Tip of the Tongue Query Elicitation for Simulated Evaluation

Yifan He*
Carnegie Mellon University
Pittsburgh, PA, USA
yifanhe@alumni.cmu.edu

To Eun Kim*
Carnegie Mellon University
Pittsburgh, PA, USA
toeunk@cs.cmu.edu

Fernando Diaz
Carnegie Mellon University
Pittsburgh, PA, USA
diazf@acm.org

Jaime Arguello
UNC Chapel Hill
Chapel Hill, NC, USA
jarguell@email.unc.edu

Bhaskar Mitra
Microsoft Research
Montréal, QC, Canada
bmitra@microsoft.com

Abstract

Tip-of-the-tongue (TOT) search occurs when a user struggles to recall a specific identifier, such as a document title. While common, existing search systems often fail to effectively support TOT scenarios. Research on TOT retrieval is further constrained by the challenge of collecting queries, as current approaches rely heavily on community question-answering (CQA) websites, leading to labor-intensive evaluation and domain bias. To overcome these limitations, we introduce two methods for eliciting TOT queries—leveraging large language models (LLMs) and human participants—to facilitate simulated evaluations of TOT retrieval systems. Our LLM-based TOT user simulator generates synthetic TOT queries at scale, achieving high correlations with how CQA-based TOT queries rank TOT retrieval systems when tested in the Movie domain. Additionally, these synthetic queries exhibit high linguistic similarity to CQA-derived queries. For human-elicited queries, we developed an interface that uses visual stimuli to place participants in a TOT state, enabling the collection of natural queries. In the Movie domain, system rank correlation and linguistic similarity analyses confirm that human-elicited queries are both effective and closely resemble CQA-based queries. These approaches reduce reliance on CQA-based data collection while expanding coverage to underrepresented domains, such as Landmark and Person. LLM-elicited queries for the Movie, Landmark, and Person domains have been released as test queries in the TREC 2024 TOT track, with human-elicited queries scheduled for inclusion in the TREC 2025 TOT track. Additionally, we provide source code for synthetic query generation and the human query collection interface, along with curated visual stimuli used for eliciting TOT queries.

CCS Concepts

• **Information systems** → **Test collections**; **Specialized information retrieval**; • **Human-centered computing** → **Interaction design**.

*Equal Contribution.

Keywords

Tip-of-the-Tongue Known-Item Retrieval; Synthetic Query Generation; Human Query Elicitation

1 Introduction

Tip-of-the-Tongue (TOT) retrieval is a known-item search task where the searcher aims to re-find a previously encountered item but struggles to formalize the query due to an inability to recall specific identifiers, such as a document title or a person's name [4, 10]. Unlike other known-item retrieval tasks, TOT information requests are typically verbose and complex, often blending semantic details about the context in which the searcher previously engaged with the item [25, 27]. These queries frequently include linguistic phenomena such as uncertainty, exclusion criteria, relative comparisons, and even false memories [4, 27], making them especially challenging for traditional retrieval systems [2, 4, 10, 26].

Addressing the challenges of TOT retrieval is critical because searchers in the TOT state experience higher level of frustration compared to other types of memory-related search failures [18]. TOT retrieval spans a wide range of domains, from casual leisure search [19], such as finding games [23], music [9], and books [10], to critical applications like email retrieval [16, 17, 24] and enterprise content search [15]. In professional settings, ineffective TOT support can lead to significant economic costs, as employees spend excessive time retrieving essential information, reducing productivity and efficiency [37].

The evaluation of TOT retrieval systems faces significant challenges due to limitations in test collection resources and current dataset construction methods. Most existing TOT queries in current datasets are sourced from community question answering (CQA) websites, which fail to fully capture the diversity of real-world TOT information needs [27]. This reliance on CQA platforms introduces domain skewness, as existing datasets and studies are heavily biased toward casual leisure topics like movies and books [11, 19], while underrepresenting areas such as people and landmarks. Even attempts to expand domain coverage [27] have had limited success,

Yifan He, To Eun Kim, Fernando Diaz, Jaime Arguello, & Bhaskar Mitra

leaving many domains underrepresented. Additionally, the labor-intensive nature of dataset construction—which relies on manual annotation, coding, and refinement [4, 9, 10, 27]—slows progress and restricts the ability to conduct large-scale and multi-domain evaluations.

Advancing TOT retrieval research requires overcoming these challenges, as the lack of diverse and representative datasets leaves retrieval systems poorly optimized for real-world scenarios, reducing their ability to handle complex and ambiguous queries effectively.

However, current approaches fail to address these issues due to data scarcity, limited scalability, and restricted scope. Search engine query logs, which could provide real-world insights, remain largely inaccessible due to corporate privacy [8] and confidentiality [31] concerns. As a result, prior efforts have relied on CQA platforms and manual labeling, but data collection restrictions from sites like Reddit, along with the scalability and domain bias issues, further limit dataset diversity and representativeness. These challenges underscore the need for more scalable and comprehensive approaches to TOT data collection and evaluation.

In response, this paper presents a novel evaluation framework for TOT retrieval systems, overcoming the limitations of CQA-based datasets by eliciting TOT queries from both large language models (LLMs) and humans. We develop an LLM-based TOT user simulator and a human query collection interface to address data scarcity, domain skewness, and scalability challenges, enabling more comprehensive evaluations.

Our research questions are as follows:

RQ 1: Can we elicit TOT queries from LLMs for effective simulated evaluation of TOT retrieval systems? Unlike prior work, we explore LLM-elicited queries for TOT evaluation, validating our simulator by measuring system rank correlation and linguistic similarity with CQA-collected queries. By testing various prompt configurations and temperatures, we demonstrate that LLM-elicited queries can effectively support TOT retrieval evaluation with high validity.

RQ 2: Can we elicit TOT queries from humans for effective simulated evaluation of TOT retrieval systems? To our knowledge, human-elicited TOT queries have not been previously used for evaluation. Similar to RQ1, we validate their effectiveness by comparing system rankings derived from human-elicited and CQA-based queries, demonstrating high alignment. Additionally, linguistic similarity analysis confirms that human-elicited queries share key characteristics with CQA-based queries, supporting their viability for TOT retrieval evaluation.

RQ 3: Can the elicitation-based evaluation methods be used in other domains underrepresented in CQA-collected test collections? While the development and validation of our methods focus on the Movie domain—where ample CQA-collected data exists—we test their applicability to Landmark and Person domains, which are less represented in existing datasets. Validation results show that elicited queries can achieve high system rank correlations in these new domains, suggesting the potential research direction on the general method of collecting TOT queries across multiple domains.

