Demographically-Inspired Query Variants Using an LLM

Marwah Alaofi
RMIT University
Melbourne, Australia
marwah.alaofi@student.rmit.edu.au

Nicola Ferro
University of Padua
Padova, Italy
ferro@dei.unipd.it

Paul Thomas
Microsoft
Adelaide, Australia
pathom@microsoft.com

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

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

Abstract

This study proposes a method to diversify queries in existing test collections to reflect some of the diversity of search engine users, aligning with an earlier vision of an 'ideal' test collection. A Large Language Model (LLM) is used to create query variants: alternative queries that have the same meaning as the original. These variants represent user profiles characterised by different properties, such as language and domain proficiency, which are known in the Information Retrieval (IR) literature to influence query formulation.

The LLM's ability to generate query variants that align with user profiles is empirically validated, and the variants' utility is further explored for IR system evaluation. Results demonstrate that the variants impact how systems are ranked and show that user profiles experience significantly different levels of system effectiveness. This method enables an alternative perspective on system evaluation where we can observe *both* the impact of user profiles on system rankings and how system performance varies across users.

CCS Concepts

• **Information systems** → **Query representation**; **Retrieval effectiveness**;

Keywords

Information retrieval; test collections; query variants; LLMs

ACM Reference Format:
Marwah Alaofi, Nicola Ferro, Paul Thomas, Falk Scholer, and Mark Sanderson. 2025. Demographically-Inspired Query Variants Using an LLM. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3731120.3744608

1 Introduction

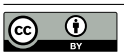
The 'ideal' test collections envisioned by Spärck-Jones and Van Rijsbergen [54] called for diversity in both documents *and* 'requests' (now known as queries) by varying content, type, source, and origin. This vision reflects Cleverdon's early emphasis on including 'representative' questions in evaluation. Although the creation of test collections – such as those developed in TREC – was largely inspired by this vision, it has not been fully realised. Most focus has been on document pooling and consistency of relevance judgments; far less attention has been given to the diversity of queries.

Past work has shown that factors like modality [12, 30], domain expertise [43, 61], age [7, 25], and language proficiency [11] influence query formulation, and consequently affect the accuracy of search results and user satisfaction. However, both capturing this variation and understanding its utility in test collection remain limited. Most system evaluation is based on so-called generic queries; how systems perform in other contexts, e.g., mobile users, or those with limited domain expertise, remains under-explored. This raises the question: how would the effectiveness of IR systems differ under a broader spectrum of queries? This calls for innovative approaches to develop cost-effective, comprehensive evaluation environments.

This work seeks to realise the original vision of the 'ideal' test collections by reintroducing the often-overlooked element of user diversity – explicitly controlling for variables that influence query formulation. The work describes an approach to expand a one query representation of information needs to better represent a diverse range of search engine users. It introduces a simulation of query *variants*: alternative queries that have the same meaning as an original (seed) query. The simulated incorporation of user properties is achieved through three LLM-based transformation methods applied to seed queries. These methods integrate profiles for specific personas (e.g., *Anne* the retired nurse), user groups (e.g., *Voice* search users), and single textual transformations (e.g., *paraphrasing*). Table 1 shows examples of variants from the three methods, along with their effectiveness scores on two IR systems.

The work aims to *validate* LLM-generated query variants and explore their *retrieval impact*. A variant is deemed *valid* if it is formulated differently yet shares the same meaning as the seed query, and if it reasonably manifests its profile. The study explores the impact of variants on the effectiveness of IR systems, and offers new insights regarding their utility in system evaluation. An overview of the experimental design is given in Figure 1. We explore the following research questions:

RQ1 Are LLM-generated query variants valid?
- **1.1** Lexical Variation: Do variants differ lexically from seed queries?
- **1.2** Semantic Similarity: Are variants semantically similar to seed queries?
- **1.3** Profile Alignment: Are variants likely to be generated by their respective profiles?



ICTIR '25, Padua, Italy

ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744608

Table 1: An example seed query from DL22 and its LLM-generated query variants and their NDCG@10 scores using two IR systems.

Profile	Variant	BM25	ANCE
Seed	**What to do if antibiotics cause nausea**	**0.23**	**0.35**
Emily *the scientist*	***Coping strategies*** *for nausea resulting from antibiotic use*	0.00	0.51
Noah *the curious child*	*How can I feel better if medicine for sickness makes me* ***throw up****?*	0.19	0.10
Non-native	*What should one do when antibiotics* ***make feel sick***	0.07	0.13
Voice	***Please find solutions*** *for dealing with nausea caused by antibiotic medication*	0.26	0.63
Paraphrasing	*What to do when antibiotics make you* ***queasy***	0.35	0.50
Naturality	***Treatment*** *for nausea from antibiotics*	0.15	0.47

RQ2 Do the generated query variants change the outcome of system evaluation?

