Multilingual and Domain-Agnostic Tip-of-the-Tongue Query Generation for Simulated Evaluation

Xuhong He
Carnegie Mellon University
Pittsburgh, PA, USA
xuhongh@cs.cmu.edu

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

Maik Fröbe
Friedrich-Schiller-Universität Jena
Jena, Germany
maik.froebe@uni-jena.de

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

Bhaskar Mitra
Independent Researcher
Tiohtià:ke / Montréal, QC, Canada
bhaskar.mitra@acm.org

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

Abstract

Tip-of-the-Tongue (ToT) retrieval benchmarks have largely focused on English, limiting their applicability to multilingual information access. In this work, we construct multilingual ToT test collections for Chinese, Japanese, Korean, and English, using an LLM-based query simulation framework. We systematically study how prompt language and source document language affect the fidelity of simulated ToT queries, validating synthetic queries through system rank correlation against real user queries. Our results show that effective ToT simulation requires language-aware design choices: non-English language sources are generally important, while English Wikipedia can be beneficial when non-English sources provide insufficient information for query generation. Based on these findings, we release four ToT test collections with 5,000 queries per language across multiple domains. This work provides the first large-scale multilingual ToT benchmark and offers practical guidance for constructing realistic ToT datasets beyond English. [1] [2]

CCS Concepts

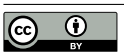
• **Information systems** → **Test collections**; **Multilingual and cross-lingual retrieval**.

Keywords

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

ACM Reference Format:
Xuhong He, To Eun Kim, Maik Fröbe, Jaime Arguello, Bhaskar Mitra, and Fernando Diaz. 2026. Multilingual and Domain-Agnostic Tip-of-the-Tongue Query Generation for Simulated Evaluation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3805712.3808626

[1] https://github.com/kimdanny/ntcir-19-tot
[2] https://zenodo.org/records/18777084



SIGIR '26, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3808626

1 Introduction

Tip-of-the-tongue (ToT) retrieval is a form of known-item retrieval in which searchers attempt to refind a specific item but are unable to recall a reliable identifying term [4]. In this state, users often remember partial, indirect, or noisy information about the target item, such as contextual details, personal experiences, or incorrect attributes. As a result, ToT queries tend to be verbose, frequently containing false memories, subjective descriptions, and exclusion criteria rather than precise identifiers. These characteristics distinguish ToT retrieval from conventional ad hoc search and pose unique challenges for retrieval systems [1].

Supporting ToT retrieval is important because such queries constitute a substantial portion of real-world search behavior [28] and are associated with high levels of user frustration when retrieval fails [17]. ToT search behavior is widely observed in community question answering (CQA) platforms such as Reddit, where users publicly articulate their refinding attempts across a wide range of topics [20]. This prevalence and diversity have motivated the development of benchmark datasets for ToT retrieval, most notably through shared evaluation efforts in the TREC forum [1–3]. More recently, advances in large language models (LLMs) have enabled scalable ToT query simulation, allowing evaluation to move beyond costly manual query collections [21].

Despite these efforts, existing ToT retrieval benchmarks remain limited in both language and domain coverage. Most prior datasets and evaluation efforts are English-centric, even though ToT search behavior and retrieval failures are not confined to English-speaking users and occur broadly across CQA platforms in many languages.[3] Furthermore, while some recent studies [2, 21] have begun to extend ToT evaluation beyond casual leisure domains, the majority of existing datasets still exhibit strong domain skewness toward entertainment-related content such as movies and music [4, 8, 11, 18]. These limitations restrict the generalizability of current ToT retrieval evaluations.

In light of these gaps, we propose a multilingual and domain-agnostic simulated ToT retrieval evaluation method with rigorous validation of the simulation process. Our approach leverages LLM-based query generation to systematically expand ToT evaluation beyond English, covering three additional East Asian languages: Chinese, Japanese, and Korean (hereafter CJK). At the same time, the

[3] For example, Naver (Korean), Yahoo! Japan (Japanese), and Baidu (Chinese) run similar CQA websites where there are constant requests on ToT item refinding.

proposed method is designed to span general-domain content rather than being restricted to narrow topical categories. This combination enables broader and more realistic evaluation of ToT retrieval systems.

