How Variability Influences Podcast Search: Queries, Transcriptions, and Judges

Watheq Mansour
The University of Queensland
Brisbane, Australia
w.mansour@uq.edu.au

J. Shane Culpepper
The University of Queensland
Brisbane, Australia
s.culpepper@uq.edu.au

Andrew Yates
Johns Hopkins University
Baltimore, United States
andrew.yates@jhu.edu

Joel Mackenzie
The University of Queensland
Brisbane, Australia
joel.mackenzie@uq.edu.au

Abstract

Podcasts have continued to grow in popularity over the last two decades, with more than 4.52 million podcasts and 584 million listeners across the globe in 2025. Developing effective search systems for web-scale podcast corpora is of vital importance. Previous research has approached the search task primarily through text representation using a single transcription of the audio content via automatic speech recognition (ASR) models. However, there is currently limited understanding about how variation in podcast representations influences ranking, retrieval, and relevance assessment.

In this paper, we provide a comprehensive analysis of the effectiveness of the podcast search problem from three different sources of variation. First, we create a set of 5,745 unique, automatically generated *query variations* derived from the TREC 2020 and 2021 podcast track topic descriptions. Second, to enable a comprehensive and reproducible analysis, we build a diverse pool composed of a set of 17 different ranked retrieval runs (*system variations*) using state-of-the-art (SOTA) information retrieval (IR) systems for all of the query variations over a set of *corpus (transcript) variations*. Third, we use five large language models (LLMs) to automatically assess the relevance of query-document pairs resulting from every query-system-transcription combination, yielding over 1.8 million judged pairs. Our analysis studies the impact of these variations on both search and judgment effectiveness, resulting in a series of practical recommendations for podcast and audio retrieval tasks.

CCS Concepts

• **Information systems** → **Query representation**; **Test collections**; **Relevance assessment**; **Multimedia and multimodal retrieval**.

Keywords

Podcast retrieval, Document variations, Query variations, Automated relevance judgments

SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809668

ACM Reference Format:
Watheq Mansour, J. Shane Culpepper, Andrew Yates, and Joel Mackenzie. 2026. How Variability Influences Podcast Search: Queries, Transcriptions, and Judges. 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, 12 pages. https://doi.org/10.1145/3805712.3809668

1 Introduction

Podcasts are a popular medium for listeners, allowing them to engage with topics and communities of interest. As of September 2025, there are 4.52 million podcasts available worldwide in over 100 languages [1], making them a valuable source of knowledge.

While several studies have investigated various aspects of at-scale podcast retrieval, including user behavior [18] and multi-modality [38], the standard approach commonly relies on indexing a single automatic speech recognition (ASR) transcript to represent the underlying podcasts. This "single-transcript" approach implicitly assumes that the ASR model is correct, ignoring the inherent noise in spoken document processing. While remarkable improvements have been made in audio-to-text encoding, aspects such as speaker dialect, gender, and rare words continue to limit the performance of podcast search systems [81]. Recently, Mansour et al. [57] demonstrated that the choice of ASR model has a significant impact on retrieval performance, primarily due to the failures made by the ASR model when transcribing named entities, poor quality audio, or strong accents. However, this study was limited to the existing TREC Podcast document pool [45, 47], which has a large percentage of (potentially relevant) unjudged documents.

To mitigate these transcription issues, Carterette et al. [19] highlighted the importance of augmenting podcast indexes with *metadata* – such as episode titles or descriptions. Since metadata is often manually curated, it can complement noisy transcripts, providing a cleaner signal to improve retrieval effectiveness.

ASR-based document representations are also used for *judgment*. For example, Mansour et al. [58] demonstrated that system orderings generated by LLM-based judgment are highly correlated to those from human experts, confirming prior studies on classic text ad-hoc search [84, 89]. Again, this study was limited to the available TREC judgment pool and existing Spotify transcripts, and the relevance assessment process relied solely on the text representations of the audio segments – audio segments and metadata were not utilized.

Finally, another critical but underexplored dimension of prior work is the effect of query variation. While recent work has demonstrated that user query formulations can contribute as much variance to a judgment pool as varying retrieval systems [12, 61], the impact of query rewriting and variation has not been systematically analyzed within the audio retrieval domain.

Collectively, these limitations suggest that findings drawn from current podcast search benchmarks may not generalize to realistic retrieval scenarios. By relying on a single, potentially erroneous transcript, ignoring the variance inherent in user query formulation, and restricting relevance judgment to text-only representations, standard evaluation paradigms likely over-fit to specific system configurations. Consequently, there remain fundamental unknowns regarding how variability in queries, transcriptions, and judges influences ranking and retrieval, necessitating a comprehensive analysis of the podcast search problem from all three perspectives. In this work, we address this gap by answering the following research questions:

- What is the relative influence of system, query, and transcription variations on the variety of the judgment pool?
- How are LLM-based assessments influenced by the underlying document representation and modality?
- How does the inclusion of podcast metadata affect LLM judgment, and does it lead to higher agreement with human experts?

We conduct a comprehensive analysis of a large Podcast test collection and study variation effects on performance through query and transcript dimensions. Figure 1 summarizes the methodology used in this study. We first use multiple ASR models to generate transcript variations. Then, we employ LLMs to generate query variations based on the topic backstories. Using the generated transcriptions and query variations, we produce 17 runs using 11 state-of-the-art (SOTA) retrieval models to create a large pool of passage judgments using five different Large Language Models. Our contributions are as follows:

- We demonstrate that the variance caused by ASR models is a critical, yet overlooked, source of pooling bias. Our analysis reveals that varying the ASR model used to create the corpus alters the pool as significantly as varying the retrieval system used, suggesting that future test collections must account for transcript variance to be reusable;
- We show that systems are highly sensitive to both the underlying transcription quality, as well as the user query, and confirm the importance of including metadata when transcription quality cannot be guaranteed;
- We show that LLM judgment is robust to variation in the transcripts, suggesting that LLM judges are consistent in the presence of noise; and
- We provide practical recommendations for multimodal evaluation, showing that audio-aware LLM judges can resolve disagreements where text-only protocols fail.

Our experiments result in a set of over 1.8 million labeled topic-document pairs from five LLMs by incorporating new query, transcription, and system variations over the TREC Podcast collections.[1] We believe that our work is a useful contribution to multiple IR research areas of interest, including (1) The representation and ranking of podcast data; (2) The use of LLMs to assess the relevance of topic-document pairs using different modalities; and (3) The effects of system, query, and document content variation on IR experiments and pooling using the Cranfield paradigm [22].

[1]The prompts used, LLMs assessments, query variants code and data, and our pool runs are all available here https://github.com/Watheq9/sigir26variability

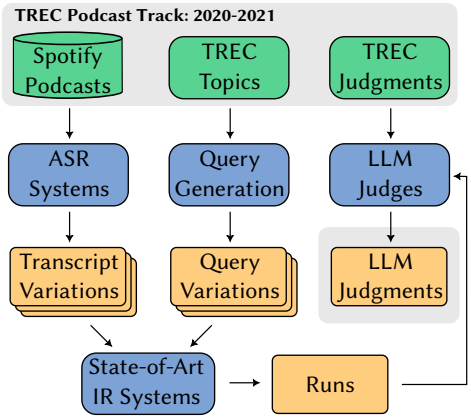


Figure 1: An overview of the analysis setup. Multiple automatic transcriptions are generated from the Spotify audio podcast collection, which are then indexed using multiple IR systems; Query variations are generated and used to perform top-k retrieval for each system; Finally, the runs are pooled, and the relevance is judged using multiple large language models.

2 Background and Related Work

We now discuss key related work, including datasets, existing work on podcast retrieval, and the use of large language models for relevance assessment.

Audio Datasets. Between 1997 and 2000, the first spoken retrieval task was offered in TREC spoken document retrieval track [36], where manual and automatic transcripts of news were collected from radio and television broadcasts. In subsequent work, several shared tasks have focused on speech retrieval of personal testimonies [64], lectures [4–7], and television shows [29, 30, 50, 77]. Jones [44] provides a comprehensive overview of spoken text retrieval in a more general context. The TREC 2020 [45] and 2021 [47] Podcast moment retrieval and summarization tasks use the Spotify podcast dataset [23], a collection of 100,000 English podcasts. In recent work, Setty and Becker [78] created a podcast fact-checking and annotation tool along with transcripts of 531 podcast episodes. However, the size of the dataset limits its usefulness for other search and recommendation tasks.