- **2.1** System Ranking: How do variants impact on relative system effectiveness rankings?
- **2.2** User Equity: Do systems serve users equally? (E.g., is a mobile user served as effectively as a voice search user?)

We make the following contributions in understanding the utility of an LLM in simulating queries and creating relevance labels: [1]

- **C1** An adaptable evaluation pipeline that, based on search engine user properties, generates query variants for fine-grained system evaluations. This assists in highlighting problems faced by certain demographics, and identifying where the system performs effectively or inadequately.
- **C2** A publicly available dataset of approximately 7K query variants for the Deep Learning Track of TREC 2021 (DL21) and Deep Learning Track of TREC 2022 (DL22), with human assessments covering a 10% sample of each profile, to support studying similarity and profile alignment.
- **C3** An evaluation of GPT-4o prompted relevance labels for DL21 and DL22. The labels – 146K query-passage pairs – are made available publicly, with 13K labels human-judged by the National Institute of Standards and Technology (NIST), enabling further research into LLM relevance labels.

2 Related Work

We examine two aspects of past work.

2.1 Users and Search Queries

The relationship between user characteristics, context, and search behaviour has been widely studied. For example, mobile search queries are often shorter than desktop ones, likely due to limited input options on mobile devices [12]. The accuracy of these queries is also compromised by the fragmented attention of users, resulting in fewer hits and lower performance [30]. On the other hand, voice queries, which are increasingly prevalent, are generally longer and more closely resemble natural spoken language [29].

Language proficiency plays a crucial role in the success of online search queries. As highlighted in a qualitative study by Chu et al. [11], non-native English speakers faced challenges in formulating queries for English content. Chinese participants had difficulty expressing specialised or academic terms and opted for simpler keywords to reduce inaccuracies. Hungarian participants exhibited a similar pattern, preferring shorter queries in English. Strategies such as translating native language queries into English, and beginning with simple keywords and progressively refining them by browsing search results, are used, which affect the accuracy of searches and increase the time taken to find relevant information.

Similarly, research on domain expertise has identified distinct differences in search behaviours between experts and novices. For example, White et al. [61] demonstrated that experts typically formulate longer queries and employ more technical vocabulary from domain-specific lexicons, 50% more likely than novices. More recently, Monchaux et al. [43] showed that psychology experts outperformed novices in finding correct answers, especially in complex searches. This success was mainly attributed to their strategy of frequently reformulating queries with domain-specific terms.

In studies examining the impact of age on search, differences have been observed between older and younger adults. For instance, Sanchiz et al. [52] demonstrated that older adults' search queries were less elaborate, significantly incorporating more keywords from the problem statements and adding fewer new keywords. They were less efficient, needing more time to complete search tasks. Children also encounter unique challenges in their search, primarily due to their in-development vocabulary and cognitive skills, resulting in misspelling and more intuitive/natural queries. Bilal and Gwizdka [7] highlighted that children predominantly use phrase and question-like queries, particularly younger children. This aligns with earlier log analysis research by Duarte Torres et al. [25], which suggests that children's queries are longer, more natural, and less effective than those of adults.

The concept of "persona" was first introduced by Cooper and Reimann [16] as a design tool and has since been useful in establishing a user-centric focus. A persona describes user characteristics and goals, ideally based on an understanding of target users. While widely used in user interaction design, limited research uses personas to help understand, guide, and evaluate the design and development of IR systems [e.g. 36].

2.2 Test Collections and Query Variants

IR system effectiveness is commonly measured using offline test collections, which typically represent each information need with one query. This *single query assumption* underpins the collection of relevance judgments [59] and has contributed to a consistent and reusable evaluation environment. However, in recent years, there is evidence of the importance of query variants on system evaluation [1, 6, 20, 46, 48].

