Synthetic Tip-of-the-Tongue Query Generation for Simulated Evaluation

TREC TOT Dataset

Yifan He To Eun Kim Fernando Diaz
{yifanhe,toeunk,fernandd}@cs.cmu.edu

Language Technologies Institute

TL;DR

This study introduces a novel **LLM-based user simulator designed to replicate the Tip-of-the-Tongue (TOT) cognitive state**, where searchers struggle to recall specific identifiers of the items they seek. By simulating how users interact with retrieval systems, our simulator generates **synthetic queries** that closely mimic human TOT queries.

The synthetic queries generated by our simulator show a **high correlation with how human TOT queries rank TOT retrieval systems**, achieving a Kendall's tau value above 0.8.

These synthetic queries have been released and are being used as test queries in the **TREC 2024 TOT track**.

Background & Motivation

Tip-of-the-Tongue (TOT) known-item retrieval refers to the process of searching for a previously encountered item when the searcher is unable to recall a specific identifier. TOT queries typically exhibit characteristics such as: 1) expressions of uncertainty, 2) exclusion criteria, 3) distorted memories, and 4) verbosity.

Problem 1: **SOTA search systems often fail to satisfy TOT searchers**. This is why many users turn to online forums like Reddit to post TOT-related questions, expressing frustration at their inability to find answers through standard web search.

Problem 2: **lack of datasets for TOT retrieval**. This stems from both 1) corporate privacy concerns and 2) the challenge of eliciting and capturing the TOT state during data collection and annotation.

We propose developing a TOT-user simulator that can be used to:

- generate synthetic TOT queries to mitigate the lack of available datasets,
- enable eyes-off offline evaluations of TOT retrieval systems, and
- support the training and evaluation of TOT retrieval agents in multi-turn conversational settings.

Method

We focus on the "Movie" domain and apply the developed simulator to the "Celebrity" and "Landmark" domains to evaluate its generalizability.

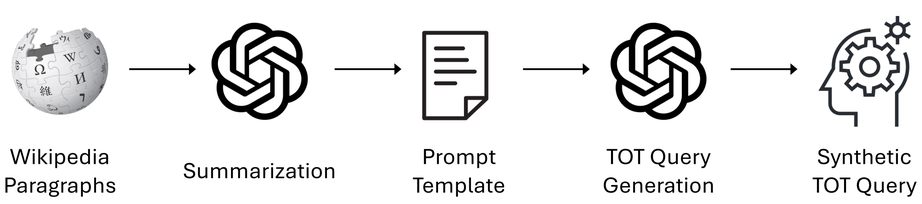

Figure 1. (User Simulation) TOT query generation process.

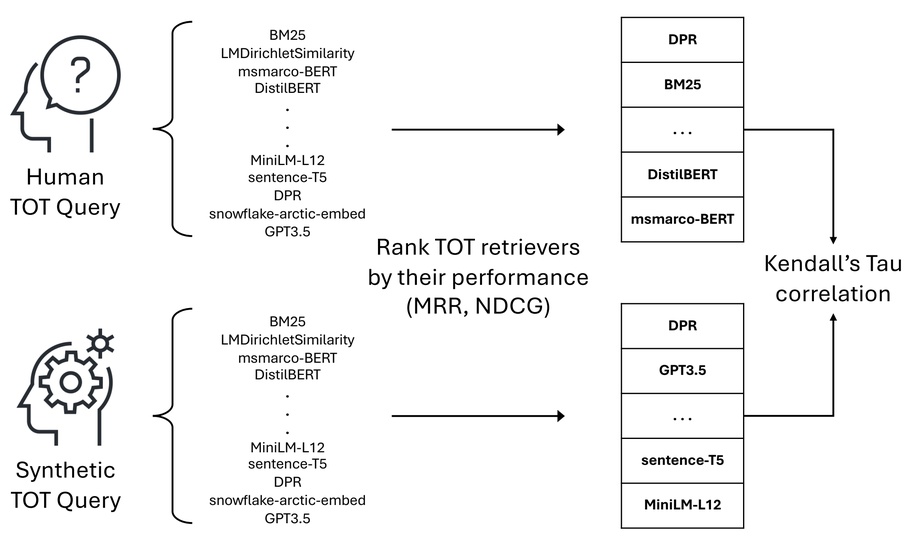

Figure 2. (Evaluation of User Simulation) We run 40 different retrievers with both human and synthetic queries, followed by ranking of retrievers by their search performance. Then, we observe the Kendall's Tau correlation between two rankings of retrievers.

Iterative Refinement of TOT User Simulator

Through **12 iterations of refinement**, we refined our prompts and adjusted the temperature parameter of GPT-4o. The final version (V6) achieved the highest τ-value and qualitatively produced queries most similar to human TOT queries. Additionally, we **fine-tuned GPT-4o using prompt V6**, using human TOT queries from the celebrity and landmark domains collected from Reddit, further improving the τ-value.

	Prompt Design				Model	
	Instruction Type	Wiki Summary	Few-Shot	Generation Requirements	Temperature	Model Fine-Tuning
V1	TOT explanation	w/o	0-shot	9 rules	0.5	No
V2	TOT explanation	w/	0-shot	9 rules	0.5	No
V3	TOT explanation	w/	6-shot	0 rules	0.7	No
V4	Searcher role play	w/	0-shot	13 rules	0.3, 0.5, 0.7	No
V5	Searcher role play	w/	0-shot	14 rules	0.1, 0.3, 0.5	No
V6	Searcher role play	w/	0-shot	7 Musts + 7 Coulds	0.3	No, Yes

Figure 3. Different prompt and modeling strategies attempted.

