OVERVIEW OF THE TREC 2024 TIP-OF-THE-TONGUE TRACK

Jaime Arguello[1], Samarth Bhargav[2], Fernando Diaz[3], To Eun Kim[3], Yifan He[3], Evangelos Kanoulas[2], and Bhaskar Mitra[4]

[1]University of North Carolina, USA, jarguell@email.unc.edu
[2]University of Amsterdam, Netherlands, {s.bhargav, e.kanoulas}@uva.nl
[3]Carnegie Mellon University, USA, diazf@acm.org, {toeunk, yifanhe}@cs.cmu.edu
[4]Microsoft Research, Canada, bmitra@microsoft.com

ABSTRACT

Tip-of-the-tongue (ToT) known-item retrieval involves re-finding an item for which the searcher does not reliably recall an identifier. ToT information requests (or queries) are verbose and tend to include several complex phenomena, making them especially difficult for existing information retrieval systems. The TREC 2024 ToT track focused on a single ad-hoc retrieval task. Participants were provided with training and development data in the movie domain. Conversely, systems were tested on data that combined three domains: movies, celebrities, and landmarks. This year, 6 groups (including the track coordinators) submitted 18 runs.

Track website: https://trec-tot.github.io

1 Introduction

Tip-of-the-tongue (ToT) known-item retrieval involves retrieving an item for which the searcher is unable to reliably recall an identifier. An example might involve resolving the name of a movie for which the searcher does not recall the title or the name of a cast member. ToT information requests (or queries) are verbose and include several complex phenomena. First, they include information about the item itself (i.e., semantic memories) as well as the context in which the searcher last engaged with the item (i.e., episodic memories). Additionally, they include language phenomena that simple keyword-matching algorithms are not equipped to handle, such as mentions of (un)certainty, exclusion criteria, relative comparisons, and false memories. Such phenomena are not prevalent in verbose queries in other retrieval scenarios.

Current IR systems are not well-suited to resolve ToT information needs. As evidence, a wide range of community Q&A sites have emerged to help people resolve their ToT information needs through the help of other people. Such Q&A sites focus on domains such as movies[1], books[2], stories[3], and songs[4].

Last year (TREC 2023), the ToT track focused on a single ad-hoc retrieval task in the movie domain [Arguello et al., 2024]. This year, the TREC 2024 ToT Track expanded the task to include three domains: movies, celebrities, and landmarks. Participants were provided with a training set of 150 queries and two development sets of 150 queries each. All 450 queries originated from the Microsoft ToT Known-Item Retrieval Dataset for Movie Identification (MS-ToT Dataset) [Arguello et al., 2021].[5] All queries originated from the "I Remember This Movie…" community Q&A site, which helps people re-find movies and television shows.[6]

Participants were provided with a test set of 600 queries. The test set included 150 new movie queries sampled from the MS-ToT Dataset. Additionally, the test set included 450 *synthetic* queries generated as described in Section 3.2—

[1]https://www.reddit.com/r/tipofmytongue
[2]https://www.goodreads.com/group/show/185-what-s-the-name-of-that-book
[3]https://scifi.stackexchange.com/questions/tagged/story-identification
[4]https://www.watzatsong.com/en
[5]https://github.com/microsoft/Tip-of-the-Tongue-Known-Item-Retrieval-Dataset-for-Movie-Identification
[6]https://irememberthismovie.com

150 movie queries, 150 celebrity queries, and 150 landmark queries. The document corpus was composed of the subset of Wikipedia (3,185,450 articles) associated with our three target domains. For each query, participants were asked to produce a ranked list of 1000 document IDs from the Wikipedia corpus. The corpus contained the correct answer for every training, development, and test query. Runs were evaluated using metrics appropriate for retrieval tasks associated with one relevant item (e.g., NDCG@1000). The TREC 2024 ToT track received 18 runs from 6 research groups.

2 Task description

The TREC 2024 ToT Track had a single ad-hoc retrieval task focusing on movie, celebrity, and landmark identification. Participants were provided with a training set with 150 queries and two development sets with 150 queries each. All training and development queries originated from the MS-ToT Dataset [Arguello et al., 2021], which contains 1,000 query-answer pairs gathered from "I Remember This Movie...", a forum where users post ToT requests for movies.

Participants were provided with a test set of 600 queries. Of these, 150 were new queries sampled from the MS-ToT Dataset. The remaining 450 test queries consisted of *synthetic* queries evenly split across three domains: movies, celebrities, and landmarks. Additionally, track participants were provided with a corpus of 3,185,450 Wikipedia articles associated with our three target domains based on their assigned Wikipedia categories. The Wikipedia corpus contained the correct answer for all training, development, and test queries distributed to participants. Given a query, participants were asked to produce a ranking of (at most) 1000 document IDs from the Wikipedia corpus, with the correct answer ranked as high as possible. Participants were allowed to use external resources such as Wikidata. The official metric was NDCG@1000.

