A Comparative Analysis of Linguistic and Retrieval Diversity in LLM-Generated Search Queries

Oleg Zendel
oleg.zendel@rmit.edu.au
RMIT University
Melbourne, Australia

Sara Fahad Dawood Al Lawati
s3919996@student.rmit.edu.au
RMIT University
Melbourne, Australia

Lida Rashidi
lida.rashidi@rmit.edu.au
RMIT University
Melbourne, Australia

Falk Scholer
falk.scholer@rmit.edu.au
RMIT University
Melbourne, Australia

Mark Sanderson
mark.sanderson@rmit.edu.au
RMIT University
Melbourne, Australia

Abstract

Large Language Models (LLMs) are increasingly used to generate search queries for various Information Retrieval (IR) tasks. However, it remains unclear how these machine-generated queries compare to human-written ones, particularly in terms of diversity and alignment with real user behavior. This paper presents an empirical comparison of LLM- and human-generated queries across multiple dimensions, including lexical diversity, linguistic variation, and retrieval effectiveness. We analyze queries produced by several LLMs and compare them with human queries from two datasets collected five years apart. Our findings show that while LLMs can generate diverse queries, their patterns differ from those observed in human behavior. LLM queries typically exhibit higher surface-level uniqueness but rely less on stopword use and word form variation. They also achieve lower retrieval effectiveness when judged against human queries, suggesting that LLM-generated queries may not always reflect real user intent. These differences highlight the limitations of current LLMs in replicating natural querying behavior. We discuss the implications of these findings for LLM-based query generation and user behavior simulation in IR. We conclude that while LLMs hold potential, they should be used with caution.

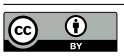
CCS Concepts

• **Information systems** → **Information retrieval query processing**; **Evaluation of retrieval results**.

Keywords

Generative AI, Large language models, Synthetic query generation, User behavior simulation, Information retrieval evaluation

ACM Reference Format:
Oleg Zendel, Sara Fahad Dawood Al Lawati, Lida Rashidi, Falk Scholer, and Mark Sanderson. 2025. A Comparative Analysis of Linguistic and Retrieval Diversity in LLM-Generated Search Queries. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761382



CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761382

1 Introduction

A defining characteristic of research in IR is its emphasis on users. Unlike other areas of computer science, where system behavior can often be formally specified, IR systems must account for user interactions, which are inherently variable and difficult to model. As a result, much of the research in this field is empirical. Prior work has shown that users are a major source of variability in retrieval effectiveness. Their behavior tends to be inconsistent, making evaluation challenging—especially in industrial contexts, where aligning offline and online evaluation remains an open problem [3, 11, 15, 17, 34].

To reduce the cost of collecting user assessments and constructing new datasets, several studies have proposed using Large Language Models (LLMs) to simulate user behavior. These studies have shown that data and annotations generated by LLMs often align well with those collected from human users, suggesting that LLMs could serve as substitutes in some IR tasks [4, 39, 55].

To examine this hypothesis, we used the UQV100 collection, which contains a diverse set of query variations generated by crowd workers for information needs originally developed for the TREC Web Track [10]. We also used a follow-up study in which crowd workers generated additional queries for the same set of topics. While collecting real-world queries provides insight into how users express their information needs, it is difficult to control for variation, as users often have diverse or ambiguous intents [6]. Since it's challenging to isolate queries that correspond to the same underlying need, we follow established methodologies that simulate a specific information need and ask users to formulate a query accordingly [10]. Collecting queries from different searchers for the same information need is how researchers have been studying query variation [3]. Query variation remains an important factor to research in academia and industry, as it has been shown to significantly impact retrieval system performance [17, 42].

In total, we evaluated seven widely used LLMs, including GPT-4o mini, Llama 3.3, and Claude 3.5, using the same information needs as the crowd studies. We prompted each LLM with multiple prompt variants to generate query variations. We then compared the generated queries to those written by crowd workers based on metrics such as the ratio of unique queries and syntactic structure. Our results show that the queries generated by the LLMs actually have a higher uniqueness ratio than those produced by human workers. Furthermore, the LLM queries exhibit a higher degree of

syntactic variability, with a greater number of unique syntactic structures. This suggests that LLMs can produce a wider range of query formulations than human workers, who tend to rely on a limited set of syntactic structures. However, it also indicates that the LLM queries are substantially different from those generated by humans, which may limit their applicability in IR research, if the goal is to simulate human behavior.

While LLMs show potential for generating queries, our findings suggest that they are not yet a suitable replacement for human users in IR research. The LLM-generated queries differ in variability and coverage, leading to results that diverge from human-generated query behavior. This suggests that although LLMs can be valuable tools in IR experimentation, they should be used with caution and not as a full substitute for human input.

2 Related Work

LLMs are increasingly used for tasks that traditionally required human annotation [19, 33, 53, 56]. A key development in this area is *InstructGPT* [37], which introduced instruction-based fine-tuning. This approach first generates labeled examples using human-written prompts, followed by supervised fine-tuning. Followed by reinforcement learning from human preference rankings, the result is models that follow instructions and generate coherent responses.

These models are now widely applied in IR tasks [22, 46] and broader machine learning applications [30]. Empirical studies show that LLMs often outperform crowdsourced human labeling across multiple evaluation tasks [22]. MacAvaney and Soldaini [33] demonstrated that LLMs can be useful to fill gaps in relevance judgments to provide reliable system rankings, a finding supported by subsequent studies [1, 47]. Microsoft has also reported using OpenAI's GPT models for relevance assessment in the Bing search engine [44]. Building on this, Upadhyay et al. [49] introduced UMBRELA, an open source toolkit that uses proprietary OpenAI models to label unjudged documents, replicating the results of Thomas et al. [44]. More recently, Upadhyay et al. [48] suggested that UMBRELA could serve as a viable replacement for human assessors. Extending this line of work, Alaofi et al. [4] demonstrated that LLMs can automatically generate queries and query variants from an information need description. Rahmani et al. [39] later proposed using LLMs to generate fully synthetic test collections, replacing human users in both query generation and relevance judgment. Along similar lines, the LiveRAG competition emerged, where questions, answers, and initial evaluations were all conducted by LLMs [12, 13, 41].

While recent advancements in LLMs have created new opportunities, their rapid adoption warrants caution. A recent study by Alaofi et al. [5] showed that LLMs can be manipulated through query stuffing, highlighting potential vulnerabilities. Despite their potential, further research is needed to understand LLM limitations.