In addition to the research contributions, the core outcomes of this work lie in its resource contributions:

- A dataset of 450 synthetically generated TOT queries across three domains: Movie, Landmark, and Person.
- A method and prompt design for synthetically generating effective TOT queries from LLMs for simulated evaluation.[1]
- A human TOT query elicitation interface designed to collect effective TOT queries for simulated evaluation along with a carefully curated visual stimuli collection, covering the Movie, Landmark, and Person domains.[2]

2 Related Work

2.1 Query Simulation & Known-Item Retrieval

Query simulation methods have been used for various purposes, including document expansion [28] and synthetic test collection generation [32]. In the context of known-item retrieval, these methods have been explored to improve retrieval strategies [29, 30] and evaluation frameworks [5, 21].

Simulating the known-item queries has long been an active research area [6, 7, 17, 24]. Early work [6] generated synthetic queries using term-based likelihood models, selecting query terms based on their likelihood within a randomly chosen document. Later studies adapted this approach for desktop search [24] and email re-finding [17], demonstrating its effectiveness for simulated evaluations of know-item retrieval models. The *validation* of these query simulators has also been a key focus. System ranking correlation [7], retrieval score distribution comparisons [6], and synthetic versus human query resemblance [24] have been used to assess their reliability.

While valuable, known-item search queries differ significantly from TOT queries, which are longer and more complex. Despite progress in simulating known-item queries, TOT retrieval remains unexplored. This paper bridges that gap by introducing novel TOT query elicitation methods and adapting established validation techniques [38] to ensure alignment with real-world queries, enabling scalable and accurate simulated evaluations.

2.2 TOT Datasets

Several datasets have been developed to support research on TOT retrieval, primarily collected from online CQA platforms and focused on specific domains. MS-TOT [4] was constructed from the *IRememberThisMovie* website and human-annotated with tags in the Movie domain. It also includes qualitative coding of TOT queries and demonstrates significant room for improvement in current retrieval technologies for such information needs. Similarly, Jørgensen and Bogers [23] collected TOT queries from Reddit's */r/tipofmyjoystick* subreddit in the Game domain, providing coded tag information. Other datasets include Reddit-TOMT [10], focused on movies and books from Reddit's */r/tipofmytongue* subreddit; TOT-Music [9], targeting the Music domain from the same subreddit; and Whatsthatbook [26], sourced from *GoodReads*, focused on the Book domain.

In response to the domain specificity of these datasets, recent efforts have aimed to expand TOT datasets across multiple areas. Meier et al. [27] expanded to general casual leisure domains using data from six Reddit subreddits, including games, books, and music, although other identified domains, such as videos and people,

---
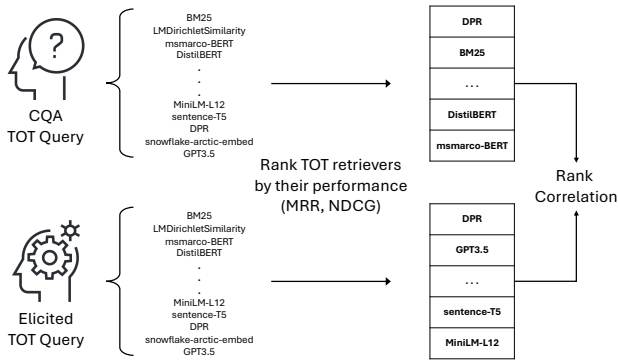
Figure 1: Validation of elicited queries using system rank correlation. We evaluate 40 different retrieval models using both CQA-based and elicited queries, ranking them based on search performance measured by MRR@1000 and NDCG@1000. We then compute Kendall's Tau and Pearson correlation to assess the agreement between the rankings derived from the two query sets.

Figure 2: Validation of automatic annotation using the MS-TOT dataset. Results show that GPT-4o-mini performs well in annotating TOT queries, closely aligning with human annotators. It achieves high accuracy and exhibits low Earth Mover's Distance (EMD), indicating strong agreement with expert annotations.

remain underrepresented. Similarly, TOMT-KIS [20] extended the collection from */r/tipofmytongue* by adapting Bhargav et al. [10]'s approach with fewer filtering restrictions, resulting in 1.28 million TOT queries. However, only 47% of these queries have identified answers, and the dataset continues to exhibit severe domain skewness toward a few topics.

In this work, we develop and validate TOT query elicitation methods using the Movie domain for robust evaluation, then expand to Landmark and Person to assess applicability across underrepresented domains.

3 Method

Overall, we elicit TOT queries from both LLMs and humans and validate them using two methods: system rank correlation (§3.2.1) and linguistic similarity (§3.2.2).
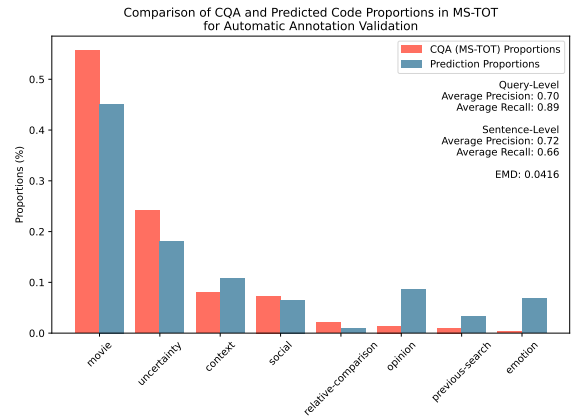
3.1 Query Elicitation

For LLM-elicited queries, we generate synthetic queries by exploring various prompting strategies. We experiment with different prompting conditions and model hyperparameters to identify the most effective prompt that yields the best validation results. This procedure is detailed in Section §4.

For human-elicited queries, we designed an interface that places participants in a TOT state using visual stimuli, such as movie stills, landmarks, and celebrity images. Participants then compose TOT queries as they would when posting on a CQA website. This procedure is described in Section §5.

3.2 Query Validation

3.2.1 System Rank Correlation. To validate the effectiveness of elicited TOT queries, we measure the correlation between its rankings of retrieval systems and rankings based on CQA queries for the same entities. This evaluation assesses whether retrieval models maintain consistent performance across different query sources but

on the same entities. If the rankings derived from elicited queries strongly correlate with those from CQA-based queries, it indicates that the elicited queries capture similar retrieval challenges and retrieval effectiveness. A high correlation suggests that our synthetic and human-elicited queries can serve as reliable substitutes for traditionally collected CQA-based queries in evaluating retrieval systems.

To compute these correlations, we run the queries on 40 different retrieval models, comprising lexical and dense retrievers. The lexical retrievers include BM25 [34] and language models with Dirichlet priors [39] using varying parameters. The dense retrievers include models of different sizes and performance levels, such as MiniLM-L6 and MiniLM-L12 [36]. To further introduce variation in systems, we include retrieval models with intentionally degraded performance by reinitializing the weights of certain layers in dense retrievers. Additionally, we incorporate an API-based closed-source LLM as one of the retrieval systems, specifically GPT-3.5-Turbo-Instruct, following the prompting format used in the baseline of the TREC 2024 TOT track to function as a ranker. Furthermore, we include the top-ranked retrieval system from the TREC 2023 TOT track [2, 12] to provide a strong performance reference point.