As previously noted, ToT queries are verbose. However, they contain complex phenomena that are not typically found in other retrieval scenarios involving verbose queries. Such phenomena include: (1) memories about the item itself, (2) contextual memories, (3) false memories, (4) mentions of uncertainty, (5) mentions of previous failed attempts to re-find the item, (6) relative comparisons that require multi-hop reasoning to be useful, and (7) social niceties.

The following are examples of the four types of queries included in the test set: MS-ToT movies, synthetic movies, synthetic celebrities, and synthetic landmarks.

MS-ToT Movies

Okay guys this is a weird one. I keep having this memory of a scene where a guy is late for class and runs down this kind of covered Pathway to the classroom building door. This is like in college or something. But when he grabs the door and tries to open it, it is locked. He then shakes his head and sits down in front of it. Then somebody else walks up, and opens the other door, he looks around to make sure nobody saw his mistake, and runs inside. I am not sure if this is in the opening credits of some TV show, or it's a scene from a movie, or what. But this has been frustrating me for about 6 days now. My wife thinks I'm crazy. Please help. This is live action, full color, probably late eighties to Mid nineties and on American TV comma whether it was a movie on TV or TV show I am not sure.

Correct Answer: Tommy Boy

Synthetic Movies

I'm trying to recall this one film I watched ages ago. It was set in this vast, open landscape, probably somewhere in the American frontier. The main character was this soldier guy who ends up at a really rundown fort. He was supposed to be there alone, I think, and he starts fixing the place up. There was this really cool part where he befriends a wolf. I remember the wolf because it was such a touching moment, and it felt like the wolf was his only friend for a while. The soldier guy also starts interacting with a nearby Native American tribe. At first, there's a lot of tension, but over time, he learns their language and customs. There's this wise elder and a fierce warrior who stand out in my memory. I also recall a romantic subplot with a woman who was somehow connected to the tribe, maybe adopted or something. The whole movie had this epic feel, with sweeping shots of the plains and intense moments like a buffalo hunt. Towards the end, things get pretty dramatic. The soldier guy gets captured by his own people, who think he's a traitor. There's a rescue scene, but he realizes he can't stay with the tribe because it puts them in danger. The movie ends on a bittersweet note with the soldier and the woman leaving, and there's some text about the fate of the tribe. Does this ring any bells for anyone? It's been bugging me for days!

Correct Answer: Dances with Wolves

Synthetic Celebrities

There's this actor I keep thinking about, and it's driving me nuts that I can't remember his name. He was pretty big in the 80s and 90s, often playing these quirky, everyman roles. I think he was in some teen movies back in the day, you know, those classic high school flicks where the guy is kind of an underdog but super charming in his own awkward way. One of those movies had him holding a boombox over his head in a really iconic scene, if that rings any bells. He also did some more serious stuff later on, like this one film where he was a hitman going back to his high school reunion. That one was a mix of dark comedy and action, and it was pretty unique for its time. I remember he was also in a movie with a lot of famous faces, something about a plane full of convicts, and he played a good guy trying to sort out the mess. His family is kind of a big deal too, I think. He has a sister who's also an actress, and they might have even been in a movie together at some point. Their last name is pretty distinctive, but it's just not coming to me right now. He's also known for being quite vocal about his political views, often stirring up conversations on social media. Does anyone know who I'm talking about? It's really bugging me, and I could use some help jogging my memory!

Correct Answer: John Cusack

Synthetic Landmarks

So, there's this place in New York City that I can't quite put my finger on. It's one of those grand, old buildings that feels like it's been around forever, you know? I remember going there a few years ago for a concert, and the whole experience was just magical. The venue itself is this massive, elegant structure with a kind of old-world charm. It's got these intricate designs on the walls and ceilings, almost like something you'd see in an old European opera house. The main auditorium is huge, with multiple levels of seating that seem to go on forever. I think it was five stories high, but my memory might be playing tricks on me. The acoustics were incredible; you could hear every note perfectly, no matter where you were sitting. I remember the seats being this rich, deep red color, adding to the overall luxurious feel of the place. There are also a couple of smaller halls within the same building. One of them is more intimate, with fewer seats, and I think it's used for smaller performances or recitals. The other one is somewhere in between, size-wise. I remember walking past these smaller halls and thinking how versatile the whole venue must be. What really stuck with me was the history of the place. I overheard someone mentioning that it was saved from demolition by a famous musician, which added a layer of significance to the whole experience. The building has this mix of old and new, with some parts feeling like they've been meticulously preserved while others have clearly been updated over the years. I was with a group of friends, and we spent some time exploring the area around the venue before the concert. It's in a pretty bustling part of the city, with lots of restaurants and shops nearby. We grabbed a bite to eat at a cozy little cafe just a block or two away, which made the whole evening feel even more special. Does anyone know the name of this place? It's driving me crazy trying to remember! Any help would be much appreciated.