Given the recent promising outcomes from using LLMs in relevance labelling [26, 56] and user query simulation [2], we are prompted to question this assumption. This leads us to consider

[1]Data is available at: https://github.com/MarwahAlaofi/demo-qv

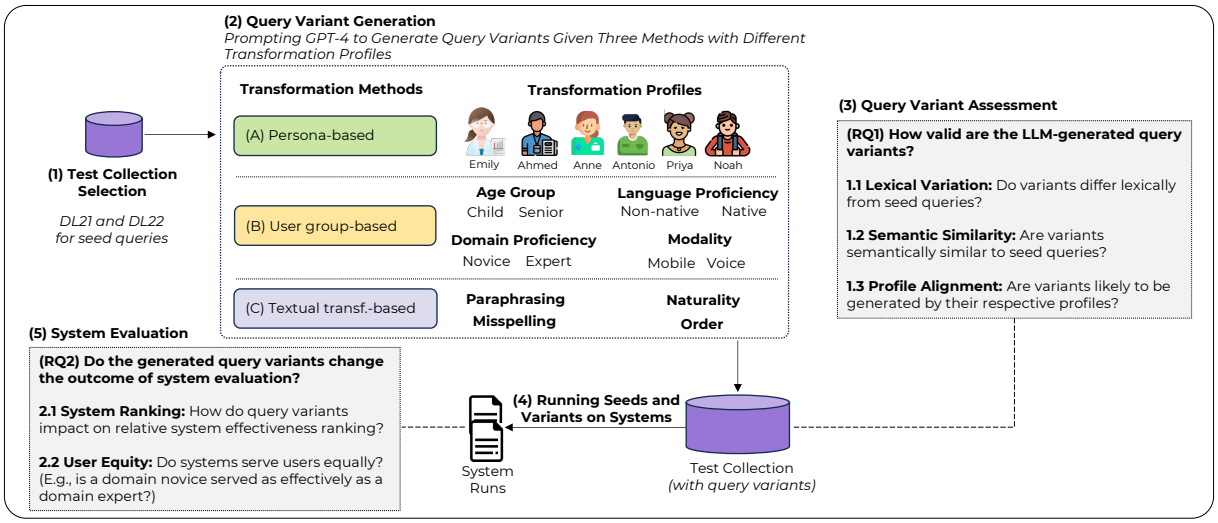


Figure 1: An overview of the experiment pipeline and the research questions.

whether we can move beyond assuming a single query will represent a large and diverse range of users, whose information-seeking behaviour varies widely.

Studies show that users issue a substantial number of query variants despite having the same information need [5, 40]. The utility of variants in identifying relevant documents dates back several decades [53] and was explored in previous TREC tracks [10]. Their effect in constructing document pools is demonstrated to be comparable to that of systems [42]. Recently, Alaofi et al. [2] used an LLM to simulate human queries for building similar document pools. Research on variants faces challenges. Methods like crowd-sourcing [5, 40] and click graphs [63] have been used to sample variants from populations and datasets. However, these methods have limitations: crowdsourcing is costly to scale, and click graphs are noisy and lack user properties and context. Given LLM capabilities in simulation, this work explores LLM ability to generate variants.

Beyond generating relevance labels in IR, several studies have used LLMs to create synthetic data, including queries [8, 9, 21, 47, 51], documents [4], and relevance explanations [27], aiming to build more effective or less biased retrieval models. The generated data is optimised to maximise scores but does not necessarily represent real users. The question of how we can use LLM capabilities to simulate the queries of various user demographics and assess their value for system evaluation is under-explored.

Drawing on the literature about search engine user characteristics, and inspired by the use of personas as a user-centered design tool, this study uses an LLM to integrate the human element into evaluation practices, offering a diverse representation of users and thus following a more 'pessimistic' approach to system evaluation [24], moving beyond average utility to examine a range of query variants of varying quality to better understand system behaviour.

3 Experiment Design

3.1 Test Collection Selection

We sourced our *seed queries*, the basis queries for generating query variants, from the passage retrieval tasks in DL21 [18] and DL22 [19]. This track is used for benchmarking ad hoc passage retrieval methods. Both years use the expanded MS MARCO dataset (v2), which contains 138 million passages [44], about 16 times larger than v1.

The DL21 track, due to the increased corpus size, included a larger number of relevant passages, leading to concerns about score saturation and re-usability [18, 60]. To address these issues in DL22, the organizers implemented several strategies, including the selection of more challenging topics, specifically, those that did not contribute to the development of the MS MARCO corpus. The DL21 and DL22 tracks include 53 and 76 queries, respectively, with relevance judgments on a four-point scale. The passage corpus for our experiments is obtained through the ir_datasets tool [38].

3.2 Query Variant Generation

We use three methods for generating query *variants* from seed queries. They follow the same approach but are differentiated by the properties of their transformation profiles.