Figure 1 illustrates our validation strategy, showing how retrieval system rankings across different query sets provide insight into the reliability and effectiveness of our elicited queries.

3.2.2 Linguistic Similarity. Elicited TOT queries can exhibit substantial variation in writing style, word choice, experiences, and the presence of distorted memories [4, 27]. This diversity is inherent to the nature of TOT queries and is a valid characteristic of real-world TOT retrieval scenarios. Consequently, evaluating linguistic similarity between CQA-based and elicited queries using traditional methods, such as vector-based semantic similarity or lexicon-based similarity, may be ineffective.

To address this, we utilize a predefined set of TOT-specific linguistic codes introduced by Arguello et al. [4]. These handcrafted codes provide sentence-level annotations of TOT queries in the Movie domain, categorizing linguistic phenomena into eight top-level groups: 'movie', 'context', 'previous-search', 'social', 'uncertainty', 'opinion', 'emotion', and 'relative-comparison'. By leveraging this framework, we compare the linguistic distribution of codes between CQA-based and elicited queries rather than relying on direct semantic or lexical overlap.

To conduct this analysis, we annotate our elicited TOT queries in the Movie domain using these linguistic codes and compare the percentage distributions of codes found in CQA-based and elicited queries. To automate this process, we develop a language model-based automatic code annotator, prompting GPT-4o-mini to produce JSON-formatted sentence-level annotations.[3]

Before applying this annotator to LLM- and human-elicited queries, we first validate its performance on the MS-TOT dataset, where sentence-level gold annotations are available. We evaluate the annotator's performance by computing precision and recall as prediction accuracy measures. Additionally, to assess annotation quality at a broader level, we compute query-level precision and recall, which measure the accuracy of identifying unique codes that appear within a multi-sentence query.

To further validate the annotator, we compare the distribution of annotated codes against the gold annotations using Earth Mover's Distance (EMD), which quantifies how different the predicted code distribution is from the reference distribution. Figure 2 presents the validation results of our automatic annotator on the MS-TOT dataset. Our evaluation shows that the annotator achieves sufficiently high accuracy and low EMD, confirming its suitability as an automatic labeler.

With this validation, we apply the automatic annotator to both LLM- and human-elicited queries in the Movie domain and analyze how their linguistic code distributions compare to those in CQA-based queries. However, we conduct this linguistic similarity validation only in the Movie domain, as there are no existing comprehensive handcrafted linguistic codes available for the Landmark and Person domains.
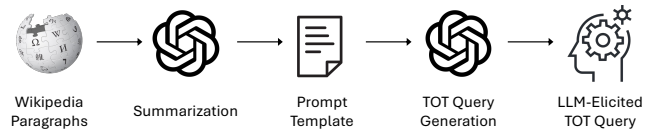
4 TOT Query Elicitation from LLMs

4.1 Synthetic Query Generation

Figure 3 illustrates the LLM-elicited TOT query generation process. Given an arbitrary entity for which we want to generate a TOT query, we retrieve the entity's Wikipedia page, summarize it into a few paragraphs, and construct a prompt template using the summary. This prompt is then used to query an LLM to generate a synthetic TOT query.

The query generation process consists of the following steps:

(1) *Sampling a target entity*:
A random entity is selected as the target for TOT query generation. In the Movie domain, the MS-TOT test set serves as the pool of entities, while for the Landmark and Person domains, we manually curated the entity pool.

Figure 3: LLM-elicited TOT query generation process. Given an arbitrary entity for which we want to generate a TOT query, we retrieve its Wikipedia page, summarize it into a few paragraphs, and construct a prompt template using the summary. This prompt is then used to query an LLM to generate a synthetic TOT query.

(2) *Retrieving and summarizing the entity's Wikipedia page*:
The entity's Wikipedia page is retrieved and used to construct a summarization prompt for the GPT-4o model.[4] To ensure the content fits within the model's token limit, the page is truncated if necessary. The model processes the page into a two-paragraph summary. For Movie domain entities, we check for a dedicated *Plot* section and ensure its content is included in the summarization prompt, as it often provides essential context for generating realistic TOT queries.

(3) *Constructing a domain- and entity-specific prompt*:
Following summarization, a domain- and entity-specific prompt is constructed, incorporating the summarized information to generate a TOT query. The same GPT-4o model is used for query elicitation. We tested various prompt configurations and temperature settings to identify the best-performing strategy in the Movie domain, then applied it to the Landmark and Person domains.[5]

(4) *Ensuring entity name anonymity in the generated query*:
An essential consideration in the query generation process is ensuring that the generated TOT query does not contain the target entity's name. To address this, we implement a name-checking step: if the generated query includes the entity name, we retry the generation using the same prompt, with a maximum of three retries before discarding the query.

4.2 Results and Analyses

RQ 1: Can we elicit TOT queries from LLMs for effective simulated evaluation of TOT retrieval systems?

We tested various configurations to identify the most effective TOT query generation approach. The final prompt structure was determined through the extensive testing of different prompt templates, system roles, instruction types, and temperature settings, as well as evaluating the inclusion of Wikipedia summaries (Table 1). For each configuration, queries were generated using entities from the test set of the TREC TOT Track 2023 [2], derived from the MS-TOT dataset [4], and evaluated based on system rank correlation and linguistic similarity.

The top-performing strategies (Experiment ID 12 and 13) utilized a role-playing scenario, where the LLM assumed the role of a TOT searcher. The prompt began with:

---

[3]GPT model's temperature is set to 0 for reproducibility and consistency.

[4]GPT4o-2024-05-13

[5]Details on the different prompt versions we tested are available at https://github.com/kimdanny/llm-tot-query-elicitation.