Correct Answer: Carnegie Hall

3 Data

3.1 Corpus Construction

Participants were provided with a corpus of 3,185,450 Wikipedia articles associated (directly or indirectly) with Wikipedia categories relevant to our three domains (i.e., movies, celebrities, and landmarks). The corpus contained the correct answer for all training, development, and test queries. Each document in the corpus included the following fields:

- **doc_id:** Unique document identifier (Wikipedia page ID).
- **title:** Wikipedia page title.
- **text:** Full text of the Wikipedia page.
- **wikidata_id:** ID of corresponding entity in Wikidata.
- **sections:** list of top-level sections with: (1) section title, (2) starting character position in the text field, and (3) ending character position in the text field.

Section information was provided in case participants wanted to focus on only certain sections of the Wikipedia page.

3.2 Synthetic Query Generation

In addition to queries from the MS-ToT Dataset [Arguello et al., 2021], we generated synthetic queries using a large language model. Queries were synthesized by first sampling a random Wikipedia article from our corpus (Section 3.1) and then using a category-specific prompt seeded with the article title to pass to a large language model.

We developed category-specific prompts that approximated the ordering of systems with non-synthetic ToT queries. First, we gathered a small set of ToT queries for each category. We used evaluation queries from TREC 2023 ToT to develop the prompt for movies. To develop the prompt for celebrity es and landmarks, we sampled queries from the `r/tipofmytongue` subreddit. All of these queries had the correct answer, allowing us to measure system performance. Second, we ran k retrieval systems for each set of non-synthetic queries. Third, we iteratively designed a prompt to generate synthetic queries associated with the same relevant items. The prompt for each category was parameterized by the relevant item name (i.e., Wikipedia page title) and the first paragraph of the Wikipedia page. Figure 1 shows the prompt used to generate synthetic landmark queries. This resulted in the same number of queries for each category, which could then be used to generate document rankings for each of the k systems. Because the relevant item was known for the synthetic queries, for each category, we could compute evaluation metrics and a ranking of systems. We iteratively designed category prompts to maximize Kendall's τ between the system ordering by synthetic and non-synthetic queries.

3.3 External Sources

Track participants were permitted and encouraged to leverage external sources beyond the Wikipedia corpus provided. However, because 150 of our test queries originated from the MS-ToT Dataset, participant groups were cautioned to not tune/train their systems using data from the MS-ToT Dataset or data scrapped from the "I Remember This Movie…" community Q&A site.

4 Results and Analysis

4.1 Participation

The TREC 2024 ToT track received a total of 18 submissions from 6 groups, including two baseline submissions from the track coordinators. Same as last year, participants were asked to report if they were certain that test data was not used to train their models. Excluding the baseline runs, this year none of the participants answered in the affirmative. This may reflect how it is becoming increasingly difficult to attest to such a claim in the absence of transparency around the data used for pretraining different LLMs that participants employed in producing their runs.

External Data Usage The submission form had an additional field for reporting if external data were also used. Nine runs, including the two baselines, used only the provided training data. Three runs used external datasets in addition to the provided training datasets, and six runs did not use the provided training data and leveraged other datasets exclusively. The "Training data" column in Table 1 reports what data was leveraged for individual runs.

Baseline usage The track coordinators submitted a BM25 baseline and a dense-retrieval baseline, which were made available to participants.[7] Baselines are marked with an asterisk in Table 1. Of the 16 submissions (excluding the baseline runs themselves), five submissions reported to have used the baseline runs either as re-ranking candidates or as negative samples. Runs which leveraged the baseline runs are marked in the column "Used baseline" in Table 1. All of these five submissions ranked among the top-six highest performing runs in this year's track.

4.2 Overall results

Table 1 contains the results of all submitted runs including the baselines. This year we observed a much wider spread of NDCG@1000 scores, as well as other metric scores, across the different runs. We plotted the distribution of metrics across different runs in Figure 2. The runs along the x-axis were sorted by their median score. Of the two submitted baselines, the dense retrieval baseline (*baseline-dense*) achieved the best performance.

The different performance metrics for the submitted runs seem to be well correlated to each as shown in Figure 3. The participating systems typically performed much better on synthetic queries for all three domains (movies, celebrities, and landmarks) compared to the MS-ToT movie queries as shown in Figure 4. However, in spite of the differences in absolute metric scores, the systems performance on the different query types are fairly well correlated as shown in Figure 5. This finding supports the use of synthetic queries for retrieval evaluation for ToT information needs.