- **Persona-based**. Six personas are chosen with varying properties of age, education, first language, and search modality preference (e.g., mobile or voice). Each persona is described by background information and some search-influencing factors.
- **User group-based**. Eight different user groups are chosen to reflect groups of users sharing one single property (e.g., age). As with personas, we use age, education, first language, and search modality preference. In contrast to the use of personas, this approach could be used to associate metric variability with those properties in isolation.
- **Textual transformation-based**. This category does not relate directly to user profiles but instead stems from a taxonomy of variants developed by Penha et al. [46]. Seed queries undergo four distinct textual transformations: paraphrasing, which involves the replacement of words; order change, where the sequence of words is shuffled; naturality, where queries are transformed into either keyword-based or natural language forms; and misspelling, where seed words are misspelled.

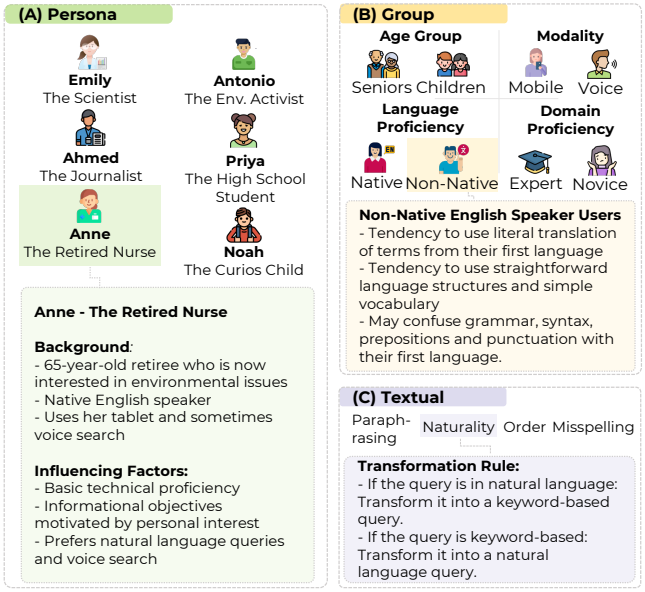


Figure 2: Transformation profiles and example descriptions.

These methods represent varying levels of abstraction over users or queries, enabling flexible evaluation depending on available user information. Persona-based variants may suit scenarios with detailed user information, user group abstractions could be applied when only partial traits are known, and textual transformations offer a simple option when no user data is available.
Figure 2 presents all of the profiles associated with each method, together with one complete example profile description for each. Other profile descriptions are publicly available to foster community research and support reproducibility.

In all methods, seed queries together with the transformation profiles are fed into an LLM, specifically the GPT-4 model, with the temperature parameter set to one. Initial experiments conducted at lower temperatures yielded more deterministic results, but these were marked by substantially less variation, and with models often getting stuck in repetitive word sequences. Initial human assessments of the temperature set to one yielded positive results, thus we continued with this setting. The model is prompted to generate three variants for each seed and transformation profile combination. The template used to create all prompts is given in Figure 3.

In addition to the profile-based variants, we generate *neutral query variants*, which are generated by GPT-4 without any transformation profiles. These variants help to isolate the impact of GPT-4's capabilities in generating variants, and separate this from the influence of the transformation profiles. For this purpose, we use a similar prompt as shown in Figure 3, but remove the instructions specific to the transformation profile (2 and 3b). The total number of variants generated by each method is given in Table 2.

3.3 Human Assessment of Query Variants

When assessing query variants, our objective is to quantify three aspects: (1) *Lexical similarity* to seed queries, where lower levels are preferable; 2) *Semantic similarity*, ensuring that variants retain

Given the seed search query: "{seed_query}", do the following:
1) Consider the query search intent.
2) Consider the {transformation_profile} described below.
3) Write three web search queries that:
   a) share the same meaning of the seed query.
   b) and are likely to be written by the {transformation_profile}.
Write the queries in a JSON array. Do not add text before or after the array.
{transformation_profile_description}

Figure 3: The prompt template used to generate variants.

Dataset	Seed	Query Variant Sets			
		Persona	Group	Textual	Neutral
DL21	53	954	1272	636	159
DL22	76	1368	1824	912	228

Table 2: Number of queries (seeds) and variants as generated by the transformation methods and a neutral setting.

the same meaning as their seed queries; and 3) *Profile alignment*, whereby variants represent their respective profiles.