	Prompt & Model Design					Validation Results		
						MRR-based	NDCG-based	Linguistic Similarity
ID	Prompt Template	System Role	Wikipedia Summary	Instruction Type	Temperature	τ/r ↑	τ/r ↑	EMD ↓
1	V0	Writing Assistant	w/o	9 rules	0.5	0.2641 / 0.6101	0.3092 / 0.4247	0.0899
2	V1	Writing Assistant	w/	9 rules	0.5	0.4923 / 0.5692	0.5077 / 0.6475	0.1012
3	V2	Searcher role play	w/	6-shot	0.5	0.4632 / 0.5391	0.4430 / 0.5492	0.0299
4	V2	Searcher role play	w/	6-shot	0.7	0.4538 / 0.4234	0.4310 / 0.4650	0.0276
5	V3	Searcher role play	w/o	6-shot	0.5	0.1524 / 0.5293	0.2045 / 0.5995	0.0584
6	V3	Searcher role play	w/o	6-shot	0.7	0.1440*/ 0.5173	0.1960*/ 0.5942	0.0534
7	V4	Searcher role play	w/	13 rules	0.3	0.6927 / 0.9143	0.6757 / 0.8975	0.0490
8	V4	Searcher role play	w/	13 rules	0.5	0.6463 / 0.9173	0.6235 / 0.8998	0.0542
9	V4	Searcher role play	w/	13 rules	0.7	0.6472 / 0.9336	0.6785 / 0.9303	0.0577
10	V5	Searcher role play	w/	14 rules	0.1	0.6719 / 0.9388	0.6500 / 0.9375	0.0542
11	V5	Searcher role play	w/	14 rules	0.3	0.6558 / 0.9396	0.6728 / 0.9342	0.0620
12	V5	Searcher role play	w/	14 rules	0.5	0.7013 / **0.9536**	0.7127 / **0.9500**	0.0590
13	V6	Searcher role play	w/	7 Musts + 7 Coulds	0.3	**0.7573** / 0.9166	**0.7194** / 0.8973	**0.0264**

Table 1: Experiments on LLM-elicited query generation in the Movie domain. τ and r represent Kendall's Tau and Pearson's r correlation values, respectively, while EMD (Earth Mover's Distance) measures linguistic similarity between CQA-based and LLM-elicited queries. All correlation values have a p-value < 0.01, except for those marked with (*).

> Let's do a role play. You are now a person who watched a movie {TOT Entity} a long time ago and forgot the movie's name.

This setup encouraged the LLM to generate queries that better captured real TOT search behavior, leading to improved retrieval rankings.

Specifically, the prompt included a summarized Wikipedia entry followed by 14 guideline rules, ensuring the searcher role-play aligned with real user behavior. For example, rules such as *"Reflect the imperfect nature of memory with phrases that express doubt or mixed recollections"* and *"Introduce a few incorrect or mixed-up details to make the recollection seem more realistic and challenging to pinpoint"* helped simulate the uncertainty and distorted memory typical of TOT searchers. Similarly, *"Share a personal anecdote related to when or with whom you watched the movie"* and *"Focus on sensory details such as the overall mood, sounds, or emotional impact of the movie, using vivid descriptions"* aimed to incorporate the personal and emotional elements often found in real TOT queries.

While Experiment 12 achieved the highest Pearson correlation, Experiment 13 resulted in the highest Kendall's Tau correlation and the lowest linguistic distance from CQA-based queries. A notable comparison is between Experiment IDs 11 and 13, which share the same temperature and 14 rules. The key difference lies in the rule structuring: while ID 11 presented all 14 rules equally, ID 13 divided them into 7 "must" rules and 7 "could" rules. This slight difference led to a significant increase in Kendall's Tau and linguistic similarity, despite a minor drop in Pearson correlation.

Through the experiments of different setups, we derived several key insights. Role-playing as a TOT searcher with explicit behavioral guidelines led to more realistically structured TOT queries. Providing generation rules proved more effective than few-shot

prompting, as seen when transitioning from Prompt Version 2 (few-shot) to Prompt Version 4 (rule-based), which significantly improved validation scores.

Including a Wikipedia summary in the prompt substantially increased system rank correlation, indicating that structured background knowledge helps the LLM generate queries more aligned with real TOT queries. Additionally, lower temperature settings (0.3, 0.5) outperformed higher ones (0.7), contrary to our initial hypothesis that increased randomness would better simulate the stochastic nature of memory retrieval. The improved performance at lower temperatures likely stems from the model following the defined generation rules more closely.

Additionally, linguistic similarity analysis using Earth Mover's Distance (EMD) resulted in a value of 0.0264 (Figure 4) in Experiment 13, showing the lowest value.

Based on findings across the two validation methods, we empirically conclude that LLM-generated TOT queries are an effective approach for designing simulated evaluations of TOT retrieval systems.

RQ 3-1: Can the LLM-elicitation-based evaluation methods be used to other domains underrepresented in CQA-collected test collections?

Based on our findings, we adopted the Experiment ID 13 configuration for the final dataset generation, extending its application to the underrepresented Landmark and Person domains. To adapt the Prompt V6 setup for these domains, we adjusted entity descriptions while preserving the core prompt structure. For instance, the rule *"Share a personal anecdote related to when or with whom you watched the movie"* was modified to *"Share a personal anecdote related to the person"* to align with the context of the Person domain.

To validate the generated queries, we compared them against CQA-collected queries. Since existing TOT test collections are largely focused on the Movie domain, no established test collection
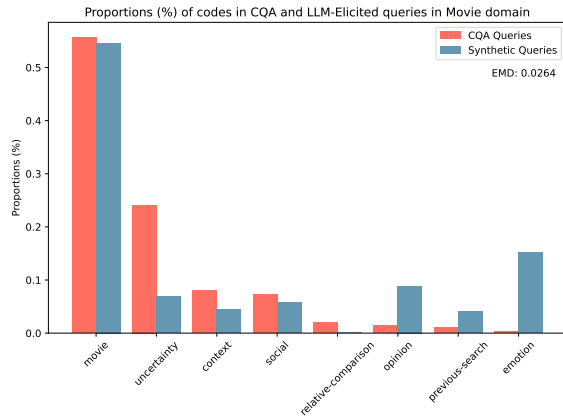
Figure 4: Proportions of labeled codes in CQA and LLM-elicited queries in the Movie domain using Prompt Version V6 (Experiment ID 13). This setup achieved the lowest frequency distribution difference (EMD) among all experiments.

	MRR-based	NDCG-based
Domain	τ/r ↑	τ/r ↑
Landmark	0.5984 / 0.8227	0.6967 / 0.9142
Person	0.6362 / 0.7231	0.5691 / 0.7996

Table 2: Kendall's Tau and Pearson's r correlation values between CQA-based queries and LLM-elicited queries in the Person and Landmark domains. τ and r represent Kendall's Tau and Pearson's r correlation values, respectively. All correlation values have a p-value < 0.01.

exists for the Landmark and Person domains. To address this, we hand-curated a test collection, resulting in 22 Landmark-domain queries and 70 Person-domain queries from the Reddit's subreddit, called */r/tipofmytongue*.

Table 2 presents the system rank correlation validation results for both domains. While the validation scores are not as strong as those observed in the Movie domain, they still demonstrate reasonably high values. This suggests that LLM-elicited query generation methods can be extended to underrepresented domains and could potentially be developed into generalizable LLM-elicitation techniques across multiple domains.