[7] Please see `https://github.com/TREC-ToT/bench/` for more details about the baselines

Let's do a role play. You are now a person who vaguely remembers a place called {ToTObject}. You are trying to recall the name of the place by posting a verbose post in an online forum like Reddit describing the place. Generate a post of around 200 words about the place {ToTObject}. Your post must describe a vague memory of the place without revealing its exact name. People on the forum must have a hard time figuring out which place you are looking for. The answer should be difficult to find in search engines, so avoid using obvious keywords. I will provide you with some basic information about the place, and you must follow the guidelines to create a post.

Information about {ToT target name}: {first paragraph of Wikipedia article}

Guidelines:
MUST FOLLOW:

[1] Reflect the imperfect nature of memory with phrases that express doubt or mixed recollections, avoiding direct phrases like "I'm not sure if it is true, but".

[2] Do not directly specify the name of the place.

[3] Refer to the places in an ambiguous way using descriptions instead of names.

[4] Maintain a casual and conversational tone throughout the post, making sure it sounds natural and engaging without using formal structures.

[5] Provide vivid but ambiguous details to stir the reader's imagination while keeping them guessing.

[6] Use the provided information only as inspiration to craft a unique and engaging narrative, avoiding any direct replication of the given phrases.

[7] Start directly with your post, avoiding formal greetings like "Hello" or "Hey everyone."

[8] Start directly with your post, without describing your state of mind like "So, there's this", "I remember", "I've been thinking about".

COULD FOLLOW:

[1] Share a personal anecdote about your time at the place and the people you were with, but avoid common phrases like "When I was young." Instead, find unique ways to set the scene.

[2] Focus on sensory details like the overall mood, sounds, and emotional impact of being in the place, using vivid descriptions.

[3] Draw comparisons with other places or familiar experiences in a nuanced way that doesn't directly echo well-known locations.

[4] Introduce a few incorrect or mixed-up details to make the recollection seem more realistic and harder to pinpoint.

[5] Describe particular scenes or moments using ambiguous terms or partial descriptions.

[6] End the post by encouraging responses with genuine, open-ended questions for help.

Generate a post based on these guidelines.

Figure 1: Prompt to Synthesize Landmark ToT Queries.

Table 1: Summary of results. Best scores are in bold. The baselines are marked with an asterisk.

Run ID	Group ID	Training data	Used baseline	NDCG@10	NDCG@1000	MRR@1000	Recall@1000
rg4o_t100_test	h2oloo	Official only	✓	**0.8042**	**0.8159**	**0.7910**	**0.9250**
fs_test	h2oloo	Official only	✓	0.6785	0.7003	0.6417	**0.9250**
pr_test	h2oloo	Other only		0.5810	0.6204	0.5513	0.9133
dpr-pnt-lst-rerank	yalenlp	Official and other	✓	0.5708	0.6049	0.5482	0.8417
dpr-lst-rerank	yalenlp	Official and other	✓	0.5010	0.5424	0.4860	0.8400
dpr-router-lst-rerank	yalenlp	Official and other	✓	0.4664	0.5098	0.4515	0.8333
webis-tot-04	webis	Other only		0.4588	0.5008	0.4241	0.8200
ThinkIR_4_layer_2_w_small	IISER-K	Official only		0.3719	0.4239	0.3435	0.8067
ThinkIR_BM25	IISER-K	Official only		0.3679	0.4142	0.3438	0.7483
ThinkIR_BM25_layer_2	IISER-K	Official only		0.3555	0.4056	0.3332	0.7517
webis-tot-02	webis	Other only		0.3193	0.3712	0.2567	0.8200
baseline-dense*	Track coordinators	Official only	✓	0.1580	0.2377	0.1431	0.7450
webis-tot-01	webis	Other only		0.0652	0.1745	0.0566	0.8200
webis-tot-03	webis	Other only		0.0534	0.1705	0.0509	0.8200
baseline-bm25*	Track coordinators	Official only	✓	0.0912	0.1484	0.0849	0.5167
webis-base	webis	Other only		0.0487	0.1311	0.0444	0.6633
rag-sequence-nq	SUNY-BingU	Official only		0.0404	0.0924	0.0361	0.4300
ThinkIR_semantic	IISER-K	Official only		0.0018	0.0234	0.0021	0.1700