Podcast Retrieval. Podcasts are an important medium for sharing and consuming content for a wide variety of special-interest topics. Given the variety of active podcasts, providing efficient and effective information-discovery tools is an important but underexplored research direction [18, 85]. Key challenges in developing effective podcast search systems include: handling rich data formats [9], the varying level and granularity of information needs (podcast level, episode level, or moment level), and the huge variety in content, to name a few [46, 87].

One popular strategy for indexing podcasts – and many other spoken document formats such as radio broadcasts – is to use *automatic speech recognition* (ASR) models to generate text transcriptions [19, 36, 79, 86]. This enables SOTA text retrieval technologies to be directly applied to podcast retrieval tasks without significantly changing the indexing or retrieval engine. This was the main strategy used by participants of the TREC 2020 and 2021 Podcast track [45, 47], which challenged the community with two distinct shared tasks – podcast moment retrieval, where the best *entry point* into a podcast is retrieved (formulated as a ranking problem); and podcast summarization, where the key ideas presented in a podcast episode are captured as a short form text snippet. Mansour et al. [57] examined the effect of changing the underlying text representation by using different ASR transcripts, and showed that the variance in transcriptions generally leads to a loss in effectiveness. We recommend interested readers see the work of both Tian et al. [85] and Jones et al. [46], who describe the unique challenges presented by podcast search and information access at scale, including issues relating to users and user behavior, robust evaluation, and multi-modality.

Sources of Pooling Variation. In Cranfield-based experiments [22], a fixed test collection is provided to a diverse set of participants who then build retrieval systems to form sets of *system runs* for the topics provided. However, the information need for each topic is usually represented by a single, short keyword query matching the need. Consequently, the coverage of documents pooled during relevance assessments is directly proportional to the number and diversity of the system runs created. Recent work demonstrates that users tend to be as diverse as retrieval systems [61]. In other words, the way in which individual users formulate queries to satisfy a given information need can result in more variance than is typically produced by using a single query with a diverse set of IR systems [12, 13, 60]. Therefore, there is a growing belief that we should rethink the "single query, multiple system" model that is currently used in exercises such as TREC, and to instead consider more descriptive information needs to model and optimize search engine quality [25]. One clear downside, however, is that greater variance means that many more relevance assessments are required. However, this additional cost could be worthwhile if it improves the reusability of test collections.

Relevance Assessment Using Large Language Models. In early 2023, Faggioli et al. [31] explored the advantages and disadvantages of using LLMs to generate relevance judgments for a test collection. Faggioli et al. experiment with two LLMs (OpenAI's GPT-3.5 and YouChat), using the TREC-8 ad-hoc retrieval task and TREC 2021 Deep Learning track (TREC-DL 2021), using two simple prompts (with minimal prompt optimization). Interestingly, the authors found that the overlap in agreement between human judgments and the LLM was 47% for GPT-3.5 and 33% with YouChat. When reassessing the original judgment pool for TREC-DL 2021, they observed a modest correlation between highly-trained human assessors and the fully automated LLM approach, which yields similar system orderings despite the lack of high overall agreement between the two approaches.

Upadhyay et al. [89] recreated the results from Thomas et al. [84] in a reproducibility study using OpenAI GPT-4o, and released an open-source toolkit called *UMBRELA* to aid researchers interested in this problem. Experiments using the TREC Deep Learning Tracks from 2019–2023 demonstrate that system rankings created from LLM labels are highly correlated to ordering produced from human ground-truth judgments. Interestingly, Upadhyay et al. showed several corner cases where LLM judgments were actually more accurate than the corresponding human judgments, which they attributed to the unreliability of human assessments or a lack of a clear description of a user's information need. By comparing LLM and human judgments, Upadhyay et al. [88] observed that *LLM judgments could replace human assessments when using many common IR effectiveness metrics, when the overall effectiveness ordering at the run level is being measured.* When comparing agreement at the judgment level, they found that human assessors apply more stringent relevance criteria than LLMs currently do. These findings corroborate the results reported by Faggioli et al. [31], where LLMs were found to be less reliable when attempting to distinguish relevance between the grades of 1 and 2 in graded relevance assessment exercises.

Other recent work explores ensembling automatic LLM judgments to improve reliability [71], and how to use LLMs to fill "holes" in judgment pools [2, 53]. The first *LLM4Eval* workshop contains additional findings and references for interested readers [70].

While there have been a range of empirical studies that support the use of LLMs for relevance assessment exercises, there are also studies that support the opposite point of view. For example, Clarke and Dietz [21] criticize claims that LLMs can replace humans [88], and provide several concrete cases where LLM-based assessments fail to reliably identify the best-performing systems. Alaofi et al. [8] demonstrate that simple keyword stuffing techniques can mislead LLMs and produce incorrect labels. Faggioli et al. [32] consider the wider role of humans and LLMs in the relevance assessment process.

In this work, we provide additional evidence in favor of employing LLMs to automate relevance assessments, but cautiously consider the limitations presented in previous work that warn against using this technique. More specifically, we use multiple LLMs with query variations and make all of the assessments we have created publicly available for future research and analysis on this important problem. We also carefully compare the automated judgments we have generated against ground truth judgments created by human assessors for our task.

3 The Variations Setup

This section describes the design choices made when constructing the variations and reports the key characteristics.

3.1 Document Corpus

First, we describe the document corpus variation process and its basic properties.

Spotify Podcast Corpus. Given our desire to use a large-scale podcast dataset, we use the Spotify podcast collection [23]. Originally published in 2020 for the TREC podcast track [45, 47], it consists of 100,000 English podcasts published between 2019 and 2020 on the Spotify platform; for more details on the specific filtering criteria applied to create this sample, please refer to the original description [23]. The episodes constitute around 60,000 hours of audio, consuming close to 4 terabytes of disk space. In addition to the raw audio, the organizers also shared an automatic transcription file using Google's Speech-to-Text and provided author-curated metadata for each episode when possible.

Transcript Variations. Mansour et al. [57] explored the effect of applying ASR systems of varying quality to the Spotify podcast corpus. In particular, they re-transcribed all 60,000 hours of audio using three new ASR models, and demonstrated that the quality of the ASR model had a large effect on search quality. However, a key problem with this work is the inherent bias in the TREC judgments to systems using the original Spotify transcript: running retrieval systems on alternative transcripts returns a large number of unjudged documents. This is a key motivation for this work. We group the four transcripts available in the literature and augment them with an additional transcript using an SOTA ASR model, resulting in *five* distinct representations of each podcast:

- `Spotify`: This is the original transcript generated by Google's speech-to-text API in 2019, which was used by most teams in the 2020 and 2021 TREC Podcast tracks.
- `WhisperX`: WhisperX is a SOTA open-source speech recognition model [15] derived from the OpenAI *Whisper* models [68], but is more efficient and effective. We generate `WhisperX` transcriptions using the `LargeV3` version and reproduce the transcriptions used in Mansour et al.'s work using the `Base` version [57].
- `Silero`: Silero is a family of pre-trained lightweight speech-to-text models suitable for real-time ASR tasks.[2] We employ the `small` and `large` models, which were used in prior work [57].

Segments and Statistics. After transcription, each podcast episode is broken into a series of *segments* that span 120 seconds, with a 60 second overlap, mirroring the process used in the original Spotify collection [23]. This results in around 3.5 million segments per transcription, with a total of 17.7 million segments in total.

To the best of our knowledge, this is the first example of a document collection that contains *document variations*. That is, the content of each document can vary between different transcription models, even though the underlying *source of truth* – in this case, the raw audio – remains the same. Figure 2 shows a concrete example of this variance.

3.2 Topics and Query Variations

Next, we outline the topics that accompany the corpus and the process used to create the query variations.

