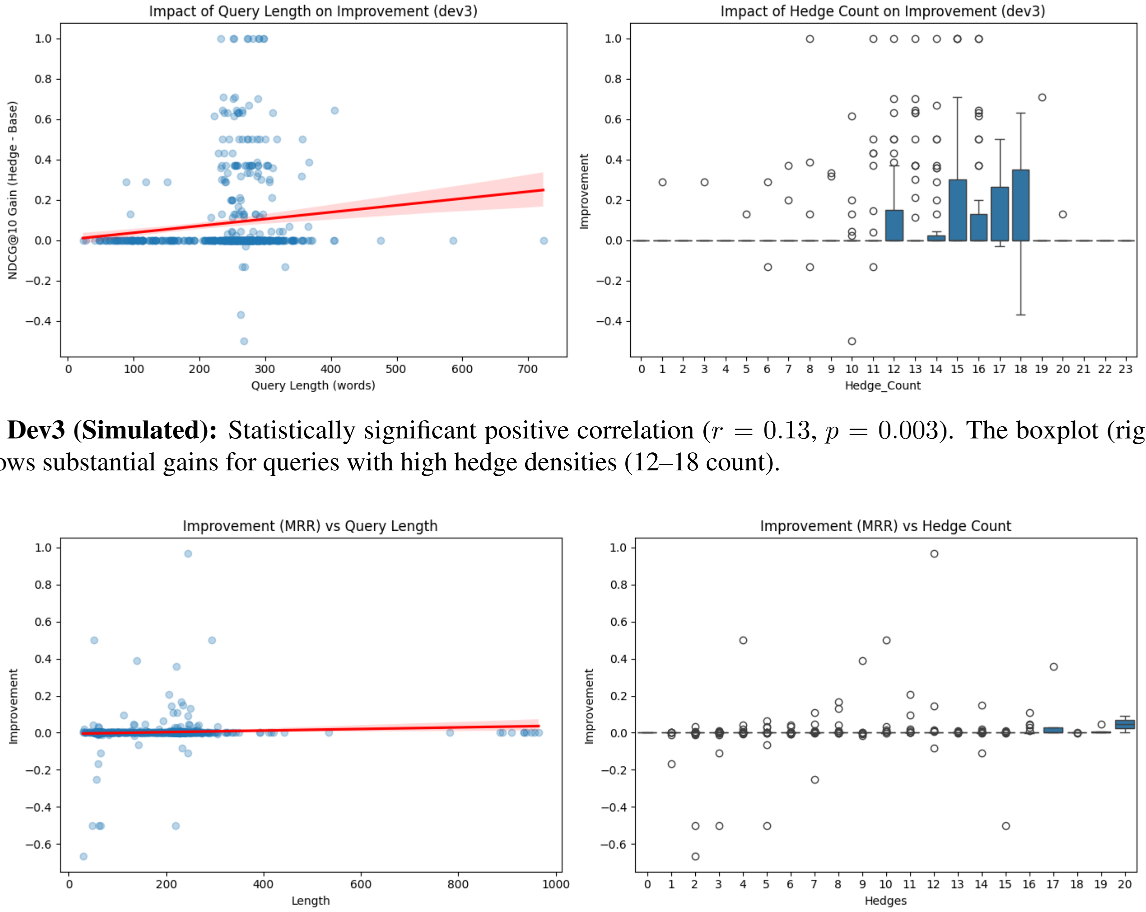
5.1 Query Characteristics

To validate whether the performance gains from Hedge Removal were driven by the specific characteristics of LLM-generated text (verbosity and synthetic uncertainty), we analyzed the relationship between Query Length and Performance Improvement (NDCG@10 Δ) across splits.

As illustrated in Figure 1a, the ‘dev3‘ split exhibits a statistically significant positive correlation between query length and performance gain ($Pearson\ r = 0.1276, p = 0.0031$). The regression line indicates that as query verbosity increases, the benefit of stripping hedge markers grows. Furthermore, the boxplot analysis for ‘dev3‘ reveals a “threshold effect”: queries containing a high density of hedges (10–18 count) show the largest median improvements. This suggests that, for this split, the method primarily functions as a denoiser for verbose, hallucinated uncertainty markers typical of LLM outputs.