5 TOT Query Elicitation from Human

The study of the tip-of-the-tongue phenomenon has been a long-standing area of research in psychology [13, 22]. Many studies have attempted to induce TOT states in participants using auditory [33] or visual stimuli [35], enabling researchers to examine recognizability (whether a subject recognizes an entity from the stimulus) and retrievability (whether they can recall the entity's name or title).

Building on these methodologies, we employ visual stimuli in the Movie, Landmark, and Person domains to develop an interface that places participants in a TOT state and allows them to compose TOT queries about the entities they struggle to recall.

In this section, we describe the design process of our interface for eliciting human-written TOT queries from trained contracted

participants, along with an analysis of the collected human-elicited queries.

5.1 Visual Stimuli Collection

To obtain images in Movie, Landmark, and Person domains, we collected entities from Wikipedia, selecting pages based on their infobox namespace and the presence of images.[6] For the Person and Landmark domains, we directly collected images from Wikipedia pages. However, for the Movie domain, we identified linked IMDb or TMDB IDs on the Wikipedia pages and retrieved high-quality backdrop or still-cut images from TMDB. This approach was necessary because Wikipedia's movie images were predominantly posters, which often contain the movie title and cannot be used to elicit a TOT state. TMDB backdrops, on the other hand, provide spoiler-free, high-quality visuals suitable for this purpose.[7] For the Landmark domain, we selected "Place" as a top-level category, and for the Person domain, we chose "Person".

Selecting appropriate infobox namespaces within these top-level categories presented challenges, as the collected images needed to be *unambiguous* and *popular* enough for human participants to recognize the entities. For example, in Landmark domain, images of random parks often depicted generic green spaces, making them ambiguous, whereas iconic parks were more likely to be recognizable. For the Person domain, ambiguity was defined as the presence of multiple people in the image, which could confuse participants. To ensure the quality of the stimuli, two authors rated infobox namespaces under the "Place" and "Person" categories on a three-point scale:

- 0: Most images would be ambiguous or unrecognizable.
- 1: Some images might be ambiguous.
- 2: Most images would be unambiguous and recognizable.

Given that our participants were based in the United States, this factor was also considered during the labeling process.

After completing the rating and ranking process, we selected one infobox namespace ('film') for Movie, 94 for Landmark (e.g., 'amusement park', 'bridge', 'ancient site'), and 123 for Person (e.g., 'comedian', 'sportsperson', 'academic').

Next, we collected Wikipedia pages associated with the selected infobox namespaces and sorted them by page view counts, using the Wikipedia database dump from October 1, 2023.[8] We selected the top 20% of pages by view count, as less popular entities were deemed unlikely to be recognized by participants.

Finally, we manually filtered the collected images to remove unsuitable content. For the Movie domain, we excluded inappropriate images, such as those featuring nudity or excessive gore, and posters containing the movie title. For the Landmark domain, we removed images where the entity name was visible, such as station names in train station pictures. For the Person domain, we excluded images with multiple individuals.[9] After filtering, we finalized a visual stimuli collection of 1,687 Movie entities, 1,946 Person entities, and 330 Landmark entities, all associated with Wikipedia pages,

---

[6]List of infoboxes: https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes.
[7]https://www.themoviedb.org/bible/image/
[8]https://dumps.wikimedia.org/enwiki/
[9]Detailed image deselection criteria can be found at https://github.com/kimdanny/human-tot-query-elicitation-mturk/blob/main/image_deselection_criteria.md.

ensuring a diverse and high-quality set of stimuli for eliciting TOT queries.

5.2 Interface Design

We designed the flow of an interface as illustrated in Figure 5.

5.2.1 Phase 1 (Recognizability). In this phase, the system presents an image drawn from the pool of visual stimuli, asking participants whether they recognize the movie, landmark, or person depicted in the image, depending on the domain. The process repeats, drawing new images until the participant confirms recognition of an entity.

To ensure a balanced selection process, images in each domain were grouped into 20 buckets based on their popularity, measured by Wikipedia page view counts. When selecting an image, the system draws one random image per bucket in sequential order. The three domains—Movie, Landmark, and Person—are covered in a round-robin fashion, ensuring that images from different domains are presented in a balanced manner.

This approach minimizes the likelihood of a participant encountering the same image multiple times within a session and prevents from disproportionately drawing easily recognizable entities, leading to a more effective and controlled TOT induction process.

5.2.2 Phase 2 (Retrievability). In this phase, we assess whether the participant can recall the name or title of the entity they recognized in Phase 1. If they cannot recall the name, it indicates that they are in a TOT state, and the interface transitions to Phase 3.

If they can recall the name, they are prompted to submit it, and the interface moves directly to Phase 4.

5.2.3 Phase 3 (TOT Query Collection). If a participant reaches Phase 3, they are prompted to write a TOT query describing the entity they recognized but could not retrieve.

To ensure high-quality queries, we analyzed word count distributions in the MS-TOT dataset and determined that a minimum of 300 characters is necessary to effectively capture the complexity of TOT queries. To encourage longer responses, we followed the approach of Agapie et al. [1] and implemented a color-coded progress bar that visually indicates the current query length. The bar starts as red when the query is below 200 characters, turns yellow at 200 characters, and finally turns green once it reaches 300+ characters, signaling that the recommended length has been met.

Once the participant submits their TOT query, the interface transitions to Phase 4.

5.2.4 Phase 4 (Entity Confirmation). In Phase 4, the system displays the Wikipedia page of the entity, allowing the participant to verify whether the entity they wrote a TOT query about or successfully retrieved is correct. This confirmation step is crucial for determining whether the collected TOT query is valid.

After completing Phase 4, the interface loops back to Phase 1, where a new image is presented to continue the process.

5.3 Results and Analyses

Seven expert contracted participants interacted with the designed interface to generate human-elicited TOT queries.

Table 3 shows the number of instances recorded at each phase for the Movie, Landmark, and Person domains, along with their

Phase	Movie	Landmark	Person	Total
Phase 1: Recognizability				
Yes	672	231	1279	2182
No	5513	2079	5159	12751
Phase 2: Retrievability				
Yes	395	82	816	1293
Yes (Correct)	314	23	721	1058
Yes (Incorrect)	81	59	95	235
No	273	149	462	884
Phase 3: Query Collection	245	119	413	777
Phase 4: Entity Confirmation				
Yes	178	57	349	584
No	61	61	59	181
N/A	6	1	5	12

Table 3: Statistics of human-elicited TOT query collection across different phases of the interface.

totals. "Phase 1 - Yes" and "Phase 1 - No" indicate whether participants recognized the presented entity. "Phase 2 - Yes" represents cases where participants recalled the entity's name, with "Correct" and "Incorrect" indicating the accuracy of their recall. "Phase 2 - No" corresponds to participants who could not retrieve the entity's name, leading to TOT query collection in Phase 3. "Phase 4 - Yes" and "Phase 4 - No" denote whether the retrieved entity was correctly confirmed, while "Phase 4 - N/A" represents cases where confirmation was not applicable.[10]

From this human TOT query collection process, we obtained 178 Movie queries, 57 Landmark queries, and 349 Person queries, resulting in a total of 584 human-written TOT queries across three domains. The following is an example of a Landmark query:

> **Human-Elicited Landmark Query**:
> i've seen a picture of this landmark but I can't quite remember where it is. it might have a garden area inside of it. you can walk from the bottom around and around to the top if i remember correctly. i don't think it's for housing. it's a sculpture. possibly located in new york or at least in the US.
> **Correct Answer: Vessel**

As shown in the example, we qualitatively identified common linguistic patterns, such as expressions of uncertainty and distorted memories, within the collected queries.