We empirically evaluate multiple ToT query generation strategies and validate the resulting simulations using system rank correlation analysis, following established meta-evaluation methodologies [13, 21]. Our results show that the proposed method maintains high system rank correlation across all evaluated languages, providing evidence for the fidelity and robustness of our simulated evaluation method. Using this method, we construct a multilingual (CJK and English), general-domain ToT retrieval test collection, which has been released for use in the upcoming NTCIR evaluation forum.

2 Related Work

ToT Datasets. Prior work on tip-of-the-tongue (ToT) retrieval has produced a range of datasets spanning both single-domain and multi-domain settings. Early efforts primarily focused on well-defined content domains. In the movie domain, Arguello et al. [4] introduced a ToT dataset derived from the *IRememberThisMovie* forum. Parallel efforts explored ToT phenomena in books [9, 27], constructing datasets from user queries on Reddit and Goodreads, and in music [8], leveraging posts from the */r/tipofmytongue* subreddit. Also, ToT datasets have been extended to the video game domain, using data from the */r/tipofmyjoystick* subreddit [23].

Beyond single-domain collections, several works have pursued broader coverage across domains. Meier et al. [28] introduced a multi-domain ToT dataset focused on casual leisure content extracted from Reddit. Bogers et al. [12] compared the effectiveness of humans and large language models in answering questions across several ToT domains.The TOMT-KIS benchmark [20], also derived from */r/tipofmytongue*, further emphasized known-item search under ToT conditions. Large-scale evaluation campaigns have additionally contributed multi-domain ToT resources. The TREC 2024 [2] and TREC 2025 [3] ToT tracks adopted a synthetic ToT query generation framework [21], enabling controlled construction of test collections across a broad set of domains. Finally, a recent effort introduced a multimodal ToT query dataset created through human annotation [14].

TREC ToT Tracks. The TREC ToT tracks have played a central role in standardizing ToT evaluation. The inaugural TREC 2023 ToT track [1] focused exclusively on the movie domain [4]. Subsequent iterations significantly broadened both methodological rigor and domain coverage. In TREC 2024 [2] and TREC 2025 [3], organizers adopted rigorous principles from the known-item query simulation literature [6, 7, 16, 25], emphasizing systematic synthetic query generation and validation. This line of work culminated in a modernized ToT query simulation and validation pipeline [21], grounded in system ranking correlation analyses.

Using this framework, TREC 2024 expanded beyond movies to include landmarks and celebrities, while TREC 2025 further generalized the approach, resulting in a test collection that spanning 53 domains. Despite this breadth, both tracks remained monolingual and English-centric, reflecting a broader limitation in existing ToT evaluation resources.

Our Contributions. Our work builds on the progressive line of ToT dataset construction and evaluation established by prior TREC efforts, while introducing several differences. Similar to recent TREC tracks, we rely on a Wikipedia-based corpus and adopt a domain-agnostic synthetic ToT query generation framework. The generation process continues to follow rigorous simulation validation practices, including system ranking correlation analysis, ensuring comparability and reliability of the resulting benchmarks.

The core contribution of this work lies in its multilingual and multi-domain design. As far as we are aware, one prior effort [14] has explored multilingual ToT benchmarking. While valuable, that dataset is still primarily English-centric: most ToT queries are written in English, with multilinguality arising mainly from answer documents in other languages, accounting for only a minority of the data. In contrast, our dataset is multilingual by construction. Each language is associated with its own corpus, and all ToT queries are written in the target language. We cover four languages—English, Chinese, Japanese, and Korean—with 5,000 queries per language, spanning Movies, People, and general-domain content found in Wikipedia.

3 Method and Experimental Setup

To support Tip-of-the-Tongue (ToT) retrieval research in more linguistically and culturally diverse communities, we construct multilingual ToT test collections in four languages: Chinese, Japanese, Korean, and English. Each language is treated independently, with both corpus and ToT queries written in the same language, resulting in four self-contained test collections.

Our construction pipeline consists of three main stages. First, we build ToT entity sampling pools from the corresponding Wikipedia corpus (§3.1). Second, we generate synthetic ToT queries using an LLM (§3.2). Third, we validate the generated queries through system ranking correlation analysis to ensure that synthetic queries induce retrieval behavior consistent with real user ToT queries (§3.3). Based on this validation framework, we conduct a series of prompt variation experiments to identify effective query generation strategies for each language (§3.4).[4]

3.1 Sampling ToT Entity Candidates

We use Wikipedia as the underlying corpus due to its broad topical coverage and its rich cross-lingual linking structure. Specifically, we rely on publicly available Wikipedia dumps archived in November 2023 and distributed via HuggingFace.[5] These dumps include complete English, Chinese, Japanese, and Korean Wikipedia pages. We sample ToT entities for generation from these corpora.