In stark contrast, Figure 1b shows that this relationship decouples on the Official Test Set. The correlation drops to $r = 0.0726$ with a p-value of 0.0704 (exceeding the standard $\alpha = 0.05$ threshold for significance). More critically, the distribution of hedges is fundamentally different: while ‘dev3‘ contained many queries with > 12 hedges, the Official Test Set is far more concise, with the vast majority of queries containing only 0–5 hedge markers.

The Official Test Set is qualitatively harder and less reliant on synthetic verbosity than the development data. Consequently, the “easy wins” achieved by simple query reduction on ‘dev3‘ do not translate linearly to the test set, where the challenge shifts from noise reduction to solving vocabulary mismatch.



(a) **Dev3 (Simulated):** Statistically significant positive correlation ($r = 0.13$, $p = 0.003$). The boxplot (right) shows substantial gains for queries with high hedge densities (12–18 count).

(b) **Test Set (Real/Hard):** The correlation weakens and loses significance ($r = 0.07$, $p = 0.07$). Note the x-axis on the boxplot: the vast majority of test queries contain fewer than 5 hedges.

Figure 1: Diagnostic analysis comparing Query Length/Hedge Count vs. Performance Gain. The strong “de-noising” effect observed in the simulated Dev3 split disappears in the Official Test Set.

Table 3: **Deep Recall Analysis (Dev3 vs. Test).** Comparing the Recall Pool (Success@100) reveals the difficulty of the Test Set. While 66% of simulated queries retrieved the target within the top 100, only 32% of real test queries did so.

Metric	**Dev3 (Simulated)**	**Test Set (Real)**
Success@10 (Precision)	0.5224	0.1817
Success@100 (Recall Pool)	**0.6623**	**0.3248**
Recall Gap	0.1399	0.1431

5.2 The “Recall Wall”: Simulated vs. Real

To diagnose the performance gap between the development and test phases, we analyzed the Success@100 (the "Recall Pool") for our best lexical run (‘bm25 hedges‘). This metric indicates the theoretical upper bound of any reranking strategy.

As shown in Table 3, ‘dev3‘ exhibits a healthy recall pool (0.6623), confirming that for verbose, simulated queries, lexical overlap is usually sufficient to retrieve the target.

In stark contrast, the Official Test Set hits a “Recall Wall” at 0.3248. This implies that for 67.5% of test queries, the correct document was not among the top 100 BM25 results. This confirms a vocabulary mismatch in the real-world data that is absent in the simulated data. The user describes the movie using

Table 4: Impact of Dense Reranking on Precision (Success@10) across splits. While it failed on the easy Dev3 split, it succeeded on the hard Test split.

Split	BM25 (Lexical)	Hybrid (Dense)	Impact
Dev3	0.5224	0.3825	-26.7%
Test	0.1817	**0.2010**	**+10.6%**

terms that simply do not appear in the metadata. This recall bottleneck explains why absolute scores on the Test Set remained low regardless of the reranking.

5.3 Dense Reranking

The following section compares a Hybrid Dense Reranker across splits.

Table 4 shows the effectiveness of the lexical and dense rankers, prioritizing lexical precision or dense recall. On the “easy” ‘dev3‘ queries, dense reranking hurt performance. This is likely because easy queries contain exact entity matches (names, unique nouns) that BM25 handles well, whereas dense embeddings “smooth over” these specific signals, causing semantic drift.

However, on the “hard” Test Set, Dense Reranking improved Precision by 10.6%. In these difficult queries, exact keywords fail. The Dense model, by capturing the “semantic vibe” or plot description, successfully salvaged relevant documents that BM25 had buried deep in the ranking (ranks 11–100).