To evaluate lexical similarity (1), we use the Jaccard Index, which calculates the ratio of common words to the total words in both seed queries and their variants. To improve accuracy, we first convert all queries to lowercase, remove punctuation, and truncate words using the Porter stemmer [49]. Other metrics, such as BLEU and ROUGE, were found to lead to congruent conclusions; for brevity, we therefore report just the Jaccard Index.

To evaluate semantic similarity and profile alignment (2 and 3), we conducted a human assessment on a randomly selected sample of 10% of the variants created using each transformation profile.

We recruited annotators from Prolific,[2], choosing those whose native language is English and who had an acceptance rate of 98% or higher. Each annotator was presented with pairs of seed queries and their corresponding variants. The pairs were divided into small batches. Two annotators were assigned to each batch. The tasks for the annotators were twofold: 1) to assess the semantic similarity of each variant to its seed query, and 2) to identify if the variant matches its corresponding profile. A pair is deemed similar if both annotators independently agree on this assessment. The accuracy of semantic similarity is then calculated by dividing the number of similar instances by the total sample size. Further details about the profile alignment task are described in the following section. The *Order* change and *Misspelling* profiles were not assessed for semantic similarity as their transformations are at the lexical level. To ensure accuracy and prevent fatigue, the batches were spread over multiple days, with each batch having a maximum of two transformation profiles. Annotators were compensated at a rate of $15 per hour, based on an estimate of the time needed to complete each from pilot studies.[3]

For quality assurance purposes, gold standard questions were integrated into the assessment process, with one such question included for every 10 query pairs. These gold questions were handcrafted to feature seed queries paired with queries that share words but differ in meaning, e.g., *"what foods should you stay away from if you have asthma","what types of food is good for fat loss?"*. While all

[2]https://www.prolific.com/
[3]This protocol was approved by RMIT University Human Research Ethics Committee.

contributed annotations were compensated, only those from annotators who correctly answered all gold questions were ultimately used for analysis.

For profile alignment, we also aim to compute three features of each transformation profile: 1) variant length; 2) readability, as measured by the Flesch-Kincaid Grade Level Score [34]); and 3) lexical diversity, to measure the variety of unique words in profile variants, calculated as the ratio of distinct words to the total word count. To prevent artificially inflating the ratio, misspelled words in the *Child* queries are corrected. We would expect, for example, profiles representing children to demonstrate higher readability and smaller vocabulary diversity compared to those representing adults. These features are reported exclusively for user-specific (persona and user group-based) profiles, as they appear relevant in that context. The Mann-Whitney U test is used to assess the pairwise differences in mean scores of these features across all profiles, with a significance level α set at 0.05.

3.3.1 Human Assessment of Profile Alignment. Assessment was structured slightly differently depending on the method used for generating query variants – although, in all cases, accuracy was calculated as the number of correct instances divided by the total sample size:

Persona-based. For each variant, annotators were asked to identify which persona was more likely to have generated the query variant. Annotators were presented with three options: the correct persona; a candidate persona randomly chosen from the remaining five; and, an option indicating that both personas are "equally likely". The variants of each persona are ultimately paired with all other personas as candidates.

The objective was not to measure how distinguishable a persona was but rather to measure if the variants appeared to have been written by the persona. A choice of "equally likely" was considered the same as the correct persona. The choice of a persona was deemed correct if both annotators agree on the same persona which matches the correct persona. Any other scenario is considered incorrect.

User group-based. For each sampled variant, annotators had three options: the correct user group, their opposite group, and "equally likely". For instance, with non-native English speakers, the options were "native", "non-native", and "equally likely".

Textual transformation-based. For *Paraphrasing* and *Naturality* transformation profiles, annotators were presented with a description of the transformation profile. They were then asked to assess whether the variants were transformed in accordance with the given description. For *Order* change and *Misspelling*, the validation was conducted automatically across the entire set of variants, unlike the 10% sample used for other transformation profiles. A variant is considered to have a different order if the arrangement of words differs from the seed, i.e., the variant is not equal to the seed but a sorted list of the seed's words should match a sorted list of its variant's words for the transformation to be considered valid. For misspellings, we utilised the Bing Spell Check API.[4] A transformation was deemed valid if the variant introduced one or more misspelled words that, when correctly spelled, are part of the seed query.

[4]https://www.microsoft.com/en-us/bing/apis/bing-spell-check-api

3.4 Running Seeds Variants on IR Systems

Experimental comparisons are conducted across 15 representative IR systems: three lexical models (TF-IDF, DLH and BM25) both with and without RM3 and BO1 query expansions; four pre-trained neural re-rankers (BERT [23], ColBERT [33], ELECTRA [13], and monoT5 [50]); one neural-augmented index (Doc2Query[45]); and one dense model (ANCE [62]). This selection enables meaningful comparisons, where practitioners may consider evaluating and comparing both traditional and more advanced, less efficient systems.