Although the final retrieval corpus used for benchmarking remains identical to the original Wikipedia corpus, we partition the dataset internally to support subsequent analysis. For each East Asian language (CJK), we divide pages into two disjoint partitions:

- *monolingual pages*, which exist only in the target language, and
- *bilingual pages*, which have links to the English Wikipedia page for the same entity.

[4] Since English has been the primary focus of prior ToT retrieval studies, our experiments place particular emphasis on the three East Asian languages.
[5] https://huggingface.co/datasets/wikimedia/wikipedia

This partitioning allows us to explicitly control the presence or absence of English-aligned entities during query generation. For English, we use the original, non-partitioned corpus. In total, this process yields seven distinct partitions: two partitions (monolingual and bilingual) for each of the three East Asian languages, and one English partition.

Following prior ToT dataset construction work [2, 21], we filter each partition by popularity, retaining the top 20% of pages ranked by page view counts, and by document length, removing entities that are overly niche or insufficiently descriptive. We then assign each page to a coarse-grained domain using Wikidata metadata, grouping entities into Movies, People, and General categories.

To ensure coverage across different familiarity levels, we perform stratified sampling by popularity within each domain. Specifically, entities are sorted by popularity, divided into 20 equal-sized buckets, and sampled uniformly to obtain 2,500 entities from the monolingual partition and 2,500 from the bilingual partition under an 8:1:1 domain ratio (80% General, 10% Movies, 10% People). For English, which does not have monolingual or bilingual partitions, we directly sample 5,000 entities from the single unpartitioned pool using the same stratification strategy. This yields 5,000 ToT entity candidates per language for Chinese, Japanese, Korean, and English.

3.2 Synthetic ToT Query Generation

For synthetic ToT query generation, we adapt the LLM-elicitation method used by He et al. [21]. For each candidate entity, we generate a synthetic ToT query using the following procedure.

(1) *Summarization of the entity's Wikipedia page*:
The full Wikipedia page text of the sampled entity is used to construct a summarization prompt for the GPT-4o model with temperature 0.5.[6] If the page exceeds the model's token limit, it is truncated accordingly. The model produces a two-paragraph summary that captures salient attributes of the entity while abstracting away surface-level lexical details.

(2) *Constructing a query generation prompt*:
Using the generated summary as contextual input, we construct a ToT user simulation prompt that instructs the model to produce a natural language ToT query of the target entity. The same GPT-4o model is used for query generation but with temperature 0.3, following the findings of He et al. [21]. We experimented with multiple prompt variants to identify the one that best approximates the system ranking derived from real ToT queries. Details of these prompt variations and their experimental evaluation are described in §3.4.

(3) *Entity name anonymity check*:
A critical requirement for ToT query generation is that they must not explicitly reveal the target entity name. To enforce this constraint, we perform a post-generation check to detect occurrences of the entity name in the generated query. If the name is detected, we retry query generation using the same prompt, for up to three attempts. Queries that still fail this anonymity check after three retries are programmed to be discarded; however, all queries passed the check within the allowed attempts.

[6] GPT4o-2024-08-06

3.3 Validation of the Simulated Queries

To validate the fidelity of our simulated ToT queries, we adopt a system ranking correlation analysis framework [6, 7, 25], which has also been used in recent ToT dataset construction work [21]. The core premise of this approach is that if a set of synthetic queries is a good proxy for real user queries, it should induce similar rankings of retrieval systems when evaluated using standard IR metrics (e.g., NDCG@k, MRR).

We first collect a set of real human-authored ToT queries, denoted as Q_{real}, by manually collecting ToT question–answer pairs from community question answering (CQA) platforms that are predominantly written in each target language: Chinese, Japanese, Korean, and English.[7] Our goal is to collect 150 ToT entities per language for validation purpose. Due to the relatively lower availability of ToT queries in some East Asian languages, when the collected set falls short of this target, we supplement it by translating English ToT queries into the target language using a TowerInstruct-7B-v0.2 machine translation model [30], a winner of multiple WMT 2024 tasks.