3 Query Generation Methods

We categorized our query generation methods into two broad types: *context-based* and *query-based*. The context-based methods condition query generation on user and topic attributes embedded in the prompts. In contrast, the query-based methods focus on the structure and quantity of the output queries, giving LLMs more autonomy in generating queries based on predefined instructions and examples.

3.1 Generating Queries

Our goal was to simulate the human-generated queries from previous crowd studies [10, 54]. We developed five query generation methods. The first two methods attempted to replicate the original data collection setup used in the UQV100 dataset. In the original UQV100 collection process, each topic was described using a short backstory intended to clearly convey the information need. These backstories were written in a uniform style and presented to crowd workers in random order. Workers were asked to write one query per backstory.

To replicate this, we first prompted the LLMs with individual backstories grouped into shuffled batches of six, repeating this process 100 times with the same instructions prompt. In a second setup, we used a fixed topic order and varied the prompt phrasing across four versions, generating 50 queries per topic. Each prompt requested a single query per topic, mirroring the one query per worker setup in the original studies. Despite the structured design, the outputs showed limited lexical diversity and differed substantially from human-written queries, leading us to discard these two methods in the following analyses.

We next developed three methods more suited to LLM capabilities, which we summarize below.

Query-based methods:

- **`CW PV`:** This method prompted the LLM with a single backstory at a time, using four prompt variants (PV1-PV4). Each variant differed in how the required number of queries was specified: PV1 requested *exactly 100* queries, PV2 *expected 100*, PV3 *expected a value between 19 and 101*, and PV4 instructed the model to generate a *random* number within that range. No batching was used.
- **`CW 500`:** Each prompt included one backstory and asked for *500* queries. This number was chosen based on empirical observations. Requests for 100 queries often resulted in fewer than expected, while 1,000 queries were returned in multiple delayed batches. Generating 500 queries was both faster and more complete. To maximize the token budget for query generation, we omitted query and topic identifiers from the prompts.

Context-based method:

- **`VC`:** Inspired by Filice et al. [21], this method varied across four different user search skill levels: expert, intermediate, beginner, and none. It also considered three levels of topic knowledge: high, medium, and low, as well as two different query styles: natural language and keyword. A total of 24 unique combinations were created for each instruction prompt, utilizing one backstory at a time, and instructing the LLM to generate between 2 to 5 queries.

We include prompts, code and results for all methods, including those that were discarded, in the GitHub repository.[1]

[1] https://github.com/rmit-ir/Query-Gen-LLM

4 Empirical Analysis

This section presents an empirical comparison between the LLM-generated queries and the human-generated queries. We begin by examining basic query characteristics, including query length. Next, we assess the degree of uniqueness in the generated queries and investigate potential sources of uniqueness, such as stemming, stopword removal, and word order. We then analyze the distribution of part-of-speech (POS) patterns to identify systematic differences in linguistic structure. Finally, we examine how these characteristics relate to variability in performance in the downstream IR task, highlighting the practical implications of query formulation on retrieval effectiveness.

4.1 Experimental Settings

The UQV100 collection is comprised of over 10,000 queries submitted by 263 crowd workers, with each worker contributing one query per topic. Following data cleaning, spelling correction, and normalization, the collection includes 5,721 unique queries. These queries are associated with the ClueWeb12 Category B document corpus.[2]

In a follow-up study, Zendel et al. [54] selected 12 topics from the original UQV100 dataset and asked crowd workers to submit queries for each. While their focus was on assessing worker performance, we use their collected queries as a comparison set – referred to as `Mturk CW`– to evaluate alongside the original UQV100 and the queries generated by LLMs. Notably, many of the submitted queries in this subsequent study were not present in the original UQV100 dataset, making them a valuable source of additional query variation. Each worker contributed one query per topic, as in the original UQV100 setup. For our study, we restrict our analyses to these 12 topics to ensure consistency and comparability across human- and machine-generated queries.

Indexing and retrieval are done using the Anserini toolkit [51]. The document collection is indexed with the Porter stemmer applied and stopwords removed, following common practice for ad-hoc retrieval. In all analyses, the original crowd-sourced queries for the 12 topics from UQV100 are referred to as `UQV100 CW`, and the queries collected in the follow-up study are referred to as `Mturk CW`.

The LLMs used for query generation in this study include Mixtral 8x7B [27], Mistral Large, Mistral 7B [26], Llama 3.3 70B, Llama 3.2 11B [45], GPT-4o mini [36], and Claude 3.5 Haiku [7]. The Mistral and Llama models are open-weight and can be run locally, supporting reproducibility. In contrast, GPT-4o mini and Claude 3.5 Haiku are proprietary models accessible only via their respective APIs. In this study, all models were accessed through the Amazon Bedrock API,[3] which provides a unified interface for interacting with multiple LLMs. The OpenAI GPT-4o mini model was accessed via the Azure API.[4] Models accessed via Bedrock were used with default configurations, without additional fine-tuning or parameter adjustments. For GPT-4o mini, 'max_tokens' was set to 7,000 and 'top_p' to 0.95 to increase diversity.

[2]See http://www.lemurproject.org/clueweb12.php/.
[3]See https://aws.amazon.com/bedrock/.
[4]See https://azure.microsoft.com/.

4.2 Query Uniqueness

Table 1 presents the total number of queries generated by each method, along with the number of queries gathered from the human-generated `UQV100 CW` and `Mturk CW` datasets. The table also reports the percentage of unique queries before and after applying three standard transformations: stemming, stopword removal, and bag-of-words (BoW).[5] These transformations help isolate the factors contributing to query uniqueness. Specifically, stemming reduces words to their base forms, stopword removal eliminates common function words, and BoW discards word order — mirroring the way many IR systems, such as BM25, process queries. The uniqueness percentage is calculated as the number of unique queries divided by the total number of queries, multiplied by 100. The *RBP(0.8)* and *Residual* columns represent retrieval effectiveness and are discussed in Section 4.5.

Analyzing the changes in uniqueness after these transformations helps determine whether differences arise from lexical choices or word order. Overall, the LLM-generated queries tend to exhibit a higher percentage of uniqueness compared to human-generated queries. Interestingly, the `UQV100 CW` and `Mturk CW` datasets – despite being collected independently, across different platforms and five years apart – display similar uniqueness patterns. Among the LLM configurations, none fully replicate the uniqueness characteristics observed in human queries. For example, the combination of Llama 3.2 11B and `CW 500` shows a uniqueness rate of 41.9% after stopword removal, which is close to that of the human queries. However, this similarity does not hold consistently across transformations. For instance, when comparing the uniqueness before and after stopword removal, we see that it has no impact on the Llama 3.2 11B + `CW 500` queries. In contrast, in the human datasets, both stemming and stopword removal reduce uniqueness, with the combination resulting in an 8-10% drop. This suggests that a substantial portion of uniqueness in human queries stems from variations in stopword usage and word forms. The only LLM configuration that shows a somewhat similar pattern is the Mistral Large + `CW PV` method, where uniqueness drops by 7% (from 67.6% to 60.5%) after stemming and stopword removal. However, it further drops to 46.4% after applying BoW. This indicates that word order accounts for a large part of the uniqueness in these queries — even more than in the human-generated queries.