6 Discussion and Conclusions

In the TREC 2025 Tip-of-the-Tongue Track, we examined the effectiveness of linguistic preprocessing and hybrid retrieval strategies across simulated and human-generated queries.

RQ1 (Linguistic Preprocessing and Simulation Gap): We evaluated two query-level interventions: hedge removal and negation penalties. Hedge removal effectiveness is strongly dependent on query characteristics. LLM-simulated queries (dev3) contained high hedge density (10–18 markers) and responded strongly to hedge removal, while the official test set contained sparse queries (0–5 hedges) with minimal benefit from preprocessing. In contrast, our negation penalty approach produced no measurable effect across all splits, suggesting that rule-based score adjustments are insufficient for handling logical constraints in sparse retrieval. We conclude that high hedge counts are an artifact of LLM generation, not a universal characteristic of human ToT states.

RQ2 (The Risk of Expansion). Pseudo-Relevance Feedback (RM3) consistently degraded performance across all conditions. In known-item retrieval with low initial precision, expansion models learn from incorrect top-ranked documents, introducing distractor terms that amplify query drift rather than resolving vocabulary mismatch.

RQ3 (Precision within Recall Constraints). Our analysis reveals a fundamental recall bottleneck: only 32% of target documents appeared in the top-100 BM25 results on the test set. Within this constraint, hybrid dense reranking improved precision by reordering semantically similar candidates. However, the approach provides no benefit on queries with strong lexical overlap, where exact term matching suffices. Dense reranking is effective only when lexical methods fail, but the target remains within the candidate pool.

Limitations. Our negation penalty intervention produced no measurable effect, suggesting that rule-based score adjustments are insufficient for handling logical constraints in sparse retrieval. Additionally, our reranking approach cannot address queries where the correct document was excluded from the initial candidate pool.

Implications for Track Design. The substantial performance gap between dev3 and the official test set raises questions about relying on LLM-simulated data for system development and evaluation. Synthetic queries exhibit linguistic patterns that respond to simple preprocessing strategies, potentially rewarding shallow optimizations over genuine semantic understanding. Future ToT evaluations could also investigate human-curated vs. LLM-generated queries and their impact on ToT evaluations, or employ adversarial simulation techniques that better reflect the cognitive difficulty of genuine ToT states.

Future Work. Addressing the recall bottleneck will likely require dense retrieval architectures capable of independent candidate generation. Additionally, investigating structured retrieval methods that can enforce logical constraints without post-hoc penalties remains an open challenge for the ToT domain.

Acknowledgments

Bruno Sotic and Jaap Kamps are supported by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

J. Arguello, A. Ferguson, E. Fine, B. Mitra, H. Zamani, and F. Diaz. Tip of the tongue known-item retrieval: A case study in movie identification. In F. Scholer, P. Thomas, D. Elsweiler, H. Joho, N. Kando, and C. Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 5–14. ACM, 2021. doi: 10.1145/3406522.3446021. URL https://doi.org/10.1145/3406522.3446021.

J. Arguello, S. Bhargav, F. Diaz, E. Kanoulas, and B. Mitra. Overview of the TREC 2023 tip-of-the-tongue track. In I. Soboroff and A. Ellis, editors, *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023), Gaithersburg, MD, USA, November 14-17, 2023*, volume 500-xxx of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2023. URL https://trec.nist.gov/pubs/trec32/papers/Overview_tot.pdf.

J. Arguello, S. Bhargav, F. Diaz, T. E. Kim, Y. He, E. Kanoulas, and B. Mitra. Overview of the TREC 2024 tip-of-the-tongue track. In I. Soboroff, H. Dang, and G. Awad, editors, *Proceedings of the Thirty-Third Text REtrieval Conference, TREC 2024, Gaithersburg, MD, USA, November 18-22, 2024*, volume 1329 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2024. URL https://trec.nist.gov/pubs/trec33/papers/Overview_tot.pdf.

J. Arguello, F. Diaz, M. Fröbe, T. E. Kim, and B. Mitra. Overview of the TREC 2025 tip-of-the-tongue track. In *Proceedings of the Thirty-Fourth Text REtrieval Conference, TREC 2025, Online, December 11-12, 2025*, NIST Special Publication. National Institute of Standards and Technology (NIST), 2025.