To obtain robust and diverse system rankings, we construct a broad set of retrieval models. For each East Asian language, we construct a common pool of 22 retrievers, consisting of: (i) seven lexical retrieval models based on BM25 [32] and language modeling with Dirichlet smoothing [34], instantiated with different hyperparameter settings; (ii) one GPT-4o-based retriever following the setup of Arguello et al. [2]; and (iii) fourteen multilingual dense retrievers spanning multiple architectures and checkpoint variants [22, 31, 33]. In addition, for each language, we include five representative language-specific dense retrieval models selected based on reported performance on the corresponding language-specific MTEB leaderboard [29]. As a result, for each Chinese, Japanese, and Korean collection, we construct a total of 27 retrieval systems.

Using the real ToT queries Q_{real}, we evaluate all retrieval systems and rank them according to one IR metric, yielding a system ranking R_{real}. We then repeat the same evaluation procedure using the synthetic query set Q_{syn}, producing a corresponding ranking R_{syn}. To quantify the agreement between the two rankings, we compute both Kendall's τ and Pearson's r correlation coefficients between R_{real} and R_{syn}. A high correlation indicates that the synthetic queries preserve the relative ordering of retrieval systems observed under real user queries, thereby validating the quality of the simulated ToT queries.

3.4 Experimental Setup & Hypotheses

We experiment with four prompting variations to study how an English-developed ToT query generation prompt can be adapted to multilingual settings. Since prior work has shown that a carefully designed English prompt is effective for ToT query elicitation, our experiments focus on understanding how this prompt should be adapted when generating ToT queries in other languages.

We use the following notation. P_{en} denotes the original English ToT query generation prompt template developed by He et al. [21]. Let t be a target language. Accordingly, q_t represents a ToT query written in language t, and Wiki_t denotes a Wikipedia page written in language t. Trans_t is a translation function that translates text

[7] We do not disclose queries collected from CQA websites.

into the target language t. inst_{en} is an English-written instruction that explicitly asks the model to produce output in language t. The operator $\oplus$ indicates string insertion into a prompt template; for example, $P_{en} \oplus \text{Wiki}_{en}$ means that the English Wikipedia text is inserted into the appropriate slot of the English prompt.

Using this notation, we define four prompting strategies:

- **Variation 1: Translated Prompt + Non-English Wikipedia**

$$\text{LLM}(\text{Trans}_t(P_{en}) \oplus \text{Wiki}_t) \rightarrow q_t$$

  In this strategy, the original English prompt is translated into the target language and combined with a Wikipedia page written in the same language. The entire prompt is therefore written solely in language t.
- **Variation 2: English Prompt + Non-English Wikipedia**

$$\text{LLM}((P_{en} \oplus \text{inst}_{en}) \oplus \text{Wiki}_t) \rightarrow q_t$$

  Here, the prompt remains in English but includes an explicit instruction to generate the output in language t. The Wikipedia content is provided in the target language, resulting in a bilingual prompt.
- **Variation 3: Translated Prompt + English Wikipedia**

$$\text{LLM}(\text{Trans}_t(P_{en}) \oplus \text{Wiki}_{en}) \rightarrow q_t$$

  This variation mirrors Variation 1 in that the prompt is translated into the target language, but the Wikipedia content is taken from the English version of the page.
- **Variation 4: English Prompt + English Wikipedia**

$$\text{Trans}_t(\text{LLM}(P_{en} \oplus \text{Wiki}_{en})) \rightarrow q_t$$

  In this approach, both the prompt and the Wikipedia content remain in English. The LLM first generates an English ToT query, which is then translated into the target language.

These prompting strategies are applied under different entity sampling scopes. Recall that for East Asian languages we construct both a *monolingual* and a *bilingual* partition, while the *full set* refers to the unpartitioned sampling pool. Variations 1 and 2 can be applied to both the full set and the monolingual partition. In contrast, all four variations are applicable to the bilingual partition, since it provides access to both non-English language and English Wikipedia pages.

Based on the four variations, we can ask two research questions:

- **RQ1**: How does the language of the query generation prompt (non-English vs. English) affect the fidelity of simulated ToT queries, as measured by system-rank correlation with real user queries?
- **RQ2**: How does the language of the source Wikipedia content (non-English vs. English) influence the quality of simulated ToT queries across different languages and sampling partitions?