We observe a substantial drop in uniqueness for the human-generated queries after each transformation. In contrast, the LLM-generated queries appear less affected by such linguistic elements, suggesting a different pattern of lexical diversity. Furthermore, no single method demonstrates consistent behavior with humans across all transformations. These results highlight that although LLMs can generate diverse queries, their uniqueness patterns differ from those observed in real user behavior.

4.3 Query Length

Figure 1 shows the distribution of query lengths across all generation methods and the human-generated queries, using box-and-whisker plots. The box represents the interquartile range (IQR), the line inside the box indicates the median, and whiskers extend to 1.5 times the IQR. Outliers beyond this range are plotted individually.

[5]Note that the BoW transformation is applied after stemming and stopword removal.

Table 1: Summary of query counts and uniqueness patterns across 12 topics. The table compares the total queries generated by seven LLMs and three generation techniques with human queries, showing the percentage of unique queries before and after applying stemming, stopword removal, and bag-of-words transformations. The columns RBP(0.8) represents retrieval effectiveness, and the Residual column represents the amount of uncertainty with the existing relevance judgments.

Model	Technique	Tot. Queries	Uniq. Queries	Stop	Stem	StopStem	BoW	RBP(0.8)	Residual
Mixtral 8x7B	CW 500	987	99.3%	99.2%	99%	98.8%	97.1%	0.15	0.41
	CW PV	3,196	73.4%	71.4%	71.7%	69.7%	64.2%	0.13	0.50
	VC	1,373	60.1%	57.6%	58%	56.6%	53.8%	0.21	0.25
Mistral Large	CW 500	6,783	14.8%	14.2%	14.6%	14%	8.52%	0.23	0.26
	CW PV	4,820	67.6%	63.9%	64.4%	60.5%	46.4%	0.21	0.26
	VC	1,438	75.4%	71%	71.9%	69.5%	63.7%	0.23	0.21
Mistral 7B	CW 500	721	100%	99.7%	100%	99.7%	99.4%	0.17	0.38
	CW PV	1,997	81.7%	79.9%	80.9%	79.2%	74.8%	0.18	0.37
	VC	1,439	62.2%	60.4%	60.7%	59.5%	57.1%	0.20	0.23
Llama 3.3 70B	CW 500	4,986	64%	64%	63.9%	63.8%	59.4%	0.12	0.52
	CW PV	11,724	39.7%	38.4%	39.1%	37.6%	33.3%	0.16	0.41
	VC	1,312	65.2%	64.4%	64.6%	63.7%	61.1%	0.21	0.29
Llama 3.2 11B	CW 500	6,034	41.9%	41.9%	41.8%	41.8%	37.2%	0.14	0.48
	CW PV	19,208	30.1%	29.5%	29.9%	29.2%	22%	0.14	0.47
	VC	1,440	58.8%	57.6%	57.6%	56.2%	52.4%	0.23	0.24
GPT-4o mini	CW 500	3,767	95.5%	92.9%	93.9%	91.7%	84.2%	0.16	0.42
	CW PV	1,362	93.3%	90%	91%	89.4%	87.4%	0.18	0.42
	VC	1,424	79.1%	76%	76.5%	74.9%	72.7%	0.23	0.21
Claude 3.5 Haiku	CW 500	1,677	71.2%	71.2%	70.9%	70.9%	69.9%	0.09	0.70
	CW PV	3,094	85.3%	84.7%	84.6%	84%	81.8%	0.09	0.71
	VC	1,383	87.2%	85.8%	86.9%	85.3%	83.4%	0.17	0.42
Human	UQV100 CW	1,305	44.8%	39.5%	39.4%	34.9%	29.7%	0.28	0.03
	Mturk CW	2,200	46.3%	41.6%	42.6%	38.2%	33.2%	0.27	0.14

The figure highlights that query lengths from the Mturk CW and UQV100 CW human datasets are more similar to each other than to those generated by LLMs. While LLM-generated queries vary across the different generation technique and models, they are generally more similar to each other than to human queries, indicating some consistency in generation patterns. LLM-generated queries tend to be longer overall, with most methods producing a median length above five words, compared to a median of four words for human queries. Among the LLM methods, CW PV (multi-prompt) generates the shortest queries, whereas VC produces the longest and most variable in length ones. In particular, the CW PV technique with Mixtral 8x7B closely matches the human query distribution, with 50% of queries between three and five words. Additionally, CW PV with LLaMA 3.3 70B and CW 500 with Mistral 7B and Mistral Large exhibit similar IQR range to the human queries, though their distributions are shifted toward longer lengths.

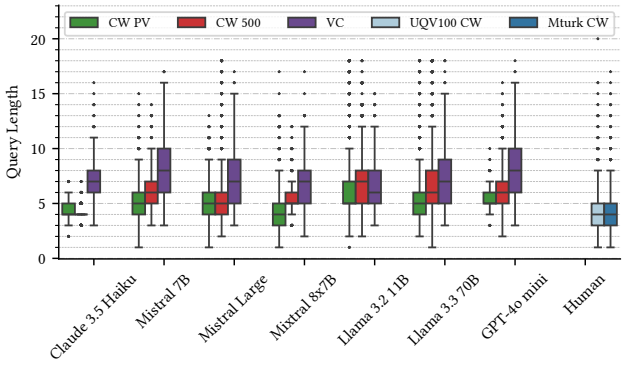


Figure 1: Boxplot showing the distribution of query lengths for each LLM and generation technique. The median is represented by the line inside the box, while the whiskers extend to 1.5 times the interquartile range (IQR).