Topic Descriptions. Given our desire to reuse existing resources as much as possible, we reuse the 2020 and 2021 TREC Podcast Track topics [45, 47]. These topics are suitable as they were specifically crafted for the podcast data available. There are 48 and 50 topics in 2020 and 2021, respectively. Each topic is classified into one of three possible types: (1) *topical*, which are targeted at discovering podcast segments related to a given topic; (2) *refinding*, which simulate a user trying to find a segment they have already listened to; and (3) *known-item*, which simulate a user looking for a segment they believe to exist, but one that they have not yet discovered.[3] In the 2020 topics, there were 35 topical, 6 re-finding, and 7 known-item, respectively. In the 2021 topics, 40 were topical, and the remaining 10 were known-item. As is typical with TREC topics, each is accompanied by a short "title" query, and a longer description outlining the information need (backstory).

[2]See: https://github.com/snakers4/silero-models
[3]Refinding and known-item were merged into known-item in 2021

Table 1: A randomly selected backstory and four random query variations generated by the different LLMs and settings.

Backstory: *There have been numerous news reports of bees and other pollinators dying. I want to learn what I can do personally to help bees and other pollinators. General discussions about habitat loss are not relevant.*

Model	Temp	Query
`GPT-3.5-turbo`	0.0	how to assist pollinators
`GPT-3.5-turbo`	0.5	supporting bee populations at home
`GPT-4o`	0.0	bee support initiatives
`GPT-4o`	0.5	helping bees in my area

Generating Query Variations. Recent work has made it clear that query variations, either generated by users [56, 61], or algorithmically [41, 65, 72], can greatly increase the number of unique documents retrieved by IR systems [12, 61], sparking further work on understanding how IR systems should handle query variations [14, 17, 35]. By providing a set of unique queries – rather than one single representative query – document pools can be made more robust by including additional documents in the judgment process. As such, we opted to generate query variations for each topic in the 2020 and 2021 TREC Podcast Track topics. To generate these variations, we followed the setup of Alaofi et al. [8], with a few important changes. Firstly, we treat each of the available TREC topic descriptions as a so-called "backstory" representation of the user information need. Then, we employed the `UQV100` collection [13] to create in-context examples for the query generation task. In particular, `UQV100` consists of 100 backstory descriptions, each of which has 58 unique user-generated queries on average. We randomly sample 5 of these backstories – each with 20 associated user queries, randomly sampled without replacement – and provide these to the LLM in our prompt template. Then, we also provide the target backstory, and ask the LLM to generate a list of query variations.[4] We used `GPT-4o` and `GPT-3.5-turbo` for this experiment, and repeated the experiment with the temperature set to 0 and 0.5, requesting twenty variations per prompt. We also ensure each of the TREC title queries is incorporated to ensure better coverage of the original podcast queries. On 25 occasions, the generative LLM did not respond with any meaningful queries, instead responding with a *"sorry, but I can't ..."* message, often citing specific guardrails [28]. Interestingly, this was only an issue with the `GPT-3.5-turbo` model, and did not occur with `GPT-4o`.

Query Properties and Statistics. After de-duplicating and combining the outputs of these experimental results, we ended up with a total of 5,745 query variations for the 100 topics. Each topic has between 26 and 75 variations, with a median of 54.

3.3 System Pooling

To make the pool as exhaustive as possible, we utilize a variety of SOTA retrieval models in two stages: (1) Initial retrieval (candidate generation), and (2) Document re-ranking. In the first stage, we include nine retrieval models from four categories of systems, and, in

[4]The full prompt template, in-context examples, and code for generating the variations are provided within the GitHub repository.

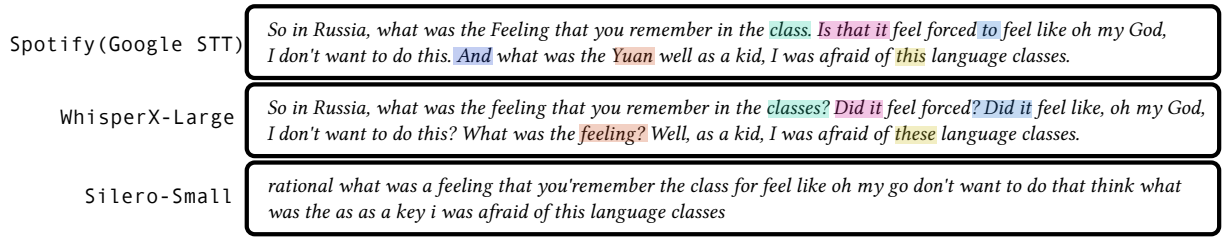


Figure 2: An example snippet created from three different ASR methods. Differences between the `Spotify` and `WhisperX` transcriptions are color-coded; the `Silero` transcription is not color-coded as it is significantly different (and found to be lower quality) than the former two.

the second stage, we re-rank eight of these runs using two different re-rankers. This leads to a pool of 17 runs for each of the five transcribed corpora.

Statistical Lexical Retrieval. We use three models from two families of weighting models. In particular, we select BM25 from the TF-IDF family [73, 74], and both DPH and PL2 from the Divergence from Randomness family [10, 11]. We use the default parameters for all models.

Term-Based Query Expansion. We use the traditional – yet strong – combination of BM25 with RM3 expansion [3]. We set the number of feedback terms and documents to 10 and 3, respectively.

Single- and Multi-Vector Dense Representations. We also include three models from two dense paradigms, namely single-vector and multi-vector dense representations. In the single-vector similarity paradigm, an embedding vector is generated using a pre-trained language model to embed the query and document, and then relevance is estimated as the dot product of these two vectors. However, in the multi-vector late interaction paradigm, an embedding vector is produced for every token in both queries and documents, creating two sets of vectors, and then the relevance is estimated between these two sets as the sum of maximum similarities between vectors [48]. We select two variants of the BGE family [20, 90] to represent single-vector similarity, BGE-large-en-1.5 and BGE-m3, due to their superior performance on MS-MARCO [16], and we choose ColBERT-v2 (the improved and efficient version of ColBERT) to represent the multi-vector late interaction paradigm [75, 76].

Learned Sparse Representations. Learned sparse models generally learn term weighting and/or expansion using pre-trained language models, and are then represented by inverted indexing techniques. SPLADE [34] is one of the strongest families of the learned sparse representations. We use two strong variants, namely SPLADE++ [33] and SPLADE-v3 [51].

Re-Ranking Models. We also employed second-stage re-ranking models in an effort to improve early precision. We used two types of re-rankers:

(1) **Point-wise:** The T5-based [69] MonoT5 cross-encoder [63]; this re-ranker was applied to the top $k = 100$ segments as retrieved by ColBERT-v2, SPLADE-v3, BM25+RM3, BM25, and DPH, respectively.
(2) **List-wise:** The LLM-based RankZephyr [67] model, with 7 billion parameters. Since this model is more expensive than the point-wise approach, we applied it on the top $k = 40$ segments as retrieved by the MonoT5 re-ranked output of ColBERT-v2, SPLADE-v3, and BM25, respectively, representing an expensive – yet hopefully accurate – three-phase ranker.

Each of these systems was executed on every query variation – over each corpus transcription – to form a comprehensive pool of documents for assessment. The retrieval depth is set to 200 for statistical lexical and term-based expansion models, and to 100 for all other models. All top-k run files are available in the GitHub repository.

Tools. We use the PyTerrier toolkit [54, 55] to index the corpora, run the lexical retrieval models, and evaluate the runs. We use RankLLM [66] for RankZephyr, and llm-rankers [92] for MonoT5.

3.4 Assessing the Judgment Pool

Given the scale of our judgment task, we employ multiple open-source and proprietary LLMs as a low-cost judgment technique. This also allows us to explore the effectiveness differences between LLMs and ensure that the judgment process is reliable.

Large Language Models. To judge the pool, we select four open-source LLMs and one proprietary model due to their strong performance on a variety of LLM-based tasks. In particular, we use OpenAI's `GPT-4o` as a closed model, as it achieves the highest agreement with human judgments in recent studies [8, 88]. For open-source LLMs, we choose the following four models:

(1) `Mistral`: A 6-bit quantized model from the wider Mistral family [43] – `Mistral-Small-Instruct-2409-Q6_K_L`;
(2) `Qwen`: A family of models from Alibaba [91] – we employed the `Qwen2.5-14B-Instruct-Q8_0` model, quantized to 8-bits;
(3) `Llama3`: Meta's `Meta-Llama-3.1-8B-Instruct-Q8_0` model [83] with 8-bit quantization; and
(4) `Gemma2`: Google's `gemma-2-9b-it-Q8_0` [82] with 8-bit quantization.