Retrieval was performed using Pyterrier [39], except for Doc2Query for which Pyserini [37] was used over a pre-built corpus with doc2query-T5 expansions.

3.5 System Evaluation

Evaluation is based on NDCG@10, the official metric for DL21 and DL22. Systems were evaluated across all variant sets: seed, persona-based, user group-based, and textual transformation sets. As anticipated, we encountered a substantial portion of missing relevance judgments in the top ten results retrieved in response to query variant sets. On average, at a cutoff of 10, 41% and 48% of relevance judgments are missing in DL21 and DL22, respectively, raising concerns about the evaluation outcomes.

With the promising results of using LLMs to provide relevance labels [26, 56], we used an LLM for relevance labelling. Based on extensive comparisons and gullibility tests of various LLMs as conducted by Alaofi et al. [3], we selected GPT-4o, which demonstrated the highest agreement with human judgments and the greatest robustness against keyword-stuffing gullibility tests given the same test collections we are using in this study.

We generate relevance labels for the top ten passages of all systems in response to all query variant sets. We used the same *Basic* prompt used in Alaofi et al. [3] and used the same 4-point scale of DL21 and DL22. We initially provided seed queries to GPT-4o to judge passage relevance, but noticed a measurable bias favouring retrieval using seed queries and thus opted to generate backstories to represent the information needs and use them instead of seed queries to assess passage relevance, following Bailey et al. [5]. Backstories are available for public use.

We assessed the performance of GPT-4o's relevance labels against the available NIST relevance judgments using Mean Absolute Error (MAE), and evaluated agreement with NIST judges using Cohen's κ [15] and Krippendorf's α [35]. When relevance scores are binarised, scores of 0 and 1 are mapped to 0, while scores of 2 and 3 are mapped to 1. The results are presented in Table 3.

GPT-4o labelling performance. The performance of GPT-4o on a binary scale (κ of 0.50 and 0.46) is comparable to human-to-human agreement levels, reported as 0.52 on the OHSUMED dataset [31] and 0.41 on the TREC-6 ad hoc track [17]. Similarly, its agreement on a graded scale (α of 0.58 and 0.62) is at the higher end of the observed human-to-human agreement range, which spans from 0.41 to 0.69 as measured on the TREC 2004 Robust Track [22]. These results indicate that the moderate level of agreement between GPT-4o and NIST judges falls within the typical range of agreement

Table 3: MAE and pairwise agreement between NIST judgments and GPT-4o labels, measured using Cohen's Kappa (κ) on a binary scale and Krippendorf's Alpha (α) on a 4-point ordinal scale.

Dataset	qrels #	Binary		Graded	
		MAE	κ	MAE	α
DL21	5141	0.25	0.50	0.69	0.58
DL22	8011	0.20	0.46	0.55	0.62

observed among human assessors, and thus, GPT-4o is used for relevance labelling.

Analysis metrics. To answer if the choice of transformation profile significantly impacts systems ranking (RQ2.1); e.g., is a system ranking for *Emily* different from that for *Ahmed*, we ran a pairwise analysis where we paired profiles that originate from the same transformation method, and with seeds. For example, pairs from the persona-based transformation method would be *Emily-seed*, *Emily-Ahmed*, *Emily-Anne*, etc. For each pair of profiles, we consider the question of whether systems are ranked differently using **Kendall** τ [32], a correlation measure used to compare rankings under different conditions – in this case, across profiles.

For each profile, we perform a two-way ANOVA (topic and system effects). We compare each pair of profiles – considering all possible pairs of systems (i.e., 105 pairs from 15 systems) – according to the following measures [28, 41]:

- **Active Agreement (AA)**: Number of instances that agree on which system is significantly better across both profiles.
- **Active Disagreement (AD)**: Number of instances that disagree on which system is significantly better across both profiles.
- **Mixed Agreement (MA)** & **Mixed Disagreement (MD)**: Similar to AA and AD, but significance observed in one profile only.
- **Passive Agreement (PA)** & **Passive Disagreement (PD)**: Similar to AA and AD, but significance not observed.

We use three-way ANOVA to examine the effect of topic, system, profile, and their interactions on how systems treat different profiles (RQ2.2). We set α to 0.05 in all of our analyses, accounting for multiple comparisons among systems with Tukey's HSD [57].