4.4 Part-of-Speech Patterns

To examine the linguistic structure of the queries, we analyzed their part-of-speech (POS) tag sequences. The goal of this analysis is to assess whether different LLMs produce queries with consistent POS distributions, and whether these distributions differ from

those found in human-generated queries. We focus on the CW PV and VC generation techniques, as CW 500 yielded a low number of queries for Mistral 7B and Mixtral 8x7B (see Table 1). Using two distinct generation techniques also allowed alignment with the two human-generated datasets. Queries were tagged using spaCy's en_core_web_trf model.[6]

We computed the most frequent POS tag sequence for each LLM and for the human-generated queries. Table 2 presents these results. While we also computed POS patterns at the generation-technique and (crowd) dataset levels, only the model-level results are shown for brevity. Interestingly, while the most common POS tag sequence varied across CW PV and VC, it remained consistent across the two crowd-sourced datasets. This suggests more stable syntactic patterns in human-authored queries. The CW PV technique produced more uniform POS patterns, resembling human queries and suggesting higher predictability. In contrast, VC yielded greater POS diversity. While the primary difference across models is in sequence length, some models also exhibit distinct syntactic variations. For example, Claude 3.5 Haiku often starts with an adjective (ADJ), whereas GPT-4o mini includes an adposition (ADP) in the third position. The most frequent POS sequence among human queries (NOUN NOUN NOUN) occurred in 12.4% of cases, compared to just 7.3% for the most common LLM sequence (Claude 3.5 Haiku). This suggests that human queries are more formulaic – often keyword-based – while LLMs generate more syntactically varied queries.

To further explore this distinction, we trained a gradient boosting classifier (CatBoost [38]) to predict the query source based on POS tag sequences of fixed length 10. Queries shorter than 10 tokens were padded, and longer ones truncated. We used 2,610 queries per class, balanced across models and techniques, and split into 80% training and 20% testing sets with stratified sampling. The classifier achieved an average F1 score of 0.41. Specifically, human-generated queries were identified with the highest F1 score of 0.53 (precision = 0.46, recall = 0.62). As shown in Figure 2, the highest diagonal value in the confusion matrix corresponds to the human class, indicating a stronger signal for identifying human-written queries. Feature importance analysis (Figure 3) revealed that the POS tags in the third and fourth positions were most predictive. These findings indicate that while POS tag sequences alone offer limited classification accuracy, they carry useful signals for distinguishing between human- and LLM-generated queries. Incorporating additional features such as embeddings or linguistic properties, may further improve classification performance, which we leave for future work.

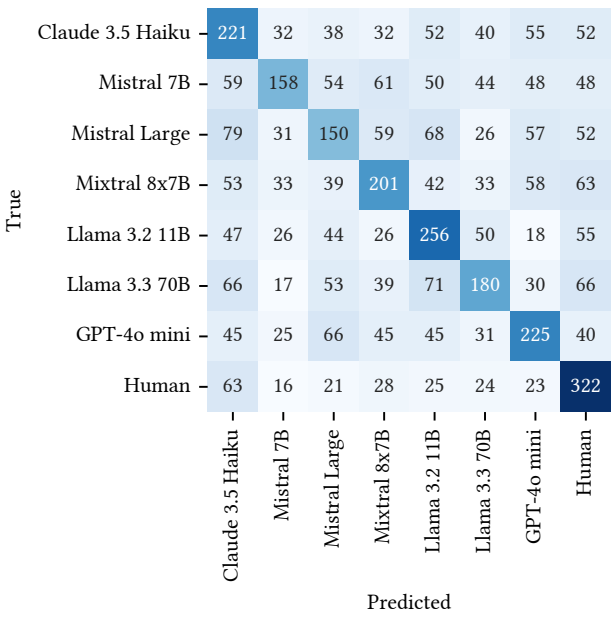


Figure 2: Confusion Matrix for classification based on POS sequence

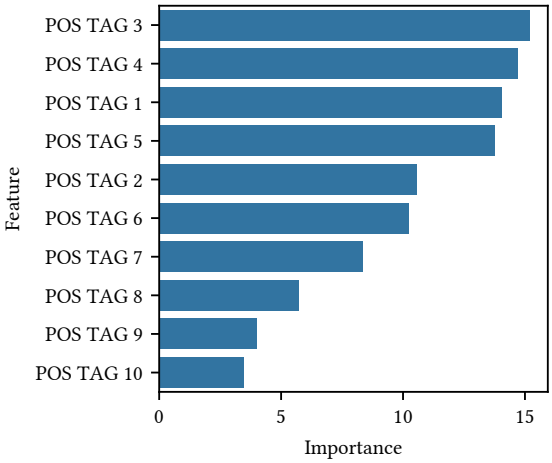


Figure 3: Feature importance for classification based on POS sequence

4.5 Retrieval Variability

So far, we have shown that there are substantial differences between human-generated and LLM-generated queries. However, differences in linguistic patterns alone are not sufficient grounds to determine the suitability of using AI-generated queries in IR. In many scenarios, the goal of generating synthetic data is to partially replace human users in the loop. Therefore, in the following analysis, we examine the impact of using LLM-generated queries on the downstream task of retrieval. We perform document retrieval using all generated queries. To reduce variability caused by superficial linguistic differences, we apply Porter stemming and remove stopwords using Anserini's default list. Retrieval is conducted using BM25 with default parameters ($k_1 = 0.9$, $b = 0.4$) on the ClueWeb12 Category B corpus. For each query, we retrieve the top 10 documents, which is a common cutoff in user-oriented IR evaluation.

To evaluate retrieval effectiveness, we compute Rank-Biased Precision (RBP) scores using a persistence parameter of 0.8 [35], along with their residuals. Scores are calculated per query, averaged across queries for each topic, and then macro-averaged over all

[6] https://spacy.io

Table 2: Most common POS tag sequences for each LLM and the human-generated queries. The table shows the top most frequent POS tag sequence, along with their respective frequency in the data and an example query.

Model	POS	Example Query	Frequency
Claude 3.5 Haiku	[ADJ, NOUN, NOUN, NOUN]	latest schizophrenia treatment options	0.073
GPT-4o mini	[NOUN, NOUN, ADP, NOUN, NOUN]	coping strategies for schizophrenia patients	0.021
Llama 3.2 11B	[NOUN, NOUN, NOUN, ADP, NOUN, ADP, ADJ, NOUN]	schizophrenia treatment options for adults with severe ocd	0.039
Llama 3.3 70B	[NOUN, NOUN, NOUN, NOUN]	schizophrenia medication side effects	0.060
Mistral 7B	[NOUN, NOUN, NOUN, NOUN]	schizophrenia medication side effects	0.038
Mistral Large	[NOUN, NOUN, NOUN, NOUN]	schizophrenia treatment drugs list	0.047
Mixtral 8x7B	[NOUN]	schizophrenia	0.052
Human	[NOUN, NOUN, NOUN]	drug treatment schizophrenia	0.124

topics. We apply the C/W/L framework [8] for score computation. The results are presented in Table 1. To be consistent with the Bailey et al. [10], the authors of UQV100, we use the relevance judgments from the UQV100 collection, which contains 4,865 topic-document pairs across the 12 selected topics.