We use `llama-cpp-python`[5] – Python Bindings for the `llama.cpp` library[6] – to conduct inference on the open-source LLMs. All models are quantized and downloaded from Bartowski's publicly available collection.[7] We also conducted audio-based experiments where SOTA audio LLMs were used to rejudge the relevance of TREC Podcast relevance judgments – more detail is provided in Section 4.4.

[5] https://github.com/abetlen/llama-cpp-python
[6] https://github.com/ggml-org/llama.cpp
[7] https://huggingface.co/collections/bartowski/

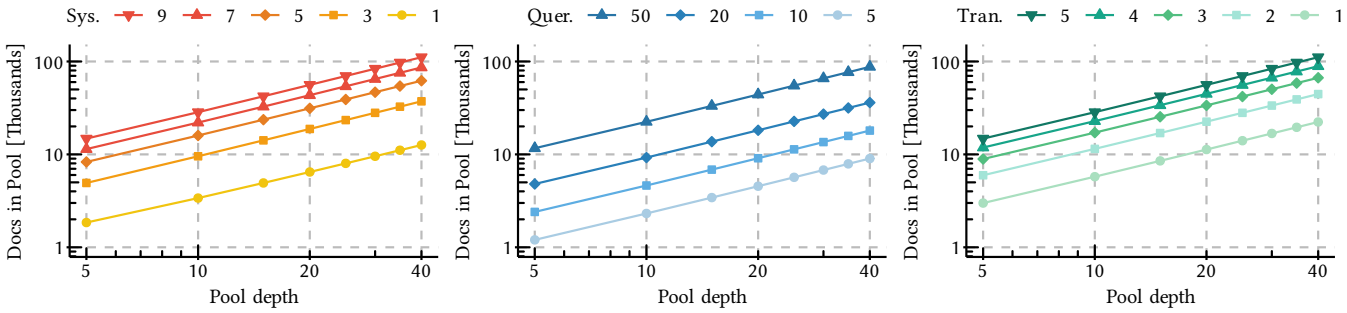


Figure 3: The effect of system (left), query (middle), and document transcription (right) variations on the size of the pool, averaged over all 100 topics. Interestingly, varying the transcription model can increase the number of pooled documents as much as including additional system or query variations. Note the logarithmic scales on both axes.

LLM Prompting and Relevance Grades. To generate new relevance judgments, we adopted the prompt used by Mansour et al. [58]. We also restrict the LLM output to JSON format, as Tam et al. [80] showed that this restriction enhances the performance during classification tasks. The relevance grades are between 0 and 3. The prompt is available in our repository.

LLM Judgments and Statistical Properties. Using the prompt described above, we deployed each of the five LLMs to judge the top-10 documents for every query-system-transcription combination. This required us to judge approximately 70,000– backstory–segment pairs *for each transcription model*, as the judgment process may be affected by the quality of the underlying transcription model used. However, only `GPT-4o` assessments for `Silero` pools are not included due to the prohibitive cost, as we found `Spotify` and both versions of `WhisperX` to be of higher quality (see Figure 2). So, using five LLMs, we have judged 72,647 and 69,173 pairs from `WhisperX` (base and large) pools, and 75,573 pairs from `Spotify` pool, and using the four open-source LLMs, we have judged 98,260 and 86,038 pairs for `Silero` (small and large) pools. In total, we have judged 1,824,157 pairs. It is worth noting that many of these pairs are derived from the same backstory and source document, but contain a different document representation produced by the underlying transcription model (as shown in Figure 2). The goal is to explore *how the document transcription quality can influence the judgment labels.*

TREC Judgments. To further validate our LLM judgment process, we have exhaustively re-judged the entire set of labels from the TREC 2020 and 2021 podcast tracks using the same LLMs, while also incorporating judgments from audio-based LLMs based on the raw audio (discussed in Section 4.4). Although the original runs contributed to TREC used a single transcription – the `Spotify` transcription, we have rejudged every pair using all five transcriptions, as we believe the transcription quality can directly influence the decision process. In Section 4.2, we use the new judgments to compare directly between the LLM judgments and the human TREC assessors.

3.5 Summary

To recap, we have five unique transcriptions of the Spotify document collection generated using various SOTA ASR models. We then used the TREC podcast topic descriptions to automatically generate a large set of query variations for each topic, and ran these query variants using a large set of SOTA retrieval systems to generate a diverse pool of runs. We then judged these runs using several different large language models, and the resulting set of judgments is curated into the variation setup of our experiments. Figure 1 provides an overview of the entire process and the respective outputs.

4 Experiments and Analysis

In this section, we describe experiments and analysis on the factors contributing to judgment pool size, the stability of judgment over varying transcripts and modalities, and the stability of systems under both query and transcript variations.

4.1 Pool Variance

Our first experiment aims to quantify the effect of including transcript variations in the document pooling process. To the best of our knowledge, we are the first to explore the effects of *corpus variations* for Cranfield experiments. Our context is podcast retrieval, and ASR models of varying quality are considered [57].

To measure the effect on collection quality, we take the set of nine first-stage retrieval systems and run all of them on all transcription models using a subset of n randomly selected query variations (with $n \in \{5,10,20,50\}$). Then, we measure the average number of unique documents added to the judgment pool when varying:

(1) The number of contributing systems (across all transcriptions and query variations);
(2) The number of contributing query variations (again, using all systems and transcriptions); and
(3) The number of contributing transcriptions (over all system and query variations).

In each experiment, we increase the number of contributing variants by sampling randomly without replacement, using the previous set as the basis for each subsequent experiment. For example, if system A was used for the "1" line in the leftmost figure, then system A would be included in the "3", "5", "7", and "9" lines as well. Figure 3 demonstrates that the pool grows as a function of the pool depth selected – that is, the number of documents taken from each ranked list and added to the judgment pool. Somewhat surprisingly, we observe that transcript variations tend to exhibit similar behavior to other sources of pool variation, including system and query changes. This experiment also validates and generalizes prior work

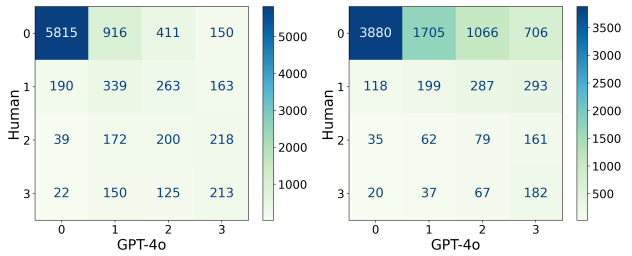


Figure 4: The confusion matrix for the relevance assessments generated by `GPT-4o` compared against human assessments for TREC 2020 (left) and TREC 2021 (right). While the agreement is high for the 2020 topics, it is much lower for the 2021 topics. All open-source models exhibited similar trends; `Mistral` was typically the most similar to the human assessors.

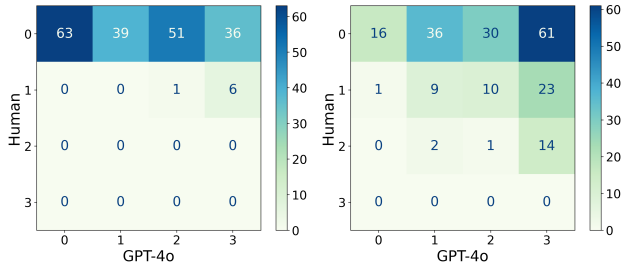


Figure 5: The confusion matrix for the relevance assessments generated by `GPT-4o` compared against human assessments for topic "13" in 2020 (left) and topic "85" in 2021 (right). We select these topics as they receive the lowest correlation scores in each year.

that suggests that queries are as important as systems in pool diversity [56, 60, 61], which can affect the reusability of document collections incorporating corpus variations.