4 Results and Discussion

The results address two main research questions: assessment of query variants and the impact of variants on system evaluation.

4.1 Query Variant Assessment

RQ1.1 Do variants differ lexically from seed queries? We quantify the lexical difference between the generated variants and their seed queries using Jaccard Index, where 0 signifies no overlap and 1 maximal overlap (not necessarily identical, due to text pre-processing, see Sec 3.3). Figure 4 (A) presents the distribution of Jaccard Index for variants generated by all profiles. The average Index for neutral variants is indicated by a red dashed line. The average Index across all profiles is 0.32. The average is lower for profiles possessing greater domain expertise, i.e., the domain *Expert* and *Emily* the scientist. The impact of using profiles in generating lexically different variants becomes evident when these are compared to neutral variants generated without profiles. Variants from all profiles, except for the *Non-native* user profile, show statistically significantly lower overlaps than the neutral variants.

RQ1.2 Are variants semantically similar to seed queries? Column one of Table 4 shows the accuracy of semantic similarity based on the human assessment of the sampled variants from each transformation profile, paired with their respective seed queries (in total, 611 pairs). Due to the high subjectivity of query semantic similarity assessment [55], we required a consensus between the two annotators for a pair to be considered semantically similar. Disagreements in variant similarity occurred in 9.82% of the pairs, primarily in cases of highly specific variants, for example *Ahmed* the journalist's query: *"budget for a work trip to Bangkok for a reporter"* for the seed query *"how much money do I need in Bangkok."* Similarly, ambiguous seed queries could lead to different interpretations by annotators. For example, for the seed query *"Collins The Good to Great"*, *Anne* the retired nurse, and who prefers voice search, and *Noah*, the child, generated the queries *"audio explanation of Jim Collins Good to Great book"* and *"what's the book Good to Great by Collins about?"*.

Another reason for this disagreement was observed when expert profiles used more scientific terms, e.g., *Emily* the scientist's query *"research studies on magnolia bark extract use for anxiety mitigation"* for the seed *"how much magnolia bark to take for anxiety"*. Country-specific expressions may also contribute to this disparity, for example, *"how to help a **jammed finger**"*, where a UK-based annotator interpreted the meaning differently than what is commonly understood in American English for a joint injury. Some disagreement is perhaps due to human errors; for example, *"Collins The Good to Great"* and its paraphrased variant *"Collins The Good to Excellent"*, was annotated as similar by one of the annotators. This could be due to a failure to recognise that "The Good to Great" is a book title or an assumption that the user may have forgotten the exact title, thereby considering it a valid paraphrase.

Generally, semantic similarity scores are high, substantially exceeding the agreement of both annotators by chance (25%), with minor shifts observed in some queries of the *Child* and *Expert* profiles. These are primarily attributable to slight oversimplifications or elaborations, which reflect the nature of these profiles, e.g., *Noah's* query *"can you **blame** the landlord if a bad guy breaks in and you get hurt"* for the seed query *"are landlords **liable** if someone breaks in a hurts tenant"*. Nevertheless, the core intent remains intact.

RQ1.3 Are variants likely to be generated by their respective profiles? Column two of Table 4 presents the accuracy of profile alignment as assessed by the annotators and Figure 4 (B, C and D) show the automatic calculation of query features for profile alignment, as described in Sec 3.3.

It is notable from Table 4 that, for persona-based profiles, *Noah* the child and *Emily* the scientist are most easily distinguishable when paired with other random profiles as set in the experiment. Other profiles also demonstrate a reasonable level of accuracy; however, challenges arise when profiles with similar characteristics are paired. In such cases, the differences between the profiles may not be sufficiently distinct to be fully exhibited in the queries. It is worth

	Profile	Similarity	Profile Alignment
Persona	Emily	0.97	0.87
	Ahmed	0.95	0.68
	Anne	0.92	0.70
	Antonio	1.00	0.62
	Priya	0.97	0.55
	Noah	0.92	0.73
Group	Child	0.67	0.74
	Senior	0.82	0.85
	Non-native	0.92	0.62
	Native	0.84	0.84
	Novice	0.84	0.70
	Expert	0.72	0.79
	Mobile	0.92	0.95
	Voice	0.97	1.00
Textual	Naturality	0.97	0.97
	Paraphrasing	0.76	0.76
	Order	-	0.69
	Misspelling	-	0.80

Table 4: Human assessment of variant similarity and alignment.