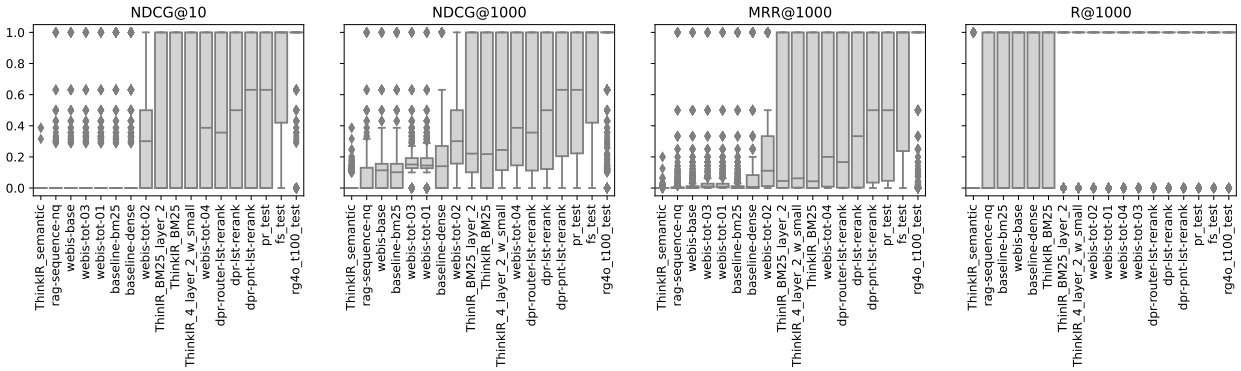

Figure 2: Metric distribution by run.

We plotted the performance of different runs across the 150 test queries of each type in Figure 6, where the x-axis is sorted by median score. From this plot, we can see that some queries were much easier to resolve compared to others, and on average the MS-ToT movie queries were more difficult than the three types of synthetic queries.

TSNE We plotted a TSNE plot of the runs in Figure 7. The TSNE reduction was performed on the NDCG@1000 scores for each topic. We can see that runs from the same group often cluster together.

Usage of external data As mentioned before, participants self-reported if external data was being used. We plotted the runs colored by whether external data was used in Figure 8. The top two runs used only the provided training data, followed by a run that exclusively uses other training data than the one provided. These are followed next by three runs that leveraged a combination of provided and other datasets for training.

References

J. Arguello, A. Ferguson, E. Fine, B. Mitra, H. Zamani, and F. Diaz. Tip of the tongue known-item retrieval: A case study in movie identification. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, CHIIR '21, page 5–14, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380553. doi: 10.1145/3406522.3446021. URL `https://doi.org/10.1145/3406522.3446021`.

J. Arguello, S. Bhargav, F. Diaz, E. Kanoulas, and B. Mitra. Overview of the trec 2023 tip-of-the-tongue track. 2024.

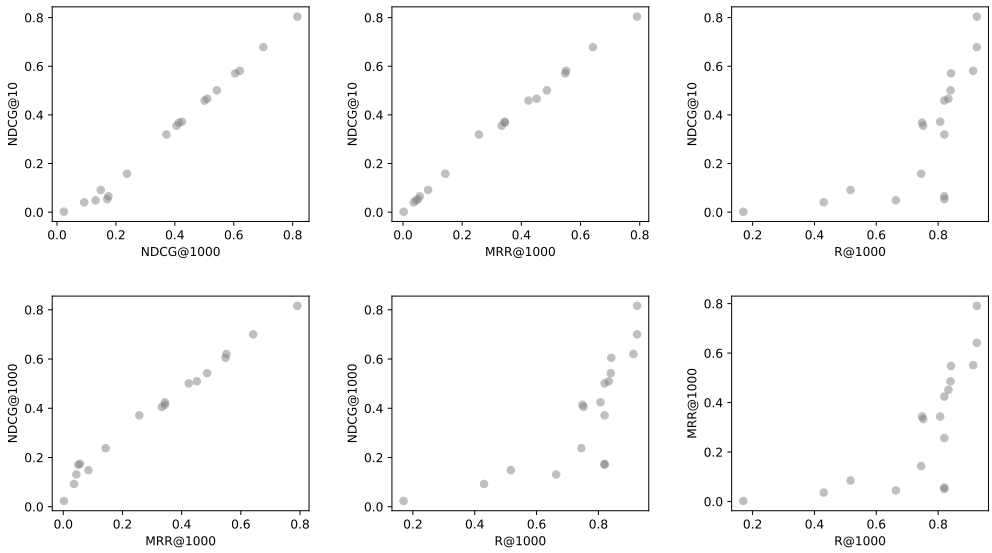

Figure 3: Metric correlations. Each dot represents the mean performance of a run, with the different metrics on the axes. We observe strong correlations between NDCG@10, NDCG@1000, and MRR. Recall also seems to correlate but less strongly with the other three metrics.