4.2 LLM Judgment Quality

Our second set of experiments addresses the suitability of deploying LLMs for the judgment task using our collection. In particular, we examine the agreement between the human TREC assessors and the rejudged labels using LLMs. We seek to answer the second research question: *how are LLM-based assessments influenced by the underlying document representation and modality?*

Confusion Matrices. Figure 4 shows the confusion matrices for `GPT-4o` assessments against TREC (human) relevance assessments for both TREC 2020 and 2021 pools. To ensure a faithful comparison against the TREC human assessors, this experiment uses the same `Spotify` transcriptions as used in the original TREC Podcast track for judgment. Although the LLM-generated labels appear to mostly agree for the 2020 test collection, the 2021 test collection provides a very different outcome. In particular, all LLM judges appear to overestimate relevance for a large fraction of the pairs that human assessors determined to not be relevant. This observation agrees with prior work [31, 88], which demonstrate that LLMs are often less strict than human assessors. However, we also note that both Upadhyay et al. [88] and Mansour et al. [58] provide evidence suggesting that LLMs can often make inferences that may be missed by humans, resulting in more accurate labels.

Pairwise and Run-Level Agreement. Table 2 and Table 3 (focusing on just the `Spotify` rows) show the inter-rater agreement between each LLM and the human assessor, using Krippendorff's α [39, 49, 59] with an ordinal scale. Damessie et al. [26] showed that a good crowd-sourcing study would be around 0.500 Krippendorff's α. For the 2020 test collection, the agreement score is exceeding crowd-sourcing in many instances and even approaching agreement levels of human experts in the best case. Run-level agreement with Kendall's τ – using NDCG@10 to rank the systems – spans from between 0.79 to 0.83 on the 2020 data, and from 0.60 to 0.70 on the 2021 data, representing a relatively strong run-level correlation (for 2020) and less correlation in 2021 as expected.

Why is agreement lower in 2021 than in 2020? Previous work demonstrated that IR experts may agree with LLMs more than the original TREC assessors on high disagreement pairs [58]. In this study, we investigate this further by computing the inter-rater agreement between the human TREC assessors and each LLM assessor on a *per topic* basis. Figure 6 represents the five agreement scores (as we used five LLMs) per topic as a box plot, clearly showing much higher agreement scores across topics in 2020 than in 2021. Another interesting observation is the high variance in 2021 topics, meaning the LLMs tend to disagree with each other more in 2021 than 2020, probably due to more ambiguous information needs in 2021.

In Figure 5, we show how the confusion matrices of the topics receiving the lowest correlation score in each year compared to the the across-topic average confusion matrices reported in Figure 4. It is evident that the distribution of the labels is different from the average. LLMs clearly overrate relevance when human marked as non-relevant, and this is more pronounced in 2021. It is worth mentioning that we checked whether the topic type (known item, topical, etc) has any impact on the agreement scores and found that all topic types have nearly the same agreement pattern, ruling this out as a cause of disagreement.

We speculate that another possible reason for 2021 assessments receiving low agreement scores is more nuanced or vague topic descriptions. For example, the word *"not"* – used to calibrate the judge on aspects that are not relevant to the information need – appears 15 times in 2020 vs. 36 times in 2021 topic descriptions. We believe that this nuance may be difficult to capture by the LLM judges. To investigate further, we annotated 3 or 4 random pairs from 3 topics – the topics receiving the lowest correlation scores – per year, totaling 39 pairs.[8] The obvious finding was ambiguity between the topic descriptions and queries, and often insufficient context to be confident with judgment (for things like "must be" or "can not"). This justifies why LLM assessments were heavily different from human ones for some topics, since we used only the topic descriptions for LLM judgment, and not the queries. Although we, as humans, accessed the query and description when annotating, we still could not be confident about what should be relevant or not relevant. For example, while the query of topic 81 in 2021 was about "men's fashion brands", the topic description did not mention anything about *fashion* or *men* at all.

[8]We release this annotation with our reasoning in the shared codebase

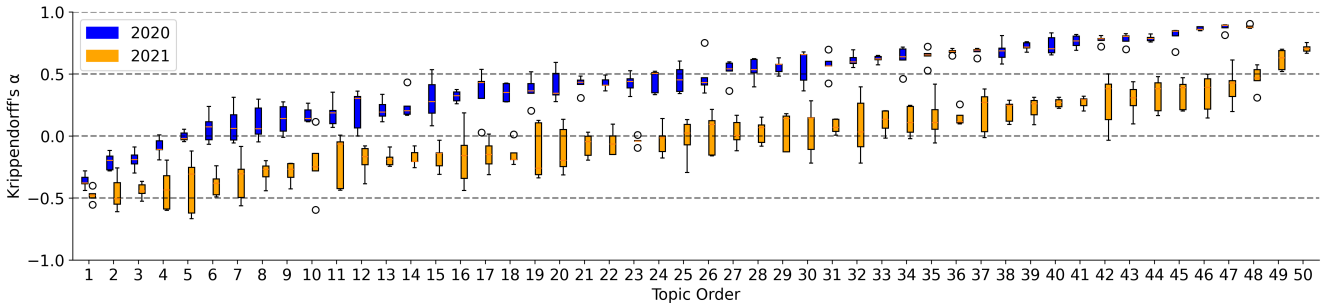


Figure 6: Inter-rater agreement as measured with Krippendorff's α between LLM and human assessments across *each topic* for 2020 and 2021 topics using the `Spotify` transcripts. Each box expresses the agreement of the LLM judge with the human assessor for each of the five LLMs. The boxes are sorted in ascending order based on the average agreement, and are grouped by year.

Table 2: Inter-rater agreement as measured with Krippendorff's α between LLM and human assessments across each pair of judgments on all ASR transcriptions for 2020 topics.

ASR	GPT-4o	Mistral	Qwen	Llama3	Gemma2
Spotify	0.589	0.596	0.495	0.545	0.529
WhisperX-large	0.587	0.586	0.503	0.547	0.535
WhisperX-base	0.590	0.596	0.502	0.553	0.532
Silero-large	-	0.583	0.509	0.550	0.534
Silero-small	-	0.576	0.504	0.519	0.533

Table 3: Inter-rater agreement as measured with Krippendorff's α between LLM and human assessments across each pair of judgments on all ASR transcriptions for 2021 topics.

ASR	GPT-4o	Mistral	Qwen	Llama3	Gemma2
Spotify	0.144	0.148	-0.043	0.144	-0.057
WhisperX-large	0.118	0.142	-0.035	0.132	-0.036
WhisperX-base	0.137	0.147	-0.029	0.140	-0.032
Silero-large	-	0.168	0.015	0.126	-0.029
Silero-small	-	0.180	0.027	0.136	-0.005

4.3 Impact of ASR Variation on Judgment

Continuing this same line of investigation, this subsection explores how the underlying text representation of the documents influences LLM judgment. Following the same experimental setup as the previous section, we re-judge the TREC Podcast 2020 and 2021 relevance pairs using the four alternative transcripts for the same audio segment – that is, the same query/segment pair is fed into each LLM, but the document segment is varied according to each ASR model. We demonstrate the per-pair agreement with humans across all transcripts on the 2020 and 2021 data in Tables 2 and 3, respectively.

Surprisingly, we see only a subtle difference in agreement scores across all used transcriptions (with the same LLM judge), even for low-quality ASR like `Silero`-small. This observation suggests that LLMs are robust in judging relevance, even in the presence of errors or noise in the transcripts. Furthermore, we note that correlation varies substantially more *across LLMs* (with a fixed ASR transcript), and that the highest agreement scores were achieved by `Mistral`, followed closely by `GPT-4o`. This variance can be attributed to underlying differences in the LLMs, such as the number of parameters, quantization levels, and training data. We also found that the distribution of labels was consistent regardless of the ASR used to make the judgments, confirming the above results.

4.4 Impact of Transcript Modality on Judgment

Our previous experiment assumed the LLM judges would operate directly on the ASR text representation of each document for judgment. However, given that the raw audio is available, it is possible to provide the audio directly to the LLM judge. Furthermore, we also experiment with the inclusion of the podcast episode metadata as further evidence and additional context for the LLM judge [19].