The `UQV100 CW` query set achieves the highest effectiveness and the lowest residual. As defined by Moffat and Zobel [35], the RBP residual quantifies uncertainty due to unjudged documents. It is the gap between the observed score (assuming unjudged documents are non-relevant) and the upper bound (assuming they are all relevant). Note that the effectiveness scores are strongly negatively correlated with the residuals (Pearson's $r = -0.98$), indicating that lower effectiveness is associated with greater uncertainty. The overall mean residual is 0.362, which exceeds any of the RBP scores. This suggests that, given the current judgments in the UQV100 collection, differences in evaluation may largely reflect inconsistencies in judgment coverage across query sets rather than true differences in retrieval quality. Furthermore, the high residuals indicate that more commonly used metrics in IR, such as NDCG, which treat unjudged documents as non-relevant, may be even more misleading. This observation aligns with prior work showing that topics and queries are the primary sources of variance in IR evaluation [9, 17, 42].

Nonetheless, it is possible to quantify retrieval variability across query generation methods by examining the diversity of retrieved results through a resampling simulation. For each topic, we randomly sample $k = 15$ queries and count the number of unique documents retrieved across those queries. This process is repeated 1,000 times. The value $k = 15$ is chosen to ensure sufficient coverage, as the minimum number of queries per topic across all sets is 21. There are 54,264 possible combinations to choose 15 queries from 21, enabling a robust estimate of variability across all methods and topics based on a sample of 1,000 combinations.

For each query, we retrieve the top 10 documents, so the theoretical maximum number of unique documents retrieved per sample is 150. However, this number is rarely achieved due to overlap in retrieved documents across queries. When the number of unique documents approaches 150, it may indicate that the queries are diverse and cover a broad range of content. Alternatively, it may suggest that the queries are poorly aligned with the topic and retrieve unrelated documents. Conversely, a low number of unique documents suggests either high similarity among queries or a narrow topic scope, leading to substantial overlap in retrieved results. We mediate this by generating new relevance judgments, which we describe in Section 4.6.

To statistically analyze retrieval diversity across query sets, we apply Tukey's HSD test for multiple comparisons and report the mean number of unique documents retrieved along with 95% confidence intervals, as shown in Figure 4. The results show that LLM-generated queries generally retrieve more unique documents than human-generated queries. The only exceptions are the combinations of Mistral Large with `CW 500` and Mistral 7B with `VC`, which retrieve more unique documents than the `UQV100 CW` set but fewer than the `Mturk CW` set. Notably, Mistral Large with `CW 500` is the only configuration that does not differ significantly from any *Human* set. In Figure 4, the methods are ordered by the mean number of unique documents retrieved. However, the lack of a consistent ranking across models and generation techniques indicates that retrieval diversity is strongly influenced by the choice of LLM, the prompt design, and their interaction. No single LLM or generation strategy performs consistently across all topics, and none can be considered a universally reliable substitute for human-generated queries.

While the `Mturk CW` queries retrieve significantly more unique documents than the `UQV100 CW` set, the difference between the two human-generated sets is smaller than their differences relative to most LLM-generated sets. This suggests that the `Mturk CW` queries, despite being collected independently, exhibit retrieval behavior more similar to `UQV100 CW` than to the LLM-generated queries, even though the two human-generated sets were created five years apart.

4.6 Retrieval Effectiveness

We have shown that the variability introduced by queries generated through different GenAI methods often differs from that of human-generated queries, with LLM-generated queries generally exhibiting greater variation. We now turn to evaluating the retrieval effectiveness of these queries. From the perspective of Query Performance Prediction (QPP), a query is considered difficult if IR systems struggle to retrieve relevant documents for it, and easy if standard systems can retrieve relevant results with high effectiveness [20, 52]. Accordingly, we examine whether LLM-generated queries tend to be easier or harder than those created by humans. Recent work by Rahmani et al. [40] suggests that retrieval systems often achieve

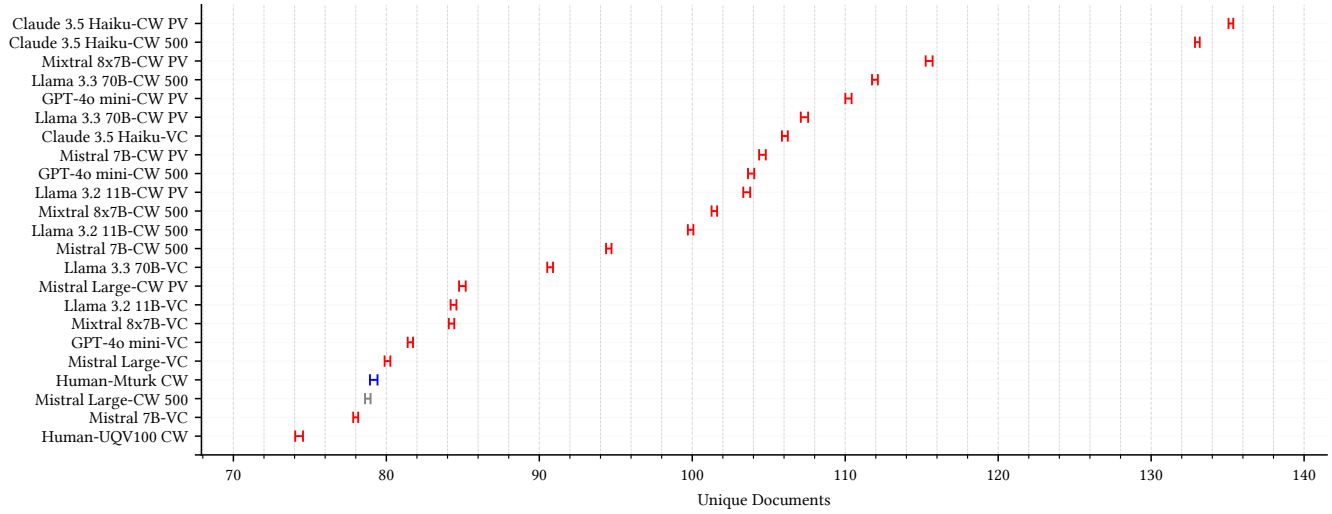


Figure 4: Mean number of unique documents retrieved per method with 95% confidence intervals from Tukey's HSD test. The `Mturk CW` queries (in blue) serve as the reference group. Intervals shown in gray indicate no statistically significant difference from the `Mturk CW` set, while red intervals denote significant differences. Statistical significance can be inferred when confidence intervals do not overlap. Methods are ordered by their mean number of unique documents retrieved.

higher NDCG@10 scores on LLM-generated queries compared to human-generated ones.