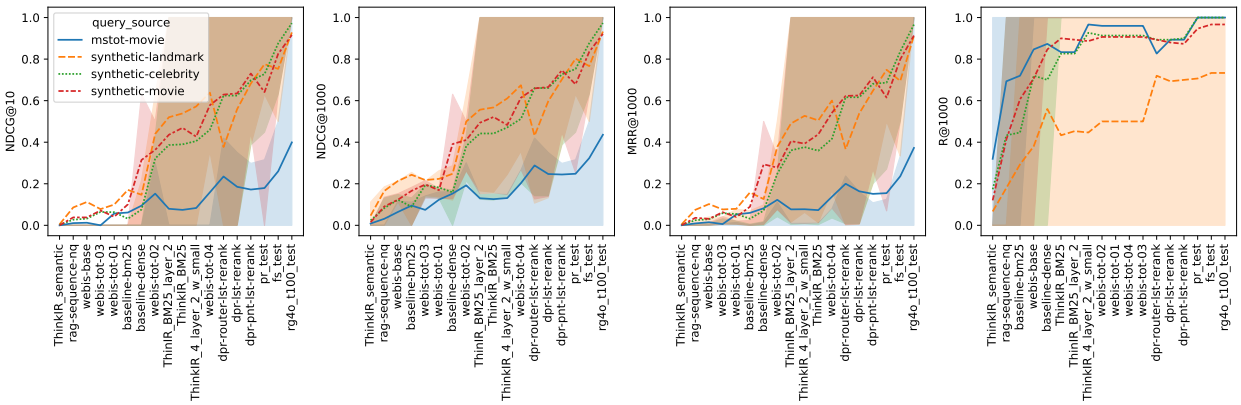

Figure 4: Mean run performance by query type.