	Movie	Landmark	Person
Recognizability	0.11	0.10	0.20
Retrievability	0.05	0.01	0.17

Table 4: Recognizability (Phase 1 - Yes / Phase 1 - Total) and Retrievability (Phase 2 - Yes Correct / Phase 1 - Total) in the human TOT query collection.

Table 4 presents the recognizability and retrievability of visual stimuli in our human experiments. Recognizability is measured as the ratio of participants recognizing an entity (Phase 1 - Yes) to the total instances in Phase 1. Retrievability is measured as the

---

[10]Due to session restarts or participants logging off, the numbers in "Phase 2 - No" and "Phase 3" do not always match exactly. Similarly, the sum of "Phase 2 - Yes" and "Phase 2 - No" does not always correspond to the number of "Phase 1 - Yes".

Phase 1
Recognizability

Phase 2
Retrievability

Phase 3
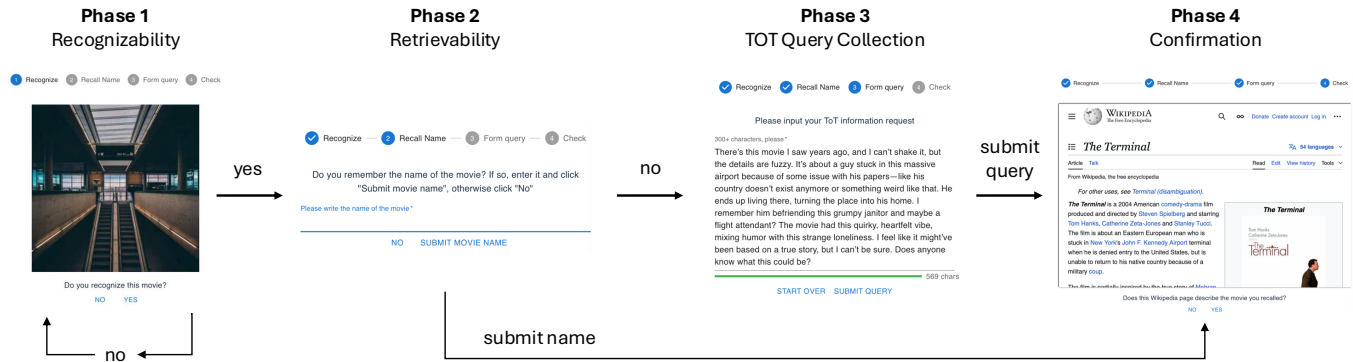TOT Query Collection

Phase 4
Confirmation

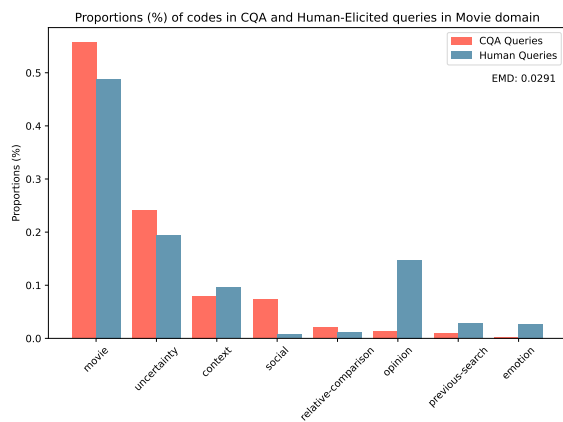Figure 5: Flowchart of the Human TOT Query Elicitation Interface.

Figure 6: Proportions of labeled codes in CQA and human-elicited queries in the Movie domain, showing a low difference in frequency distribution (EMD).

Domain	MRR-based $\tau/r \uparrow$	NDCG-based $\tau/r \uparrow$
Movie	0.6111 / 0.5882	0.6438 / 0.6889

Table 5: Kendall's Tau and Pearson's r correlation values between CQA-collected queries (TOMT-KIS) and human-elicited queries in the Movie domain. τ and r represent Kendall's Tau and Pearson's r correlation values, respectively. All correlation values have a p-value < 0.01.

RQ 2: Can we elicit TOT queries from humans for effective simulated evaluation of TOT retrieval systems?

Our results indicate that human-elicited TOT queries can serve as an effective resource for simulated evaluation of TOT retrieval systems. The system rank correlation between human-elicited queries and CQA-collected queries (TOMT-KIS) in the Movie domain shows high agreement (Table 5). Additionally, linguistic similarity analysis using Earth Mover's Distance (EMD) resulted in a value of 0.0291 (Figure 6), which is comparable to the lowest EMD achieved during the LLM-elicitation experiments.

The higher proportion of opinion-related expressions in human-elicited queries may be influenced by differences in query intent. In CQA settings, users often structure their queries to maximize answerability, prioritizing factual details over personal reflections. In contrast, our elicitation method placed participants in a TOT state without the expectation of receiving an answer from community, which may have encouraged them to articulate their thoughts more freely. This introspective process could have naturally led to a greater inclusion of subjective impressions and emotions in their queries.

RQ 3-2: Can the human-elicitation-based evaluation methods be used to other domains underrepresented in CQA-collected test collections?

Since validation was restricted to the Movie domain due to the lack of sufficient CQA-collected queries in other domains, additional validation methods may be necessary to fully address the research question. However, the low EMD score (0.0291) in the Movie domain suggests that human-elicitation methods effectively capture authentic TOT queries, at least within this domain. Given that the elicitation process was designed to be domain-agnostic, follows established psychological principles of TOT states, and does

proportion of correctly recalled entity names (Phase 2 - Yes Correct) relative to all instances in Phase 1. These values provide insights into how well different domains induce TOT states in participants with our visual stimuli set.

To validate with system rank correlation, we needed overlapping entities between CQA data and our human-elicited TOT dataset. Since MS-TOT and manually collected Reddit posts had negligible overlap in the Landmark and Person domains, we turned to TOMT-KIS for matching entities.