4 Results and Analyses

4.1 System Rank Correlation

Results for the CJK languages are reported in Tables 1, 2, and 3. We evaluate all sampling partitions, including the full set, which serves as a global reference for average performance across entity types. We primarily base our analysis on Kendall's τ correlation, as linear relationship-based measures such as Pearson's r are more sensitive to outliers and may yield misleading interpretations in this setting. Pearson's r is reported for completeness.

For Chinese (Table 1), strong rank correlations are observed across all experimental settings. In particular, English-Wikipedia-based generation in the bilingual partition achieves the highest correlations, while Chinese-language source generation with translated prompts also performs competitively on the full set.

For Japanese (Table 2), overall correlation levels are lower than those observed for Chinese and Korean. Across the full and bilingual partitions, configurations using English prompts consistently outperform their Japanese-language prompt counterparts.

For Korean (Table 3), correlations are uniformly high across settings. Prompt's instruction language seems to be less important, while Korean Wikipedia content achieves clear win against English Wikipedia-based generation.

To select the best query generation strategy for monolingual and bilingual partition, we compute the mean Kendall's τ across the three base retrieval metrics (NDCG@100, NDCG@1000, and MRR). The best-performing configurations according to this mean τ are boldfaced in the tables. Based on this criterion, the selected strategies are: for Chinese, Variation 1 for the monolingual partition and Variation 3 for the bilingual partition; for Japanese, Variation 2 for both monolingual and bilingual partitions; and for Korean, Variation 2 for the monolingual partition and Variation 1 for the bilingual partition.

RQ1 (Prompt Language). The effect of prompt language is language-dependent rather than universal. For Japanese, English-written prompts yield higher system-rank correlations than non-English language prompts, suggesting that English instructions provide more stable or effective guidance for the LLM during ToT query generation. In contrast, for Chinese, and Korean non-English language prompts perform competitively and do not exhibit a consistent disadvantage. Overall, prompt language influences ToT simulation fidelity, but the optimal choice depends on the target language.

RQ2 (Source Wikipedia Language). The impact of source Wikipedia language is also language-specific. English Wikipedia content leads to the highest correlations for Chinese in the bilingual setting, indicating that richer English articles can improve ToT simulation when non-English language resources are comparatively less detailed. However, for Japanese and Korean, non-English Wikipedia content outperforms English, confirming our observation that these corpora are sufficiently rich unlike those of Chinese, and that English content does not provide additional benefits. Consequently, the choice of source Wikipedia language should be made on a per-language basis rather than assumed to generalize across languages.

4.2 Domain-Level Analysis

To examine how simulation fidelity varies across entity types, we analyze validation performance by domain. Across all three languages, the General domain consistently achieves high system-rank correlations, with Kendall's τ frequently exceeding 0.7. For Chinese, the Movies domain yields the highest correlations among all domains. In contrast, for Japanese and Korean, both the Movies and People domains exhibit slightly lower correlations than the General domain.

Table 1: Chinese (ZH): system rank correlation results across 8 experimental settings.

	Experimental Configuration			NDCG@100		NDCG@1000		MRR		Metric Mean	
ID	Candidate Partition	Prompt	Wiki	τ	r	τ	r	τ	r	$\bar{\tau}$	$\bar{r}$
1	Full Set	Chinese (Trans)	Chinese	0.7805	0.8647	0.7880	0.8823	0.6390	0.7852	0.7358	0.8441
2	Full Set	English	Chinese	0.6562	0.6967	0.6963	0.7395	0.5989	0.6135	0.6505	0.6832
3	Monolingual	Chinese (Trans)	Chinese	0.7040	0.8159	0.6963	0.8784	0.6370	0.7124	**0.6791**	0.8022
4	Monolingual	English	Chinese	0.6313	0.6483	0.6848	0.7521	0.5415	0.5288	0.6192	0.6431
5	Bilingual	Chinese (Trans)	Chinese	0.6619	0.8717	0.7098	0.8772	0.6218	0.8167	0.6645	0.8552
6	Bilingual	English	Chinese	0.5931	0.6987	0.6332	0.7266	0.6178	0.6313	0.6147	0.6855
7	Bilingual	Chinese (Trans)	English	0.7650	0.9222	0.7364	0.9306	0.7421	0.8971	**0.7478**	0.9166
8	Bilingual	English	English	0.7536	0.9164	0.7479	0.9210	0.6676	0.9150	0.7230	0.9175

Table 2: Japanese (JA): system rank correlation results across 8 experimental settings.