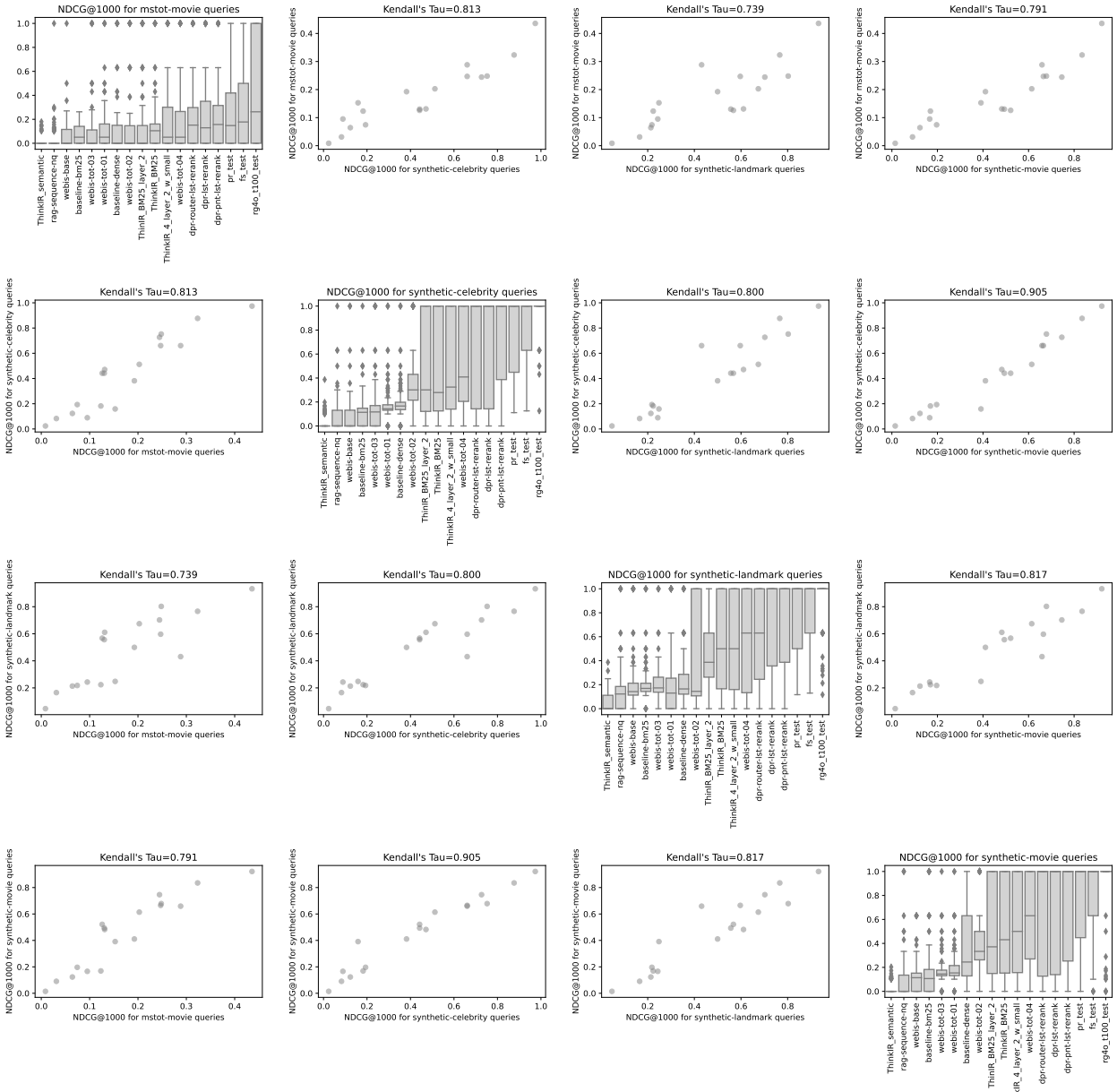

Figure 5: System performance correlation across query types.

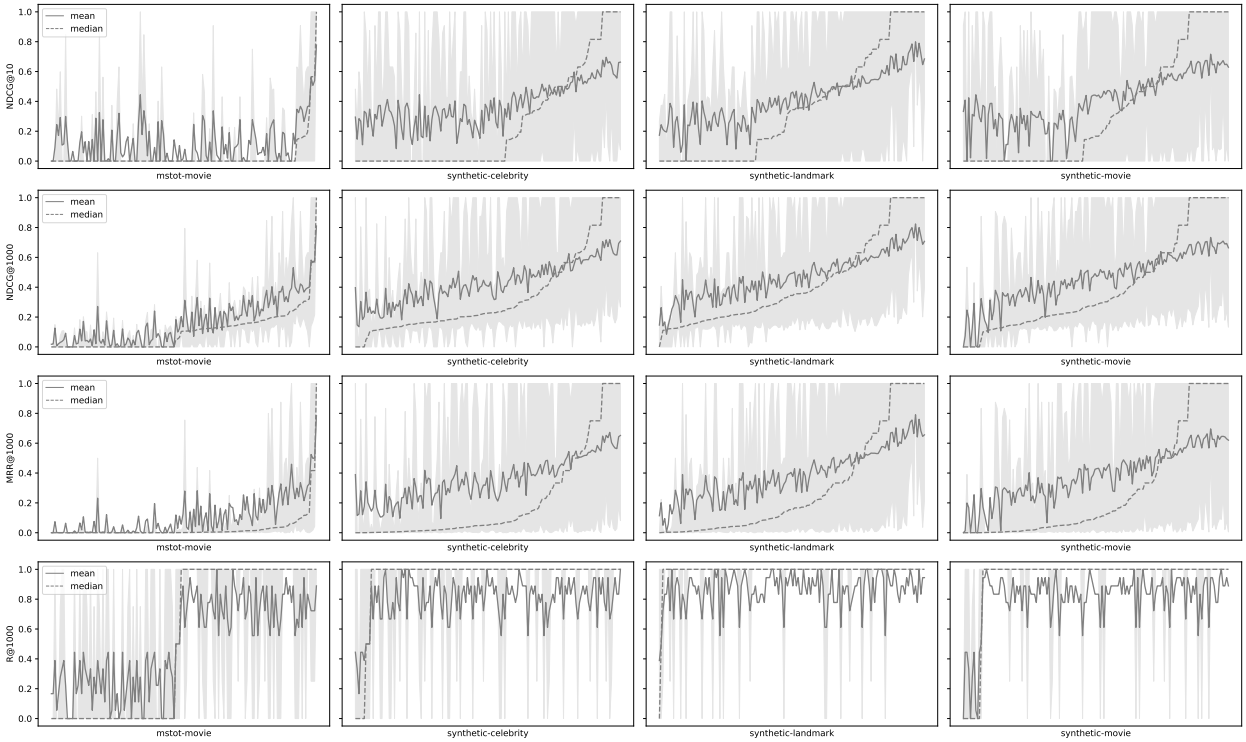

Figure 6: Metric distribution by query.