To support this evaluation, we perform automatic relevance assessments using OpenAI's GPT-4o model,[7] which is regarded as a State Of The Art (SOTA) model for this task [5, 43, 49]. To ensure consistency and reduce potential bias, we assess all topic-document pairs, including those previously judged in the UQV100 collection. We construct a judgment pool by selecting topic-document pairs that appear in the retrieval results of at least two different queries. Each query contributes a ranked list, resulting in a pool of 18,472 topic-document pairs (out of 57,966 total) for automatic relevance assessment. Although GPT-4o can process inputs up to 128,000 tokens, we truncate each document to 1,500 words to reduce computation time and minimize context degradation issues, such as "lost in the middle" [28, 32].[8] The annotation process took approximately 6 hours and incurred a total cost of US$150.87.

While prompt design plays an important role in the quality of generated judgments, Alaofi et al. [5] found that the `gpt-4o` model is generally robust to variations in prompt formulation. Therefore, we adopt a single prompt in this study rather than exploring alternative designs. Specifically, we use the original UQV100 relevance judgment instructions provided to crowd workers, with minor adjustments to ensure suitability for GPT-based annotation. These adjustments were made with the assistance of ChatGPT, which helped format the instructions in markdown and improve clarity and conciseness. The final prompt is available in our code repository.[9]

To evaluate agreement between LLM- and human-generated labels, we compute Cohen's κ [14] and Krippendorff's α (ordinal) [31]. Cohen's κ measures exact agreement on categorical labels, while Krippendorff's α accounts for the severity of ordinal disagreements. That is, the penalty for a disagreement between labels is proportional to the distance between them. Among the 2,788 topic-document pairs that have both UQV and LLM labels (out of 4,865 UQV-labeled items), we observe moderate agreement: $\kappa = 0.291$ and $\alpha = 0.477$. While prior work has reported higher agreement levels between human and LLM judgments, to the best of our knowledge, this is the first study to assess such agreement on the ClueWeb12-B collection. This dataset contains long and noisy general web documents, which pose greater challenges for LLM-based relevance assessment. Most existing studies focus on cleaner text passages and typically rely on binary relevance labels, which are generally easier for LLMs to assess [43]. Although the observed agreement is only moderate, it provides a sufficiently reliable signal for comparing the relative retrieval effectiveness (or difficulty) of different query sets.

We compute RBP scores for each query using the LLM-generated relevance judgments, following the same procedure described in Section 4.5. The results are shown in Figure 5. The overall mean residual value is 0.118, indicating that the LLM-generated judgments are more complete than the original UQV100 judgments, which had a mean residual of 0.362. This lower residual suggests that the LLM-based assessments offer a more complete signal about the relative retrieval effectiveness of the queries. Figure 5 presents the RBP(0.8) scores for each model, grouped by query generation technique. We observe consistent trends across models: the `CW PV` and `CW 500` techniques generally yield lower RBP scores than `VC`, implying that queries generated using `CW PV` and `CW 500` are harder for the retrieval system to satisfy. In contrast, the `VC` technique

[7] `gpt-4o (2024-08-06)`.
[8] For reference, the average Wikipedia article in 2025 contains 696 words: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia.
[9] The prompt and model configurations are available in the code repository.

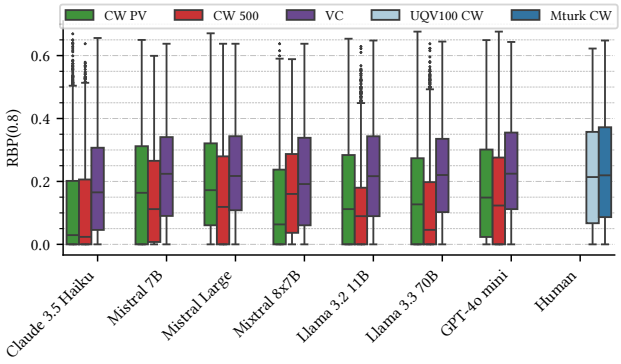


Figure 5: RBP(0.8) distribution per model and generation technique. The boxplot shows the IQR with the median line inside the box, and whiskers extending to 1.5 times the IQR. Outliers beyond this range are plotted individually.

produces queries that are more effective at representing the information need, leading to higher retrieval performance.

Among the LLM-based methods, the `VC` technique produces RBP score distributions that are most similar to those of the human-generated queries, though with lower variability – evidenced by a smaller interquartile range across models. This suggests that `VC` generates more consistent queries across different LLMs, while `CW PV` and `CW 500` exhibit greater variation in query difficulty depending on the model. These results indicate that `VC` aligns more closely with the diversity observed in human queries, although it does not fully replicate them.

While these findings should be interpreted with caution – as LLM-based relevance judgments may not fully correspond to human assessments and could introduce bias [18] – they suggest that `VC` is a promising approach for generating queries that better reflect human search behavior. It is important to note that our analysis is limited to a single retrieval system (BM25); results may differ when using other retrieval models, such as dense retrievers or neural re-ranking systems. We leave this exploration to future work, as it lies beyond the scope of the present study.

5 Discussion

Some prior studies have reported that LLM-generated queries lead to higher retrieval evaluation scores compared to human queries [4, 40]. However, these findings are primarily based on the TREC-DL collections, which consist of short passages and typically include only a single query per topic. In contrast, our study uses the ClueWeb12-B13 collection, which contains long, noisy web documents and we consider multiple queries per topic, following the approach used by Bailey et al. [10]. These differences in collection characteristics may explain the discrepancies observed in retrieval effectiveness results.

We acknowledge that query formulation is inherently variable among users. The ability to express an information need through a written query is shaped by contextual and individual factors, including ambiguity [6], anomalous states of knowledge [3], demographic background, and device usage [2]. These contextual influences highlight the complexity of modeling user queries. And were not necessarily captured in the datasets we used, which were collected under controlled conditions with crowd workers. As a result, the queries in the `UQV100 CW` and `Mturk CW` datasets may not fully represent the rich diversity of human search behavior.