	Experimental Configuration			NDCG@100		NDCG@1000		MRR		Metric Mean	
ID	Candidate Partition	Prompt	Wiki	τ	r	τ	r	τ	r	$\bar{\tau}$	$\bar{r}$
1	Full Set	Japanese (Trans)	Japanese	0.4631	0.9437	0.5183	0.9329	0.3748	0.9613	0.4521	0.9460
2	Full Set	English	Japanese	0.5611	0.8906	0.5940	0.8940	0.4833	0.8993	0.5461	0.8946
3	Monolingual	Japanese (Trans)	Japanese	0.4680	0.8659	0.5426	0.8524	0.4331	0.8998	0.4812	0.8727
4	Monolingual	English	Japanese	0.4986	0.7758	0.5871	0.8194	0.4733	0.7560	**0.5197**	0.7837
5	Bilingual	Japanese (Trans)	Japanese	0.4332	0.9205	0.5222	0.9082	0.3602	0.9349	0.4385	0.9212
6	Bilingual	English	Japanese	0.5358	0.9118	0.5989	0.9014	0.4914	0.9263	**0.5420**	0.9132
7	Bilingual	Japanese (Trans)	English	0.4136	0.9112	0.5316	0.9031	0.3902	0.9320	0.4451	0.9154
8	Bilingual	English	English	0.4878	0.9366	0.5645	0.9384	0.5086	0.9542	0.5203	0.9431

Table 3: Korean (KO): system-rank correlation results across 8 experimental settings.

	Experimental Configuration			NDCG@100		NDCG@1000		MRR		Metric Mean	
ID	Candidate Partition	Prompt	Wiki	τ	r	τ	r	τ	r	$\bar{\tau}$	$\bar{r}$
1	Full Set	Korean (Trans)	Korean	0.7626	0.9668	0.8127	0.9696	0.6589	0.9746	0.7447	0.9703
2	Full Set	English	Korean	0.7133	0.9505	0.7683	0.9516	0.6377	0.9554	0.7064	0.9525
3	Monolingual	Korean (Trans)	Korean	0.5703	0.8413	0.6119	0.8690	0.5471	0.7689	0.5764	0.8264
4	Monolingual	English	Korean	0.6045	0.8240	0.7054	0.8754	0.5486	0.7036	**0.6195**	0.8010
5	Bilingual	Korean (Trans)	Korean	0.7989	0.9600	0.8184	0.9672	0.5690	0.9649	**0.7288**	0.9640
6	Bilingual	English	Korean	0.7191	0.9467	0.7931	0.9486	0.5876	0.9521	0.6999	0.9491
7	Bilingual	Korean (Trans)	English	0.7422	0.9496	0.7723	0.9516	0.6143	0.9577	0.7096	0.9530
8	Bilingual	English	English	0.7636	0.9370	0.7723	0.9432	0.6434	0.9551	0.7264	0.9451

Table 4: Statistics of each CJK ToT Test Collection.

Category	Distribution (Count / Percentage)
Total Queries	5,000 per language (15,000 total)
Linguistic Split	Monolingual (2,500), Bilingual (2,500)
Domain Split	Movies (10%), People (10%), General (80%)
Dataset Split	Train (80%), Dev (10%), Test (10%)
Corpus Size	C (1.38M), J (1.39M), K (648K)

4.3 Constructing Test Collections

Using the selected query generation strategies, we generate ToT queries for the 5,000 entity candidates per language. For each CJK language, if an entity belongs to the monolingual partition, we apply the strategy corresponding to the highest mean τ among Exp 3 or Exp 4. If an entity belongs to the bilingual partition, we apply the best-performing strategy among Exp 5, Exp 6, Exp 7, or Exp 8, as determined by mean τ. For English, we directly follow the original English ToT query generation procedure proposed by He et al. [21], generating queries from 5,000 sampled entities without partitioning.

Dataset statistics for each CJK collection are summarized in Table 4. Together, these procedures result in four final ToT test collections: Chinese, Japanese, Korean, and English.