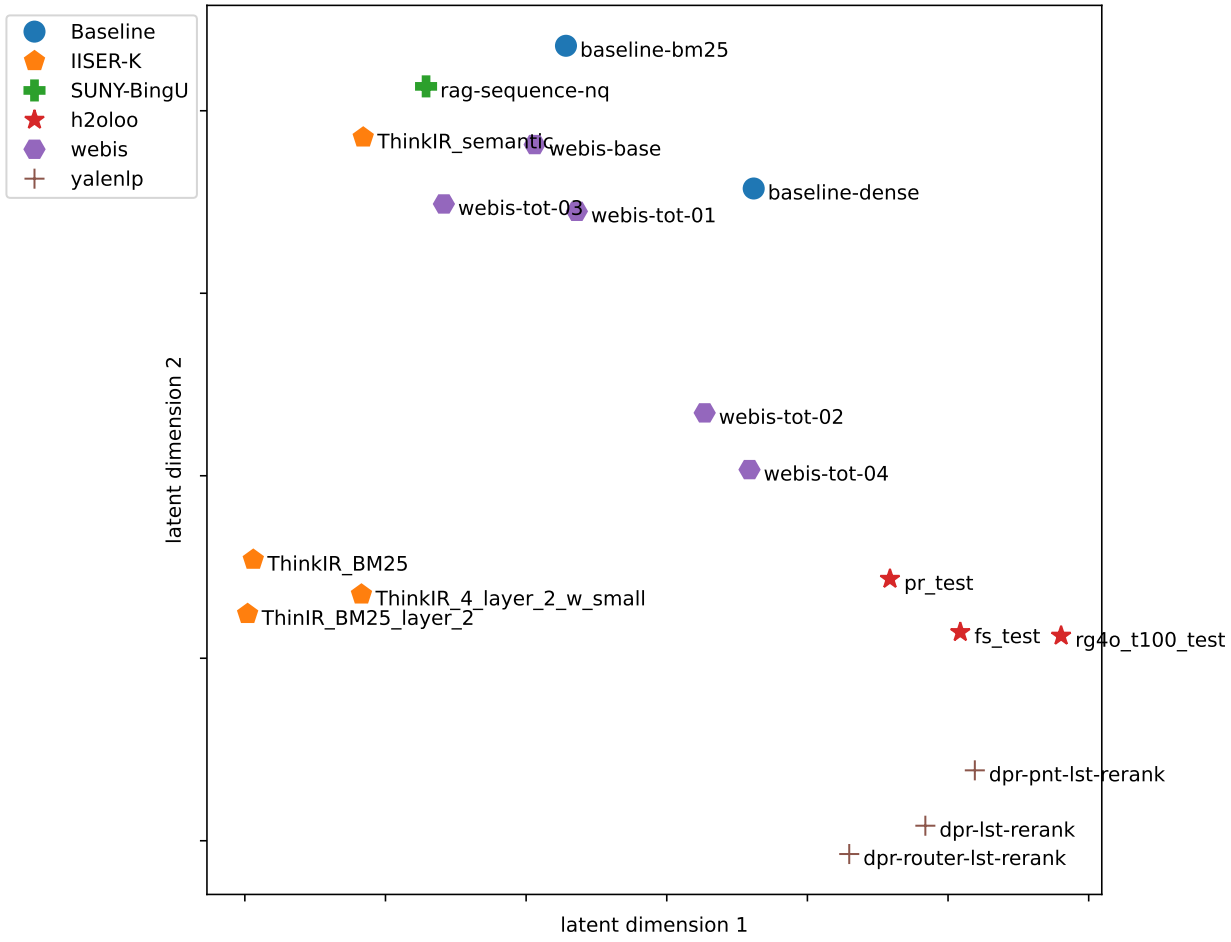

Figure 7: TSNE plot based on the NDCG@1000 scores of different runs.

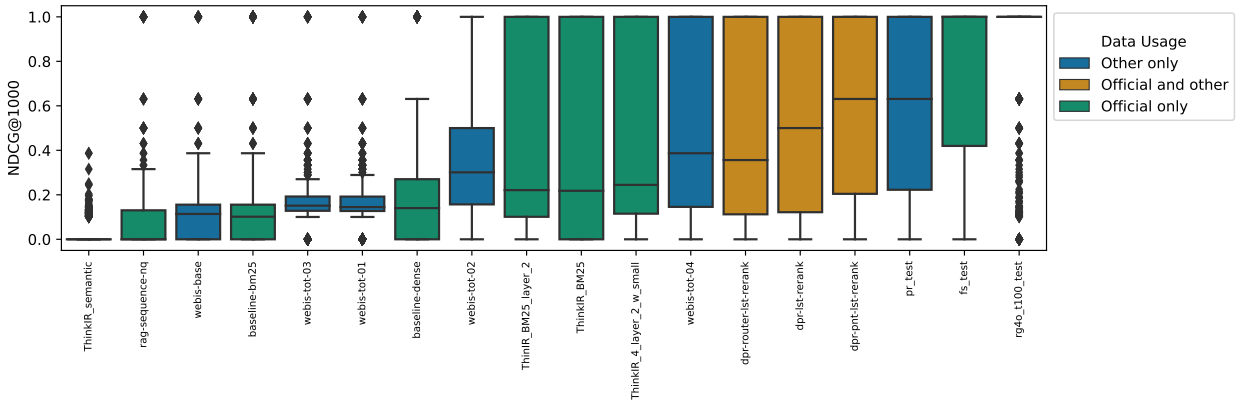

Figure 8: Distribution of NDCG@1000 scores for different runs, with colors indicating if external data was used to produce the run.