Another limitation of this study is the age of the datasets: the `UQV100 CW` queries were collected over a decade ago, and the `Mturk CW` queries in late 2021. Although both are reasonable proxies for human-generated queries, they are not drawn from live query logs and may miss recent shifts in search practices, especially with the rise of generative AI. Future work could incorporate newer human-generated queries to better reflect current search behavior.

Despite these constraints, the datasets also offer certain advantages. Access to real-world, live query logs is typically restricted to major search providers, and publicly available logs tend to be heavily skewed toward short, navigational queries (e.g., "YouTube" or "Amazon"). In contrast, the queries collected from crowd workers are more representative of exploratory search tasks, where the user's goal is to discover and synthesize new information. Such tasks are of particular interest in IR research, and these datasets continue to be widely used for training and evaluation purposes.

An additional strength is the timing of the data collection. Both sets of human-generated queries were produced before the widespread availability of LLMs. This ensures that the queries reflect unaided human input. This is important in light of recent findings by Veselovsky et al. [50], who report that LLMs are now commonly used in crowd work. While their study shows that LLM-assisted responses are generally high-quality, it also finds that such responses tend to be more homogeneous than those written without LLM assistance. Thus, our use of pre-LLM query sets provides a clearer baseline for understanding differences between human and LLM-generated queries.

We further believe that several additional factors influencing query variation remain underexplored. These may include cognitive or physical disabilities, neurodivergence, and physiological states [24, 25]. While such aspects may have been considered in disciplines such as psychology and the social sciences [3], they are seldom addressed explicitly in information retrieval research. It is also likely that other relevant factors have yet to be identified.

To address the complexity inherent in understanding human search behavior, researchers have attempted to create synthetic data [21, 39], personas, and simulations of user behavior [55]. While these approaches provide valuable frameworks, they raise concerns regarding the extent to which simulated users accurately reflect real human behavior. Assessing the validity of such simulations remains an ongoing challenge. Consequently, conducting different types of studies—separately and in combination—is essential. Each methodology offers unique insights and, collectively, can enhance our understanding of how LLMs can be leveraged to generate synthetic queries and ultimately simulate user behavior more effectively.

User behavioral variability in the context of IR system use has long been recognized in the information-seeking literature [2, 16, 23, 25, 29]. Research on search behavior has demonstrated clear value in improving search systems and enhancing user experience. Nonetheless, due to the inherent complexity of human behavior, many open questions remain. Although LLMs may contribute meaningfully to

end-to-end IR pipelines, the diversity observed in LLM-generated queries does not necessarily reflect the variability seen in human querying behavior. This mismatch may stem from the factors discussed above. As such, relying solely on LLM-generated queries for training or evaluation may result in systems that are less effective for real-world users.

6 Conclusions

This paper examined the effectiveness of LLM-based query generation by comparing machine-generated queries with human-generated queries from the `UQV100 CW` and `Mturk CW` datasets. The analysis considered both linguistic characteristics: query length, uniqueness, and part-of-speech (POS) patterns; and retrieval performance: variability and effectiveness. While LLMs can produce queries that superficially resemble human-written ones, notable differences remain. In general, LLMs tend to generate longer queries with more diverse POS distributions and fewer duplicates. In terms of retrieval, LLM-generated queries often retrieve a more diverse set of documents than their human-generated counterparts, with less variance in effectiveness. This suggests that they may be expressing information needs in systematically different ways than humans.

Among the evaluated methods, the `CW PV` approach—based on multiple prompt variants—produced queries most similar to human-generated ones in terms of length. The `CW 500` method generated the most unique queries, exceeding the variability observed in actual human queries. The `VC` method, which incorporates contextual user and topic information, yielded the longest queries and exhibited the greatest variation in both length and POS structure. A noteworthy finding is that smaller, open-weight models often match or exceed the performance of proprietary and larger models when evaluated on their similarity to human queries. This observation challenges the common assumption that larger models inherently produce more human-like outputs. These results suggest that model size alone is not a reliable indicator of quality in the context of query generation, and that methodological choices play a critical role in shaping the output characteristics of LLMs.

Our findings suggest that while LLMs can assist in generating queries for different IR tasks, they should not be considered as replacements for human-generated queries. Queries produced by LLMs often exhibit different patterns of variability and may lack the subtle nuances found in human queries, potentially misaligning with real user information seeking behavior. Therefore, their use should be carefully assessed within the context of the specific task and dataset to ensure alignment with the intended retrieval objectives.

Future research should examine task-specific scenarios and integrate contextual or user-centered signals to better identify when LLM-generated queries are useful, when they may pose risks, and how to systematically validate their suitability. This includes studying their impact across different retrieval tasks, such as multi-agent IR and Retrieval-Augmented Generation (RAG), where automatic query generation may support AI agents in conducting research or enhance retrieval through diversification and fusion methods. It is also important to consider diverse user populations and evaluation settings. Establishing evidence-based guidelines for using LLMs in query generation will be essential to ensure their responsible integration into IR systems. Ultimately, a clearer understanding of the inherent trade-offs will enable more effective and accountable applications of LLMs in search technologies.

Acknowledgments

This research was partially supported by the Australian Research council through the ARC Centre of Excellence for Automated Decision-Making and Society CE200100005 and Discovery Project DP190101113, and was undertaken with the assistance of computing resources from the RMIT Advanced Cloud Ecosystem (RACE) Hub.

GenAI Usage Disclosure

The authors used generative AI (GenAI) tools during the research. Detailed descriptions of their use are provided in the paper. GenAI tools were employed to assist with coding; however, the authors remain fully responsible for all content.

References

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

[2] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*. Springer, 135–159.

[3] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2850–2862. doi:10.1145/3477495.3531711

[4] Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can Generative LLMs Create Query Variants for Test Collections? An Exploratory Study. In *Proc. SIGIR*. 1869–1873.

[5] 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 *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2024)*. 32–41.

[6] Abhijit Anand, Jurek Leonhardt, Venktesh V, and Avishek Anand. 2024. Understanding the User: An Intent-Based Ranking Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. ACM, Boise ID USA, 5323–5327. doi:10.1145/3627673.3679166

[7] Anthropic. 2025. Claude 3.5 Haiku. https://www.anthropic.com/claude/haiku

[8] Leif Azzopardi, Paul Thomas, and Alistair Moffat. 2019. cwl_eval: An Evaluation Tool for Information Retrieval. In *Proc. SIGIR*. 1321–1324.