5 Discussion

Resource Contribution. This work makes a concrete contribution to the community by releasing a new multilingual Tip-of-the-Tongue (ToT) dataset and the accompanying infrastructure needed to support future research. Specifically, we release the training and development sets for all supported languages together with this paper, along with the full source code used to generate the dataset

and supporting resources. To align with the established evaluation practices, we plan to release the test sets following the official NTCIR-19 timeline. This schedule places the test set release close to the SIGIR 2026 conference dates, enabling coordinated evaluation, benchmarking, and community-wide participation.

Availability. The released resources are fully available to reviewers and researchers at the time of submission. The training and development sets for all languages, along with their associated corpora, are publicly released and accessible via a permanent archival link.[8] In addition, comprehensive documentation describing the dataset construction, statistics, and usage is provided through the official track webpage.[9] The test sets for all languages will be released during the summer in accordance with the NTCIR-19 evaluation schedule. To support transparency and reproducibility, the source code used for dataset generation and processing is also publicly available through an open-access repository.[10] All released resources are distributed under open licensing terms, allowing unrestricted use by both academic researchers and industry practitioners.

Utility. The dataset is designed to be straightforward to use. Detailed documentation of the dataset provenance, preprocessing, and aggregation steps are provided in this paper and on the track webpage. As a result, users with standard information retrieval or natural language processing expertise should be able to adopt the resource without specialized domain knowledge. To further lower the barrier to entry, we provide usage examples through the README file in the source code repository. These examples demonstrate how to load the dataset and reproduce baseline experiments. In addition, the dataset is integrated into the well-established `ir_datasets` Python framework, enabling practitioners to load the data with a single line of Python code and seamlessly incorporate it into existing IR pipelines.

Novelty and Predicted Impact. This resource extends a well-established line of work on Tip-of-the-Tongue queries [4] and known-item retrieval [6, 7, 25] by broadening its scope to multilingual and general-domain synthetic ToT query generation for simulated evaluation and benchmarking. As such, the contribution opens new opportunities for studying ToT retrieval under multilingual, cross-lingual, and culturally diverse conditions.

ToT queries themselves fall under the broader category of complex information needs, which still remain challenging for modern retrieval systems [24]. The proposed dataset therefore provides a valuable testbed for advancing research on complex query understanding and retrieval.

Moreover, multilingual retrieval [19] remains an active and growing research area, as also evidenced by recent competition [26]. Our dataset allows researchers to easily construct multilingual and cross-lingual retrieval benchmarks by merging corpora across languages. Because the corpus is partitioned not only by language but also by the existence of corresponding English pages, researchers can naturally define relevance across languages, including settings where multiple bilingual documents are considered relevant for a single query.

The dataset also supports research on multicultural and multilingual LLM development. Multilinguality inherently encompasses cultural variation [5], and Tip-of-the-Tongue expressions may reflect culturally specific descriptions and associations. The same underlying concept may be described in systematically different ways across Chinese-, Japanese-, Korean-, and English-speaking populations [10], making ToT queries a particularly rich signal for studying cultural variation in language use.

Taken together, these factors suggest that the dataset has long-term value and that the community of researchers using it is likely to grow over time.

Implications of Results. Our results demonstrate that multilingual ToT query simulation requires language-specific design choices rather than uniform reuse of English-centric pipelines. The analyses for RQ1 and RQ2 show that both prompt language and source Wikipedia language substantially influence simulation fidelity, but their effects vary across languages and entity types. Non-English language Wikipedia content is generally essential for accurately capturing language- and culture-specific ToT behavior, while English Wikipedia can be beneficial when non-English resources are less detailed. These findings suggest that effective multilingual ToT dataset construction should rely on adaptive, language-aware prompting and content selection strategies instead of direct translation or one-size-fits-all approaches.

Limitation and Future Work. One limitation of the current dataset is its focus on East Asian languages. However, the underlying methodology for query generation is not language-specific and could be extended to other linguistic and cultural contexts. Expanding the dataset to additional languages therefore represents a natural direction for future work. In addition, modern information access systems are increasingly interactive and multi-turn in nature [15]. Tip-of-the-Tongue experiences in real-world settings often unfold over multiple interactions rather than a single query [28]. Developing multi-turn ToT retrieval system and test collections is an important and promising direction, and we view the current resource as a foundational step toward supporting such future evaluations.

6 Conclusion

We present the first large-scale multilingual Tip-of-the-Tongue retrieval benchmark for Chinese, Japanese, and Korean, alongside a comparable English collection. Through systematic validation, we show that realistic ToT simulation requires language-aware design choices in prompt and source selection. These datasets will be used as official test collections in NTCIR-19, supporting future research on multilingual and cross-lingual ToT retrieval.