As TOMT-KIS lacks explicit entity names, we matched entities by checking if the target name appeared as a substring in the answer posts. Only one valid match ("Fort Ticonderoga") was found in the Landmark domain, and nearly all matched queries in the Person domain were about movies or songs, as TOMT-KIS did not collect person-specific queries. Thus, we had to exclud both domains from validation. For the Movie domain, we filtered TOMT-KIS queries for relevance by requiring the word "movie" while excluding "song", removing answers with multiple guesses, eliminating queries based solely on video links, and selecting the longest text for duplicate identifiers. This process yielded 303 Movie-domain queries, which we used for system rank correlation validation.

not rely on movie-specific heuristics, these methods are likely to generalize to underrepresented domains, making them a promising direction for future TOT retrieval evaluations.

6   Discussion

Resource Contribution. Using the LLM-elicitation method from Experiment ID 13 (Prompt Version 6), we generated synthetic TOT queries for entities collected through the visual stimuli selection process, resulting in 1,687 queries in the Movie domain, 330 in the Landmark domain, and 1,946 in the Person domain. Additionally, through the human-elicitation method, we collected 584 human-elicited TOT queries spanning all three domains.

The full release of these queries is scheduled for the TREC 2025 TOT track, where they will be included as part of the official test collection. However, for TREC 2024 [3], and in this work, we have released 450 synthetic queries (150 per domain) from the full set of generated queries. Each query in the dataset is accompanied by its corresponding Wikidata ID, domain name, and entity name, ensuring clear entity association for retrieval experiments and analysis. Alongside these queries, we provide the source code for query generation and experimentation, as well as the visual stimuli entity set with corresponding image URLs and Wikidata ID. We also release the MTurk-based human query collection interface, allowing researchers to replicate or extend the human TOT query elicitation process.

Availability. At the time of review, LLM-elicited queries are publicly available as part of the TREC 2024 TOT track [3] test collection and can also be accessed at the track website.[11] The human-elicited queries will be released as part of the TREC 2025 TOT track, aligning with the SIGIR 2025 conference. Although the human-elicited queries are not yet publicly available, we have released the source code for the human query collection interface used in pilot testing on Amazon MTurk.[12] This allows researchers to explore and reproduce the query elicitation process for future development.

Both datasets are, and will continue to be, freely available under open licensing terms, ensuring unrestricted access for academic researchers and industry practitioners to support research and development in TOT retrieval.

Utility. The resource is well-documented and designed for easy integration into retrieval experiments. Queries are provided in pure text format for compatibility with retrieval models, and a baseline implementation with tools for data loading and retrieval is available in the public repositories. Additionally, this paper details the data provenance, processing, and experimentation steps, ensuring that future researchers can expand the dataset or adopt the TOT query elicitation method for other domains, supporting reproducibility and innovation.

Novelty and Predicted Impact. Our work represents a major shift in TOT query collection methodology, moving beyond CQA-based datasets to LLM- and human-elicited queries, providing a scalable and flexible alternative for TOT dataset creation. Unlike previous approaches, our method eliminates the need for manual labeling, avoids data restrictions, and mitigates domain skewness

in CQA datasets, which overrepresent casual leisure topics like movies and books. By incorporating underrepresented domains such as Person and Landmark, our dataset extends beyond traditional leisure-focused queries, supporting the development of general-domain TOT retrieval systems while enabling simulated evaluation independent of CQA constraints.

While TOT retrieval builds on known-item retrieval research, our dataset and methodology provide new tools for evaluating and training retrieval systems to handle TOT queries more effectively. We anticipate its long-term value and plan to incrementally expand domain coverage through future TREC tracks, ensuring broader applicability and comprehensive evaluation in TOT retrieval research.

Methodological Implications. While our findings show that both LLM- and human-elicited queries are effective, they serve complementary roles: LLMs offer scalability and efficiency, whereas human queries may provide authentic linguistic patterns and user behaviors. A hybrid approach can balance dataset diversity and efficiency, leading to more comprehensive evaluations of TOT retrieval systems.

Beyond TOT retrieval, our elicitation methods could support vague or exploratory search scenarios, where users struggle to articulate precise queries. Additionally, LLM-based query generation could aid low-resource domains, simulating real-world search behaviors where query logs are scarce or unavailable, broadening its impact across information retrieval research.

Limitation and Future Work. A limitation of our current LLM-elicitation method is that prompts are domain-specific, limiting their generalizability. Future work should develop generalized prompting strategies to elicit TOT queries across diverse search contexts without extensive manual tuning. Additionally, expanding the methodology to multi-turn or multi-modal interactions [14] could better reflect real-world TOT search behavior, where users iteratively refine queries as they recover missing information.

7   Conclusion

This work introduces a novel approach to TOT query elicitation, leveraging LLMs and human participants to move beyond the limitations of CQA-based datasets. Through system rank correlation and linguistic similarity validation, we demonstrate that LLM- and human-elicited queries can effectively support the simulated evaluation of TOT retrieval systems. Our findings highlight the potential for expanding TOT retrieval research into underrepresented domains while ensuring scalability and reproducibility. The released datasets and source code provide a foundation for future research, enabling further advancements in TOT retrieval evaluation and system development.

Acknowledgments

---

[11] https://github.com/kimdanny/llm-tot-query-elicitation
[12] https://github.com/kimdanny/human-tot-query-elicitation-mturk

References

[1] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. 2013. Leading people to longer queries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3019–3022. doi:10.1145/2470654.2481418

[2] Jaime Arguello, Samarth Bhargav, Fernando Diaz, Evangelos Kanoulas, and Bhaskar Mitra. 2023. Overview of the trec 2023 tip-of-the-tongue track. TREC.

[3] 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 *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024), Gaithersburg, MD, USA, November 15-18, 2024.* https://trec.nist.gov/pubs/trec33/papers/Overview_tot.pdf

[4] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 5–14. doi:10.1145/3406522.3446021

[5] Leif Azzopardi and Maarten de Rijke. 2006. Automatic construction of known-item finding test beds. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Seattle, Washington, USA) *(SIGIR '06)*. Association for Computing Machinery, New York, NY, USA, 603–604. doi:10.1145/1148170.1148276

[6] Leif Azzopardi, Maarten de Rijke, and Krisztian Balog. 2007. Building simulated queries for known-item topics: an analysis using six european languages. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 455–462. doi:10.1145/1277741.1277820

[7] Krisztian Balog, Leif Azzopardi, Jaap Kamps, and Maarten De Rijke. 2006. Overview of webclef 2006. In *Workshop of the Cross-Language Evaluation Forum for European Languages.* Springer, 803–819.