A. S. Brown. A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109(2):204–223, 1991. doi: 10.1037/0033-2909.109.2.204. URL https://doi.org/10.1037/0033-2909.109.2.204.

C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50, 2012. doi: 10.1145/2071389.2071390. URL https://doi.org/10.1145/2071389.2071390.

M. Dehghani, S. Abnar, and J. Kamps. The healing power of poison: Helpful non-relevant documents in feedback. In S. Mukhopadhyay, C. Zhai, E. Bertino, F. Crestani, J. Mostafa, J. Tang, L. Si, X. Zhou, Y. Chang, Y. Li, and P. Sondhi, editors, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 2065–2068. ACM, 2016. doi: 10.1145/2983323.2983910. URL https://doi.org/10.1145/2983323.2983910.

D. Elsweiler, M. Harvey, and M. Hacker. Understanding re-finding behavior in naturalistic email interaction logs. In W. Ma, J. Nie, R. Baeza-Yates, T. Chua, and W. B. Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* *SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 35–44. ACM, 2011. doi: 10.1145/2009916.2009925. URL https://doi.org/10.1145/2009916.2009925.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987. doi: 10.1145/32206.32212. URL https://doi.org/10.1145/32206.32212.

M. M. Hall, H. C. Huurdeman, M. Koolen, M. Skov, and D. Walsh. Overview of the INEX 2014 interactive social book search track. In L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, editors, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*, pages 480–493. CEUR-WS.org, 2014. URL https://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-HallEt2014.pdf.

M. Koolen, G. Kazai, J. Kamps, A. Doucet, and M. Landoni. Overview of the INEX 2011 books and social search track. In S. Geva, J. Kamps, and R. Schenkel, editors, *Focused Retrieval of Content and Structure, 10th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2011, Saarbrücken, Germany, December 12-14, 2011, Revised Selected Papers*, volume 7424 of *Lecture Notes in Computer Science*, pages 1–29. Springer, 2011. doi: 10.1007/978-3-642-35734-3_1. URL https://doi.org/10.1007/978-3-642-35734-3_1.

M. Koolen, G. Kazai, J. Kamps, M. Preminger, A. Doucet, and M. Landoni. Overview of the INEX 2012 social book search track. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012. URL https://ceur-ws.org/Vol-1178/CLEF2012wn-INEX-KoolenEt2012.pdf.

M. Koolen, G. Kazai, M. Preminger, and A. Doucet. Overview of the INEX 2013 social book search track. In P. Forner, R. Navigli, D. Tufis, and N. Ferro, editors, *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013. URL https://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-KoolenEt2013b.pdf.

M. Koolen, T. Bogers, M. Gäde, M. M. Hall, H. C. Huurdeman, J. Kamps, M. Skov, E. Toms, and D. Walsh. Overview of the CLEF 2015 social book search lab. In J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings*, volume 9283 of *Lecture Notes in Computer Science*, pages 545–564. Springer, 2015. doi: 10.1007/978-3-319-24027-5_51. URL https://doi.org/10.1007/978-3-319-24027-5_51.

M. Koolen, T. Bogers, M. Gäde, M. M. Hall, I. Hendrickx, H. C. Huurdeman, J. Kamps, M. Skov, S. Verberne, and D. Walsh. Overview of the CLEF 2016 social book search lab. In N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, volume 9822 of *Lecture Notes in Computer Science*, pages 351–370. Springer, 2016. doi: 10.1007/978-3-319-44564-9_29. URL https://doi.org/10.1007/978-3-319-44564-9_29.

J. Liu, Q. Zong, W. Wang, and Y. Song. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–221, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.18. URL https://aclanthology.org/2025.acl-short.18/.

B. L. Schwartz. *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*. Psychology Press, 2002. doi: 10.4324/9781410604019. URL https://doi.org/10.4324/9781410604019.