Acknowledgments

We thank Kimihiro Hasegawa and Masao Someki for Japanese CQA query collection.

[8] https://zenodo.org/records/18777084
[9] https://ntcir-tot.github.io
[10] https://github.com/kimdanny/ntcir-19-tot

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 *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023)*. 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 *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024)*. https://trec.nist.gov/pubs/trec33/papers/Overview_tot.pdf

[3] Jaime Arguello, Fernando Diaz, Maik Fröebe, To Eun Kim, and Bhaskar Mitra. 2026. Overview of the TREC 2025 Tip-of-the-Tongue track. *arXiv preprint arXiv:2601.20671* (2026).

[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 (CHIIR '21)*. Association for Computing Machinery, 5–14. doi:10.1145/3406522.3446021

[5] Larissa Aronin and Muiris Ó Laoire. 2004. Exploring Multilingualism in Cultural. *Trilingualism in family, school, and community* 43 (2004), 11.

[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, 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] 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 (SIGIR '23)*. Association for Computing Machinery, 2506–2510. doi:10.1145/3539618.3592086

[9] 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, 48–56. doi:10.1145/3488560.3498421

[10] Shaily Bhatt and Fernando Diaz. 2024. Extrinsic Evaluation of Cultural Competence in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 16055–16074. doi:10.18653/v1/2024.findings-emnlp.942

[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] Toine Bogers, Maria Gäde, Mark M. Hall, Marijn Koolen, Vivien Petras, and Mette Skov. 2026. Tip-of-the-Tongue Search in the Wild: Analyzing Human and LLM Performance and Success Factors on Complex Search Requests. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval, CHIIR 2026, Seattle, WA, USA, March 22-26, 2026*. ACM, 162–171. doi:10.1145/3786304.3787864

[13] Ben Carterette. 2009. On rank correlation and the distance between rankings. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 436–443. doi:10.1145/1571941.1572017

[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] J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker. 2018. Research Frontiers in Information Retrieval: Report from the Third Strategic Workshop on Information Retrieval in Lorne (SWIRL 2018). *SIGIR Forum* 52, 1 (Aug. 2018), 34–90. doi:10.1145/3274784.3274788

[16] 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, 25–34. doi:10.1145/2009916.2009924

[17] 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.

[18] 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.

[19] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive Multilingual Text Embedding Benchmark. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=zl3pfz4VCV

[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] 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* (Padua, Italy) *(SIGIR '25)*. Association for Computing Machinery, New York, NY, USA, 3398–3407. doi:10.1145/3726302.3730335

[22] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118 [cs.IR] https://arxiv.org/abs/2112.09118

[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 (FDG '20)*. Association for Computing Machinery, Article 40, 4 pages. doi:10.1145/3402942.3402971

[24] Julian Killingback and Hamed Zamani. 2025. Benchmarking Information Retrieval Models on Complex Retrieval Tasks. *arXiv preprint arXiv:2509.07253* (2025).

[25] 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 (CIKM '09)*. Association for Computing Machinery, 1297–1306. doi:10.1145/1645953.1646117

[26] Dawn Lawrie, Sean MacAvaney, James Mayfield, Luca Soldaini, Eugene Yang, and Andrew Yates. 2026. WSDM CUP 2026: Multilingual Retrieval. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining* (USA) *(WSDM '26)*. Association for Computing Machinery, 1394–1395. doi:10.1145/3773966.3778021

[27] 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*. Association for Computational Linguistics, 5521–5533. doi:10.18653/v1/2023.findings-emnlp.367

[28] 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 (HT '21)*. Association for Computing Machinery, 143–154. doi:10.1145/3465336.3475096

[29] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037. doi:10.18653/v1/2023.eacl-main.148

[30] Ricardo Rei, José Pombal, Nuno M Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, et al. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*. 185–204.

[31] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[32] S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland) *(SIGIR '94)*. Springer-Verlag, Berlin, Heidelberg, 232–241.

[33] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual Universal Sentence Encoder for Semantic Retrieval. arXiv:1907.04307 [cs.CL] https://arxiv.org/abs/1907.04307

[34] 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, 334–342. doi:10.1145/383952.384019