Since the previously used LLMs do not natively support audio, we used the multimodal `Gemini-2.5-Pro` [24] (the top-performing audio LLM on Stanford HELM leaderboard[9] [52]) as a SOTA closed model, and the audio-based `Audio-Flamingo-3-hf` [37] and `Kimi-Audio-7B-Instruct` [27] as SOTA open models. Following the previous experimental setup, we then re-judged the TREC Podcast assessments using a variety of settings.[10]

Table 4 shows the results on the 2020 data.[11] Firstly, it is evident that the open-source LLMs have very weak levels of agreement against the TREC assessments. This is likely due to the window-based context applied by both models, which may reduce the quality for longer segments such as those used in our experiments. We find that these models are more inclined to over-rate relevance as compared to the other text-based LLMs, with a much higher proportion of '1' labels (and only around 30% of the entire set of pairs judged as non-relevant). Secondly, we observe that the `Gemini` model is quite consistent irrespective of the modality applied, with the strongest agreement to the human assessors in the text+metadata setting.

Next, we explored the *consistency* of the `Gemini` model under different modalities by measuring Krippendorff's α between all pairs of modalities. Interestingly, the consistency levels were all in the

[9] https://crfm.stanford.edu/helm/audio/latest/#/leaderboard
[10] Our prompt was modified slightly to reflect these changes in modality. All prompts are available in our codebase.
[11] Given the aforementioned issues with the stability of the 2021 data, we omit it for brevity. All data is available in our codebase.

Table 4: 2020 judgments correlation between human and text and audio LLMs. The prompts and metadata are always provided as text. The text modality uses the `Spotify` transcript. `GPT-4o` is used as a representative from the prior experiment in Table 2.

Model	Modality	Metadata	Krippendorff's α
`GPT-4o`	Text		0.589
`Flamingo-3`	Audio		0.113
`Kimi-Audio-7B`	Audio		0.239
`Gemini-2.5-pro`	Audio		0.585
`Gemini-2.5-pro`	Audio	✓	0.591
`Gemini-2.5-pro`	Text		0.588
`Gemini-2.5-pro`	Text	✓	0.605

high reliability range; the lowest value observed between the *Audio+Metadata* and *Text* modalities ($\alpha = 0.868$); and the highest value ($\alpha = 0.909$) was observed between the *Text* and *Text+Metadata* setting.

Detailed analysis of the extreme disagreements between the TREC assessors and the LLM judges revealed some interesting failure cases. For example, one segment for the query "*Nefertiti*" had a human assessment of 0 (not relevant), an audio-based label of 1 (partially relevant), and a text-based assessment of 3 (highly relevant). The audio-based assessment was justified by the fact that the "first minute of this audio clip is advertisements" which demonstrates utility in the audio modality to capture temporality – the text-based assessment does not require listening through a minute worth of ads. Interestingly, the justification provided by the same LLM under the text modality noted "a brief advertisement" was observed before the segment content.

The sum of these findings is that LLM judgment is not greatly sensitive to the modality of the document, and that the judgments made by `Gemini` remain broadly consistent across modalities. However, different presentation modalities can give rise to different interpretations of relevance. It would be useful for future work to examine the alignment between the true user experience and the way the relevance assessment is conducted.

4.5 Impact of ASR Variation on Systems

In the previous subsection, we demonstrated that the transcript variation has no impact on LLM-generated judgments. Next, we are interested in exploring whether this variation influences the retrieval effectiveness, and how judging the holes in the original collection changes the IR system performance.

Towards answering this question, we evaluate our pool systems using the original TREC Podcast labels and the `Mistral` labels. We report the results using NDCG@10 [42] and RBP with $\phi = 0.9$ [62]– see Table 5. The first interesting observation is that all systems receive a significant increase in RBP score regardless of the transcript used. This is mainly attributed to judging previously unjudged documents, and demonstrates the importance of reporting residuals when unjudged documents are present. Another observation is the remarkable decline of NDCG@10 scores across all transcripts. This is primarily attributed to the fact that when LLM judged the holes, it marked many segments with varied graded relevance, increasing the ideal DCG (denominator of NDCG) and causing a decline in the scores.

Comparing across the `WhisperX-large` and `Silero-small` transcriptions, we observe a remarkable decline in overall system performance for `Silero-small`. For example, the RBP score for the system ranked 2 with human on `WhisperX-large` is 0.538, and it goes down to 0.426 on `Silero-small`. In turn, these changes propagate to system ordering as shown by the rank delta columns, suggesting that the ASR model can have a significant impact on *system orderings*, suggesting careful consideration of the chosen ASR transcripts when designing retrieval pipelines.

4.6 System Effectiveness under Variation

Our final experiment seeks to understand why systems vary under both query and transcript variations. To achieve this, we ran all query variants across all systems using the five available transcripts – resulting in a total of $5745 \times 17 \times 5 = 488{,}325$ NDCG@10 observations – and then examined the distribution of per-topic performance.

Figures 7a and 7b present two interesting failure cases. First, in Figure 7a, it is evident that query variance has a huge impact on systems' performance. While the query provided by TREC (67: *"pros and cons of ubi"*) is difficult for many systems in the pool with variability across transcripts, the query variant "67_2071" (*"universal basic income viewpoints"*) is much easier irrespective of the transcript applied. Variance in both query variations and transcripts is observed across the board.

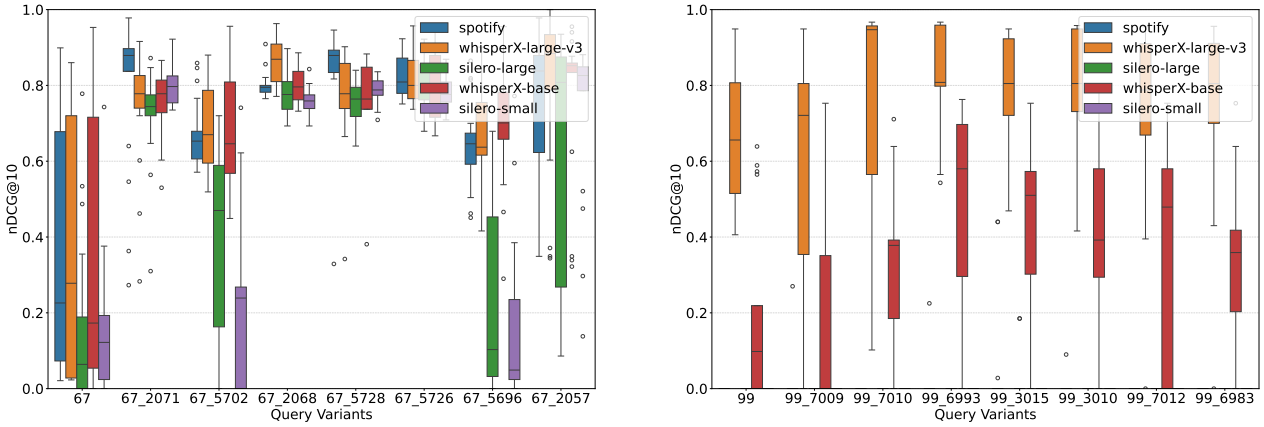
Figure 7b captures a more interesting failure case for query *99 "Samin Nosrat."* Surprisingly, nearly all systems received a zero score on the `Spotify` and `Silero` transcripts, independent of the query variant; scores are highly varied on the remaining transcripts, with `WhisperX-large` resulting in higher scores than `WhisperX-base`. Investigating further, we found that the entity *Samin Nosrat* was never transcribed correctly in the `Spotify` or `Silero` transcripts, so systems could not retrieve relevant segments. Since this topic and transcript were used in the official TREC 2021 Podcast track, we examined the overall system performance of the official runs. As expected, all systems failed to retrieve any relevant segment except for a few runs that used the metadata in their pipelines. For example, Hofstätter et al. [40] (the top team) prepended the episode title to all transcribed segments before indexing. Since the metadata was written by humans and the query words appear in the episode name, metadata can solve the transcription issue for this topic.

5 Conclusions

This study challenges the prevailing *single-transcript* paradigm in spoken document retrieval, using the TREC podcast track as a starting point. By conducting a large-scale sensitivity analysis across five ASR models, 6,000 query variations, and 1.8 million relevance judgments, we have demonstrated that current evaluation standards are brittle in the face of transcription noise. Our analysis reveals the following critical implications for the community: (a) We illustrate that the quality of the transcription model has a significant impact on the downstream systems, potentially leading to biased test collections if not handled with care; (b) We show that variations in queries, systems, and transcripts have considerable impacts on pools' diversity, suggesting a reusable audio collections should consider these variations to create a diverse and deep pool; (c) We demonstrate that LLM-based assessment remains robust to noise in transcriptions, even

Table 5: Evaluation of the pool (with NDCG@10 and RBP with $\phi = 0.9$) created using human and `Mistral` assessments on representative transcripts for the 2020 title queries. Systems are ordered on the human assessments with NDCG@10. Residual values for new transcripts are omitted because they are always approximately zero – the pool has no unjudged documents. The Kendall's τ between the human and `WhisperX-large` orderings, and between the human and `Silero-small` orderings, is 0.750, and 0.662, respectively.