[9] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User Variability and IR System Evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 625–634. doi:10.1145/2766462.2767728

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

[11] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2017. Retrieval Consistency in the Presence of Query Variations. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 395–404. doi:10.1145/3077136.3080839

[12] David Carmel, Simone Filice, Guy Horowitz, Yoelle Maarek, Oren Somekh, and Ran Tavory. 2025. The LiveRAG Challenge at SIGIR 2025. In *Proc. SIGIR*. 4199–4201. doi:10.1145/3726302.3733591

[13] David Carmel, Simone Filice, Guy Horowitz, Yoelle Maarek, Oren Somekh, Ran Tavory, Mehdi Ghissassi, Edo Liberty, and Roy Miara. 2025. SIGIR 2025–LiveRAG Challenge Report. *arXiv preprint arXiv:2507.04942* (2025).

[14] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[15] Ellese Cotterill. 2024. How to improve search without looking at queries or results. https://www.canva.dev/blog/engineering/how-to-improve-search-without-looking-at-queries-or-results/

[16] Carlos A. Cuadra and Robert V. Katter. 1967. OPENING THE BLACK BOX OF 'RELEVANCE'. *Journal of Documentation* 23, 4 (April 1967), 291–303. doi:10.1108/eb026436

[17] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2022. Topic Difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems* 40, 1 (Jan. 2022), 1–36. doi:10.1145/3470563

[18] Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and Guidelines for the Use of LLM Judges. In *Proc. ICTIR*. 218–229. doi:10.1145/3731120.3744588

[19] 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. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proc. ICTIR*. 39–50.

[20] Guglielmo Faggioli, Oleg Zendel, J. Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Inf. Retr.* 25 (2022), 94–122.

[21] Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Diverse Q&A Benchmarks for RAG Evaluation with DataMorgana. arXiv:2501.12789 (Jan. 2025). doi:10.48550/arXiv.2501.12789

[22] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *Proc. Nat. Acad. Sci.* 120 (2023), e2305016120.

[23] Ahmed Hassan, Rosie Jones, and Kristina Lisa Klinkner. 2010. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, New York New York USA, 221–230. doi:10.1145/1718487.1718515

[24] Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing Information Seeking Processes with Multiple Physiological Signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 1006–1017. doi:10.1145/3626772.3657793

[25] Kaixin Ji, Damiano Spina, Danula Hettiachchi, Flora Dilys Salim, and Falk Scholer. 2023. Examining the impact of uncontrolled variables on physiological signals in user studies for information processing activities. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1971–1975.

[26] Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2023. Mistral 7b. arxiv. *arXiv preprint arXiv:2310.06825* 10 (2023).

[27] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* (2024).

[28] Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. One Thousand and One Pairs: A "Novel" Challenge for Long-Context Language Models. *arXiv preprint arXiv:2406.16264* (2024).

[29] Diane Kelly. 2007. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval* 3, 1—2 (2007), 1–224. doi:10.1561/1500000012

[30] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Proc. NeurIPS*, Vol. 35. 22199–22213.

[31] Klaus Krippendorff. 2022. *Content Analysis: An Introduction to Its Methodology* (fourth ed.). SAGE Publications, Inc.

[32] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

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

[34] Alistair Moffat, Falk Scholer, Paul Thomas, and Peter Bailey. 2015. Pooled Evaluation Over Query Variations: Users are as Diverse as Systems. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1759–1762. doi:10.1145/2766462.2767728

[35] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* 27, 1 (2008), 27 pages. doi:10.1145/1416950.1416952

[36] OpenAI. 2025. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[37] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions With Human Feedback. In *Proc. NeurIPS*, Vol. 35. 27730–27744.

[38] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Proc. NeurIPS*. 6639–6649.

[39] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In *Proc. SIGIR*.

[40] Hossein A. Rahmani, Xi Wang, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, and Paul Thomas. 2025. SynDL: A Large-Scale Synthetic Test Collection for Passage Retrieval. arXiv:2408.16312 (Jan. 2025). doi:10.48550/arXiv.2408.16312 arXiv:2408.16312 [cs].

[41] Kun Ran, Shuoqi Sun, Khoi Nguyen Dinh Anh, Damiano Spina, and Oleg Zendel. 2025. RMIT-ADM+S at the SIGIR 2025 LiveRAG Challenge. *arXiv preprint arXiv:2506.14516* (2025).

[42] Lida Rashidi, Justin Zobel, and Alistair Moffat. 2024. Query Variability and Experimental Consistency: A Concerning Case Study. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval* (Washington DC, USA) *(ICTIR '24)*. Association for Computing Machinery, New York, NY, USA, 35–41. doi:10.1145/3664190.3672519

[43] Julian A Schnabel, Johanne Trippas, Falk Scholer, and Danula Hettiachchi. 2025. Multi-stage Large Language Model Pipelines Can Outperform GPT-4o in Relevance Assessment. In *Proc. WWW* (Sydney, Australia). ACM, New York, NY, USA.

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

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[46] Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv preprint arXiv:2304.06588* (2023).

[47] Shivani Upadhyay, Ehsan Kamalloo, and Jimmy Lin. 2024. LLMs Can Patch Up Missing Relevance Judgments in Evaluation. *arXiv preprint arXiv:2405.04727* (2024).

[48] 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).

[49] 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).

[50] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. 2025. Prevalence and Prevention of Large Language Model Use in Crowd Work. *Commun. ACM* 68, 3 (2025), 42–47. doi:10.1145/3685527

[51] Peilin Yang, Hui Fang, and Jimmy Lin. 2018. Anserini: Reproducible Ranking Baselines Using Lucene. *J. Data and Information Quality* 10, 4 (2018), 20 pages.

[52] Oleg Zendel, J. Shane Culpepper, and Falk Scholer. 2021. Is Query Performance Prediction With Multiple Query Variations Harder Than Topic Performance Prediction?. In *Proc. SIGIR*. 1713–1717. doi:10.1145/3404835.3463039

[53] Oleg Zendel, J. Shane Culpepper, Falk Scholer, and Paul Thomas. 2024. Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing. In *Proc. CHIIR*. 340–345. doi:10.1145/3627508.3638322

[54] Oleg Zendel, Melika P Ebrahim, J Shane Culpepper, Alistair Moffat, and Falk Scholer. 2022. Can Users Predict Relative Query Effectiveness?. In *Proc. SIGIR*. 2545–2549. doi:10.1145/3477495.3531893

[55] Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. USimAgent: Large Language Models for Simulating Search Users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Washington DC USA, 2687–2692. doi:10.1145/3626772.3657963

[56] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large Language Models for Information Retrieval: A Survey. *arXiv preprint arXiv:2308.07107* (2023).