Let's do a role play. You are now a person who watched a movie {ToTObject} a long time ago and forgot the title of the movie.
...
I will provide you a basic information about the movie, and you have to follow the guidelines to generate a post.

Information about {ToTObject}:
{WikiSummary}

Guidelines:
MUST FOLLOW:
1. Reflect the imperfect nature of memory with phrases that express doubt or mixed recollections.
...
7. Provide vivid but ambiguous details to stir the reader's imagination while leaving them guessing.

COULD FOLLOW:
1. Share a personal anecdote related to when or with whom you watched the movie. Think of unique ways to set the scene.
...
7. Focus on sensory details such as the overall mood, sounds, or emotional impact of the movie, using vivid descriptions.

Generate a post based on these guidelines.

Figure 4. Prompt Version 6

Results

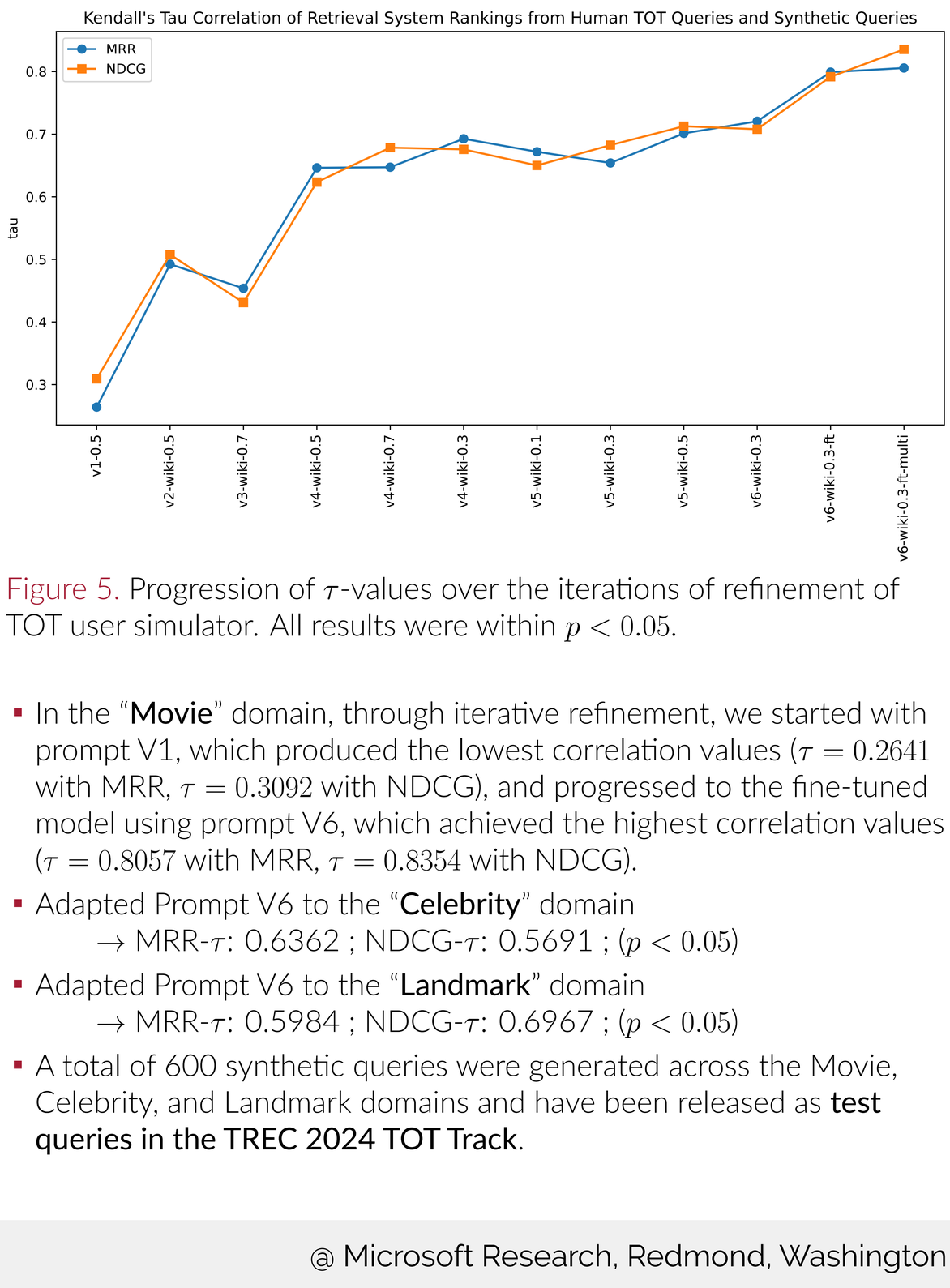

Figure 5. Progression of τ-values over the iterations of refinement of TOT user simulator. All results were within $p < 0.05$.

- In the "**Movie**" domain, through iterative refinement, we started with prompt V1, which produced the lowest correlation values ($\tau = 0.2641$ with MRR, $\tau = 0.3092$ with NDCG), and progressed to the fine-tuned model using prompt V6, which achieved the highest correlation values ($\tau = 0.8057$ with MRR, $\tau = 0.8354$ with NDCG).
- Adapted Prompt V6 to the "**Celebrity**" domain
  $\rightarrow$ MRR-τ: 0.6362 ; NDCG-τ: 0.5691 ; ($p < 0.05$)
- Adapted Prompt V6 to the "**Landmark**" domain
  $\rightarrow$ MRR-τ: 0.5984 ; NDCG-τ: 0.6967 ; ($p < 0.05$)
- A total of 600 synthetic queries were generated across the Movie, Celebrity, and Landmark domains and have been released as **test queries in the TREC 2024 TOT Track**.