System	Human		`WhisperX-large`			`Silero-small`		
	NDCG@10	RBP	Δ Rank	NDCG@10	RBP	Δ Rank	NDCG@10	RBP
(1) BM25+MonoT5+RankZephyr	0.471	0.309 + 0.199	↓2	0.309	0.524	↓1	0.272	0.457
(2) ColBERT-v2+MonoT5+RankZephyr	0.458	0.299 + 0.280	–	0.315	0.538	↓1	0.257	0.426
(3) SPLADE-v3+MonoT5+RankZephyr	0.451	0.289 + 0.304	↑2	0.316	0.541	↑2	0.290	0.471
(4) DPH+MonoT5	0.440	0.293 + 0.162	–	0.291	0.501	↓1	0.252	0.430
(5) BM25+RM3+MonoT5	0.434	0.288 + 0.159	↓2	0.285	0.495	↓1	0.251	0.430
(6) BM25+MonoT5	0.430	0.288 + 0.160	↓2	0.285	0.495	↓1	0.250	0.428
(7) ColBERT-v2+MonoT5	0.410	0.274 + 0.246	↑2	0.287	0.501	↓2	0.238	0.406
(8) ColBERT-v2	0.400	0.268 + 0.234	↓1	0.283	0.486	↓3	0.225	0.368
(9) SPLADE-v3+MonoT5	0.398	0.271 + 0.252	↑4	0.288	0.507	↑5	0.260	0.438
(10) SPLADE-v3	0.396	0.265 + 0.271	–	0.279	0.485	↓1	0.233	0.388
(11) SPLADE++	0.384	0.251 + 0.298	↓1	0.259	0.455	↓1	0.212	0.351
(12) DPH	0.341	0.240 + 0.114	↓2	0.241	0.421	↓1	0.203	0.343
(13) PL2	0.340	0.245 + 0.115	↑1	0.242	0.427	↓3	0.188	0.331
(14) BM25	0.340	0.244 + 0.095	↓1	0.237	0.420	↓3	0.187	0.331
(15) BM25+RM3	0.330	0.238 + 0.138	↑1	0.239	0.428	↓2	0.201	0.351
(16) BGE-large-en-v1.5	0.320	0.221 + 0.498	↑5	0.270	0.478	↑8	0.240	0.408
(17) BGE-m3	0.258	0.178 + 0.576	–	0.230	0.408	↑3	0.203	0.343



(a) Query 67: *pros and cons of ubi*

(b) Query 99: *Samin Nosrat*

Figure 7: NDCG@10 scores across 8 random variants of two queries and five transcripts. The official title query is the leftmost in each plot. In the left plot, we observe large effects due to the query variant used, and that this effect is not consistent across transcription models. In the right plot, we observe that transcription errors can result in no relevant documents being retrieved regardless of the query variant used.

when the quality is markedly reduced, and that the inclusion of metadata can improve the quality of both ranking and judgment, especially for named-entity queries and when metadata quality is reliable; (d) We provide empirical evidence that multimodal judgment – where the LLM evaluates the raw audio rather than the transcript itself – can successfully resolve some edge cases where the ASR model fails, though the judgment remains mostly consistent between modalities.

The resources generated by our work can be used for a number of IR tasks, including ad-hoc podcast retrieval, efficiency studies, exploring and validating document judgment using large language models, and experimentation on retrieval consistency, where both query and document content can dramatically vary for the same underlying information. Ultimately, our work demonstrates that as podcast search matures, our evaluation instruments must evolve from simple text-matching tasks to robust, multimodal listening tasks.

Acknowledgments. We thank the reviewers for their recommendations to improve the work. This work was partially supported by a Google Research Scholar grant. The authors acknowledge the peoples of the Woi Wurrung and Boon Wurrung language groups of the eastern Kulin Nation on whose unceded lands ACM SIGIR 2026 was hosted, and the Yuggera and Turrbal peoples of Meanjin where this research was carried out. We pay our respects to their Elders past, present, and emerging.

References

[1] 2025. Podcast Statistics 2025. https://podcastatistics.com. (Accessed Jan 2026).

[2] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can We Use Large Language Models to Fill Relevance Judgment Holes?. In *Proc. EMTCIR*.

[3] Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proc. TREC*.

[4] Tomoyosi Akiba, Hiromitsu Hishizaki, Hiroaki Nanjo, and Gareth J. F. Jones. 2014. Overview of the NTCIR-11 SpokenQuery&Doc Task. In *Proc. NTCIR*. 350–364.

[5] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo, and Yoichi Yamashita. 2013. Overview of the NTCIR-10 SpokenDoc-2 Task. In *Proc. NTCIR*. 573–587.

[6] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara, and Tomoko Matsui. 2011. Overview of the IR for Spoken Documents Task in NTCIR-9 Workshop. In *Proc. NTCIR*. 223–235.

[7] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, and Gareth J. F. Jones. 2016. Overview of the NTCIR-12 SpokenQuery&Doc-2 Task. In *Proc. NTCIR*.

[8] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. 2024. LLMs can be Fooled into Labelling a Document as Relevant: best café near me; this paper is perfectly relevant. In *Proc. SIGIR-AP*. 32–41.

[9] Abigail Alexander, Matthijs Mars, Josh C. Tingey, Haoyue Yu, Chris Backhouse, Sravana Reddy, and Jussi Karlgren. 2021. Audio Features, Precomputed for Podcast Retrieval and Information Access Experiments. In *Proc. CLEF*. 3–14.

[10] Gianni Amati. 2003. *Probabilistic Models for Information Retrieval based on Divergence from Randomness.* Ph. D. Dissertation. Department of Computing Science, University of Glasgow.

[11] Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Sys.* 20, 4 (2002), 357–389.

[12] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User Variability and IR System Evaluation. In *Proc. SIGIR*. 625–634.

[13] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proc. SIGIR*. 725–728.

[14] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proc. SIGIR*. 395–404.

[15] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proc. Interspeech*.

[16] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv:1611.09268v3* (2018).

[17] Rodger Benham, Joel Mackenzie, Alistair Moffat, and J. Shane Culpepper. 2019. Boosting Search Performance Using Query Variations. *ACM Trans. Inf. Sys.* 37, 4, Article 41 (2019), 25 pages.

[18] Jana Besser, Martha Larson, and Katja Hofmann. 2010. Podcast search: User goals and retrieval technologies. *Online information review* 34, 3 (2010), 395–419.

[19] Ben Carterette, Rosie Jones, Gareth F. Jones, Maria Eskevich, Sravana Reddy, Ann Clifton, Yongze Yu, Jussi Karlgren, and Ian Soboroff. 2021. Podcast Metadata and Content: Episode Relevance and Attractiveness in Ad Hoc Search. In *Proc. SIGIR*. 2247–2251.

[20] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Proc. ACL (Findings)*. 2318–2335.

[21] Charles LA Clarke and Laura Dietz. 2024. LLM-based relevance assessment still can't replace human relevance assessment. *arXiv preprint arXiv:2412.17156* (2024).

[22] Cyril Cleverdon. 1967. The Cranfield Tests on Index Language Devices. *Proc. Aslib* 19, 6 (1967), 173–194.

[23] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proc. COLING*. 5903–5917.

[24] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[25] J. Shane Culpepper. 2018. Single Query Optimisation is the Root of All Evil. In *Proc. DESIRES*.

[26] Tadele T Damessie, Thao P Nghiem, Falk Scholer, and J Shane Culpepper. 2017. Gauging the quality of relevance assessments using inter-rater agreement. In *Proc. SIGIR*. 1089–1092.

[27] Kimi Team: Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. Kimi-Audio Technical Report. *arXiv preprint arXiv:2504.18425* (2025).