[8] Michael Barbaro, Tom Zeller, and Saul Hansell. 2006. A face is exposed for AOL searcher no. 4417749. *New York Times* 9, 2008 (2006), 8.

[9] Samarth Bhargav, Anne Schuth, and Claudia Hauff. 2023. When the Music Stops: Tip-of-the-Tongue Retrieval for Music. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2506–2510. doi:10.1145/3539618.3592086

[10] Samarth Bhargav, Georgios Sidiropoulos, and Evangelos Kanoulas. 2022. 'It's on the tip of my tongue': A new Dataset for Known-Item Retrieval. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 48–56. doi:10.1145/3488560.3498421

[11] Toine Bogers, Maria Gäde, Mark Hall, Marijn Koolen, Vivien Petras, and Mette Skov. 2025. Exploring the Zero-Shot Known-Item Retrieval Capabilities of LLMs for Casual Leisure Information Needs. In *CHIIR 2025: Proceedings of the 2025 Conference on Human Information Interaction and Retrieval.*

[12] Luís Borges, Rohan Jha, Jamie Callan, and Bruno Martins. 2024. Generalizable Tip-of-the-Tongue Retrieval with LLM Re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2437–2441. doi:10.1145/3626772.3657917

[13] Deborah M Burke, Donald G MacKay, Joanna S Worthley, and Elizabeth Wade. 1991. On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of memory and language* 30, 5 (1991), 542–579.

[14] Sky CH-Wang, Darshan Deshpande, Smaranda Muresan, Anand Kannappan, and Rebecca Qian. 2025. Browsing Lost Unformed Recollections: A Benchmark for Tip-of-the-Tongue Search and Reasoning. *arXiv preprint arXiv:2503.19193* (2025).

[15] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 72–79. doi:10.1145/860435.860451

[16] David Elsweiler, Mark Baillie, and Ian Ruthven. 2008. Exploring memory in email refinding. *ACM Trans. Inf. Syst.* 26, 4, Article 21 (Oct. 2008), 36 pages. doi:10.1145/1402256.1402260

[17] David Elsweiler, David E. Losada, José C. Toucedo, and Ronald T. Fernandez. 2011. Seeding simulated queries with user-study data for personal search evaluation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Beijing, China) *(SIGIR '11)*. Association for Computing Machinery, New York, NY, USA, 25–34. doi:10.1145/2009916.2009924

[18] David Elsweiler, Ian Ruthven, and Christopher Jones. 2007. Towards memory supporting personal information management tools. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 924–946.

[19] David Elsweiler, Max L Wilson, and Brian Kirkegaard Lunn. 2011. Understanding casual-leisure information behaviour. In *New directions in information behaviour.* Vol. 1. Emerald Group Publishing Limited, 211–241.

[20] Maik Fröbe, Eric Oliver Schmidt, and Matthias Hagen. 2023. A Large-Scale Dataset for Known-Item Question Performance Prediction.. In *QPP++@ ECIR.* 13–19.

[21] Matthias Hagen, Daniel Wägner, and Benno Stein. 2015. A corpus of realistic known-item topics with associated web pages in the ClueWeb09. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015, Vienna, Austria, March 29-April 2, 2015. Proceedings 37.* Springer, 513–525.

[22] Gregory V Jones. 1989. Back to Woodworth: Role of interlopers in the tip-of-the-tongue phenomenon. *Memory & Cognition* 17 (1989), 69–76.

[23] Ida Kathrine Hammeleff Jørgensen and Toine Bogers. 2020. "Kinda like The Sims... But with ghosts?": A Qualitative Analysis of Video Game Re-finding Requests on Reddit. In *Proceedings of the 15th International Conference on the Foundations of Digital Games* (Bugibba, Malta) *(FDG '20)*. Association for Computing Machinery, New York, NY, USA, Article 40, 4 pages. doi:10.1145/3402942.3402971

[24] Jinyoung Kim and W. Bruce Croft. 2009. Retrieval experiments using pseudo-desktop collections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (Hong Kong, China) *(CIKM '09)*. Association for Computing Machinery, New York, NY, USA, 1297–1306. doi:10.1145/1645953.1646117

[25] Jin Ha Lee, Allen Renear, and Linda C Smith. 2006. Known-item search: Variations on a concept. *Proceedings of the american society for information science and technology* 43, 1 (2006), 1–17.

[26] Kevin Lin, Kyle Lo, Joseph Gonzalez, and Dan Klein. 2023. Decomposing Complex Queries for Tip-of-the-tongue Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5521–5533. doi:10.18653/v1/2023.findings-emnlp.367

[27] Florian Meier, Toine Bogers, Maria Gäde, and Line Ebdrup Thomsen. 2021. Towards Understanding Complex Known-Item Requests on Reddit. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (Virtual Event, USA) *(HT '21)*. Association for Computing Machinery, New York, NY, USA, 143–154. doi:10.1145/3465336.3475096

[28] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. From doc2query to docTTTTTquery. *Online preprint* 6, 2 (2019).

[29] Paul Ogilvie and Jamie Callan. 2003. Combining document representations for known-item search. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* (Toronto, Canada) *(SIGIR '03)*. Association for Computing Machinery, New York, NY, USA, 143–150. doi:10.1145/860435.860463

[30] Rohan Phanse, Gabrielle Kaili-May Liu, and Arman Cohan. 2024. Yale NLP at TREC 2024: Tip-of-the-Tongue Track. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024), Gaithersburg, MD, USA, November 15-18, 2024.* https://trec.nist.gov/pubs/trec33/papers/yalenlp.tot.pdf

[31] Barbara Poblete, Myra Spiliopoulou, and Ricardo Baeza-Yates. 2010. Privacy-preserving query log mining for business confidentiality protection. *ACM Trans. Web* 4, 3, Article 10 (July 2010), 26 pages. doi:10.1145/1806916.1806919

[32] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2647–2651. doi:10.1145/3626772.3657942

[33] David M Reefer, Mary K Kevari, and Daniel LF Kramer. 1995. Name that tune: Eliciting the tip-of-the-tongue experience using auditory stimuli. *Psychological Reports* 77, 3_suppl (1995), 1379–1390.

[34] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, 109–126.

[35] Daniel Tranel, Nonso Enekwechi, and Ken Manzel. 2005. A test for measuring recognition and naming of landmarks. *Journal of Clinical and Experimental Neuropsychology* 27, 1 (2005), 102–126.

[36] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 485, 13 pages.

[37] Martin White. 2015. *Enterprise search: enhancing business performance.* " O'Reilly Media, Inc.".

[38] Bernard P Zeigler, Herbert Praehofer, and Tag Gon Kim. 2000. *Theory of modeling and simulation.* Academic press.

[39] Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 334–342. doi:10.1145/383952.384019