[28] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. 2024. Building Guardrails for Large Language Models. *arXiv preprint arXiv:2402.01822* (2024).

[29] Maria Eskevich, Robin Aly, Roeland Ordelman, David N Racca, Shu Chen, and Gareth J. F. Jones. 2015. SAVA at MediaEval 2015: Search and anchoring in video archives. In *Proc. MediaEval*.

[30] Maria Eskevich, Gareth J. F. Jones, Shu Chen, Robin Aly, Roeland Ordelman, and Martha Larson. 2012. Search and hyperlinking task at MediaEval 2012. In *Proc. MediaEval*.

[31] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, et al. 2023. Perspectives on large language models for relevance judgment. In *Proc. ICTIR*. 39–50.

[32] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2024. Who Determines What Is Relevant? Humans or AI? Why Not Both? *Comm. ACM* 67, 4 (2024), 31–34.

[33] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proc. SIGIR*. 2353–2359.

[34] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proc. SIGIR*. 2288–2292.

[35] Jorge Gabín, Javier Parapar, and Craig Macdonald. 2024. Beyond Questions: Leveraging ColBERT for Keyphrase Search. *Info. Proc. & Man.* 63, 2, Part B (2024).

[36] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The TREC spoken document retrieval track: A success story. In *Content-Based Multimedia Information Access-Volume 1*. 1–20.

[37] Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. In *Proc. NeurIPS*.

[38] Helia Hashemi, Aasish Pappu, Mi Tian, Praveen Chandar, Mounia Lalmas, and Benjamin Carterette. 2021. Neural Instant Search for Music and Podcast. In *Proc. KDD*. 2984–2992.

[39] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89.

[40] Sebastian Hofstätter, Mete Sertkan, and Allan Hanbury. 2021. TU Wien at TREC DL and Podcast 2021: Simple Compression for Dense Retrieval. In *Proceedings of Text REtrieval Conference (TREC)*.

[41] Yuta Imasaka and Hideo Joho. 2024. Effect of LLM's Personality Traits on Query Generation. In *Proc. SIGIR-AP*. 249–258.

[42] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.* 20, 4 (2002), 422–446.

[43] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[44] Gareth J. F. Jones. 2019. About sound and vision: CLEF beyond text retrieval tasks. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF* (2019), 307–329.

[45] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu TREC. 2020. TREC 2020 Podcasts Track Overview. In *Proc. TREC*.

[46] Rosie Jones, Hamed Zamani, Markus Schedl, Ching-Wei Chen, Sravana Reddy, Ann Clifton, Jussi Karlgren, Helia Hashemi, Aasish Pappu, Zahra Nazari, Longqi Yang, Oguz Semerci, Hugues Bouchard, and Ben Carterette. 2021. Current Challenges and Future Directions in Podcast Information Access. In *Proc. SIGIR*. 1554–1565.

[47] Jussi Karlgren, Rosie Jones, Ben Carterette, Ann Clifton, Edgar Tanaka, Maria Eskevich, G. Jones, and Sravana Reddy. 2021. TREC 2021 Podcasts Track Overview. In *Proc. TREC*.

[48] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proc. SIGIR*. 39–48.

[49] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition).* Sage Publications.

[50] Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth Jones. 2011. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *Proc. MediaEval*.

[51] Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. SPLADE-v3: New baselines for SPLADE. *arXiv preprint arXiv:2403.06789* (2024).

[52] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove,

Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Trans. Mach. Learn. Res.* (2023).

[53] Sean MacAvaney and Luca Soldaini. 2023. One-Shot Labeling for Automatic Relevance Estimation. In *Proc. SIGIR*. 2230–2235.

[54] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval using PyTerrier. In *Proc. ICTIR*.

[55] Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. PyTerrier: Declarative Experimentation in Python from BM25 to Dense Retrieval. In *Proc. CIKM*. 4526–4533.

[56] Joel Mackenzie and Alistair Moffat. 2021. Modality Effects When Simulating User Querying Tasks. In *Proc. ICTIR*. 197–201.

[57] Watheq Mansour, J Shane Culpepper, and Joel Mackenzie. 2025. Examining the Impact of Transcript Variation on Podcast Search and Re-ranking. In *Proc. ECIR*. 118–127.

[58] Watheq Mansour, J. Shane Culpepper, Joel Mackenzie, and Andrew Yates. 2026. Revisiting Human-vs-LLM judgments using the TREC Podcast Track. In *Proc. ECIR*. 436–443.

[59] Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX* 12 (2024).

[60] Alistair Moffat. 2016. Judgment Pool Effects Caused by Query Variations. In *Proc. ADCS*. 65–68.

[61] Alistair Moffat, Falk Scholer, Paul Thomas, and Peter Bailey. 2015. Pooled Evaluation Over Query Variations: Users are as Diverse as Systems. In *Proc. CIKM*. 1759–1762.

[62] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.* 27, 1, Article 2 (2008), 27 pages. doi:10.1145/1416950.1416952

[63] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Proc. EMNLP (Findings)*. 708–718.

[64] Pavel Pecina, Petra Hoffmannova, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. 2008. Overview of the CLEF-2007 cross-language speech retrieval track. In *Proc. CLEF*. 674–686.

[65] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *Proc. ECIR*. 397–412.

[66] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. *arXiv preprint arXiv:2309.15088* (2023).

[67] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and Robust Zero-Shot Listwise Reranking is a Breeze! *arXiv preprint arXiv:2312.02724* (2023).

[68] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*. 28492–28518.

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[70] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Report on the 1st Workshop on Large Language Model for Evaluation in Information Retrieval (LLM4Eval 2024) at SIGIR 2024. *SIGIR Forum* (2024).

[71] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, and Bhaskar Mitra. 2024. JudgeBlender: Ensembling Judgments for Automatic Relevance Assessment. In *Proc. WWW*. 1268–1272.

[72] David Rau and Jaap Kamps. 2024. Query Generation Using Large Language Models. In *Proc. ECIR*. 226–239.

[73] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trnd. Inf. Retr.* 3, 4 (2009), 333–389.

[74] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC-3)*. 109–126.

[75] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. PLAID: An Efficient Engine for Late Interaction Retrieval. In *Proc. CIKM*. 1747–1756.

[76] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proc. NAACL*. 3715–3734.

[77] Sebastian Schmiedeke, Peng Xu, Isabelle Ferrané, Maria Eskevich, Christoph Kofler, Martha A Larson, Yannick Estève, Lori Lamel, Gareth J. F. Jones, and Thomas Sikora. 2013. Blip10000: A social video dataset containing spug content for tagging and retrieval. In *Proc. MMSys*. 96–101.

[78] Vinay Setty and Adam James Becker. 2025. Annotation Tool and Dataset for Fact-Checking Podcasts. *arXiv preprint arXiv:2502.01402* (2025).

[79] Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *J. Assoc. Inf. Sci. Technol.* 68, 9 (2017), 2101–2115.

[80] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? A study on the impact of format restrictions on performance of large language models. *arXiv preprint arXiv:2408.02442* (2024).

[81] Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*. 53–59.

[82] Gemma Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118* (2024).

[83] Llama 3 team. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024).

[84] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. In *Proc. SIGIR*. 1930–1940.

[85] Mi Tian, Claudia Hauff, and Praveen Chandar. 2022. On the challenges of podcast search at Spotify. In *Proc. CIKM*. 5098–5099.

[86] Manos Tsagkias, Martha Larson, and Maarten de Rijke. 2008. Term clouds as surrogates for user generated speech. In *Proc. SIGIR*. 773–774.

[87] Manos Tsagkias, Martha Larson, Wouter Weerkamp, and Maarten de Rijke. 2008. PodCred: A framework for analyzing podcast preference. In *Proc. ACM Workshop on Info. Cred. on the Web*. 67–74.

[88] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2024. A Large-Scale Study of Relevance Assessments with Large Language Models: An Initial Look. *arXiv preprint arXiv:2411.08275* (2024).

[89] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. 2024. UMBRELA: UMbrela is the (Open-Source Reproduction of the) Bing RELevance Assessor. *arXiv preprint arXiv:2406.06519* (2024).

[90] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general Chinese embedding. In *Proc. SIGIR*. 641–649.

[91] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report.

[92] Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models. In *Proc. SIGIR*. 38–47.