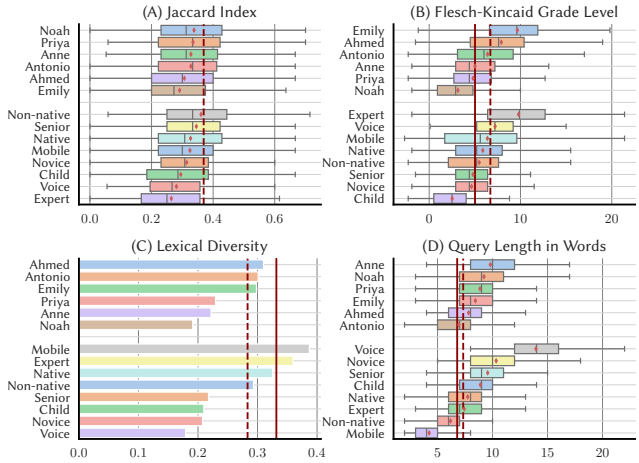


Figure 4: Distribution of profile variant properties ordered by mean for each transformation method. Vertical lines show the mean value for neutral variants (dashed) and seed queries (solid).

noting, though, that all profiles have an alignment accuracy that is substantially higher than the random chance of the agreement of both annotators (44.44%), and most inaccuracies stem from disagreements, which we consider incorrect—for instance, for *Priya*, only 7.89% of pairs were misclassified by agreement, while 36.84% resulted from disagreement.

User group-based profiles seem to have a slightly higher accuracy in comparison, which could be due to being paired with opposite profiles. The *Order* change transformation, though simple, does not perform as expected. It adds or removes words in the process, but those changes are minor (mostly in stop words).

In assessing readability, as illustrated in Figure 4 (B), where scores correspond to the school grade level required for understanding, significant variations in means across persona-based profiles are observed. This is true for all pairs except *Anne* and *Priya*. As expected, *Emily*'s queries have the lowest readability, while *Noah*'s have the highest. Other profiles do not show much deviation from the readability means of the seed and neutral queries. Contrary to expectations, the queries produced by non-native English speakers, *Ahmed* and *Antonio*, are not more readable than those by the native speaker, *Anne*. Similar results are observed for the user group-based profiles: the *Expert* and *Child* profiles produce queries that are significantly less and more readable, respectively, compared to all other profiles, while *Native* and *Non-native* are similarly readable.

Figure 4 (C) and (D) show the lexical diversity and the distribution of query length in words for all profiles. As expected, child profiles have lower lexical diversity when compared to most of the other profiles; similarly, *Non-native* is lower than *Native*. *Anne*, *Noah*, and *Priya*, whose profiles prefer natural language queries, have significantly longer queries when compared to other profiles. Not surprisingly, the *Voice* and *Mobile* profiles produced significantly longer and shorter queries, respectively.

While this analysis does not verify a direct correspondence between the generated variants and "real world" users, a qualitative examination of the query variants reveals some capabilities not only at the linguistic level but also in terms of information-seeking behaviour. For example, for the seed query *"was Friedrich Nietzsche an atheist," Ahmed* the journalist's query aims to find primary sources: *"writings of Nietzsche indicating his atheism."* Cases of failures exist, too, with a few having distorted meanings or using unusual query openers.

The results show that we can produce query variants that map their profiles and have properties that align with how those profiles are defined. While one might ask if the variants represent "real" users, this is a common concern about many evaluation campaign tasks: in TREC, queries are often sourced from a limited and non-representative group of individuals, with typically no validation to assess how realistically they simulate a broader population of searchers. The method we present here builds on prior research in search behaviour to capture common and meaningful query formulation patterns, and explicitly tests the validity of the generated variants. This goes beyond how traditional test collections are built – including those that capture human variants – where diversity among participants is hard to obtain and only automated or semi-automated methods are used to assess the quality of queries produced by crowdworkers [5, 40]. Large-scale query log sampling may reflect user diversity but has many limitations. Searchers are anonymous, have varied interests and unclear intent, making it difficult to control for demographics and topics.

4.2 System Evaluation

RQ2.1 How do query variants impact system ranking? Figure 5 shows the performance of all systems in response to seed queries and transformation profiles across DL21 and DL22. Performance varies widely depending on the transformation; notably, the *Child* and *Misspelling* profiles lead to the lowest scores across all systems.

Table 5 shows an ANOVA analysis for NDCG@10 of the topic, system, profile factors, and their interactions. The effects are all statistically significant. Profile has a large effect size, implying substantial changes in the overall system performance across profiles.

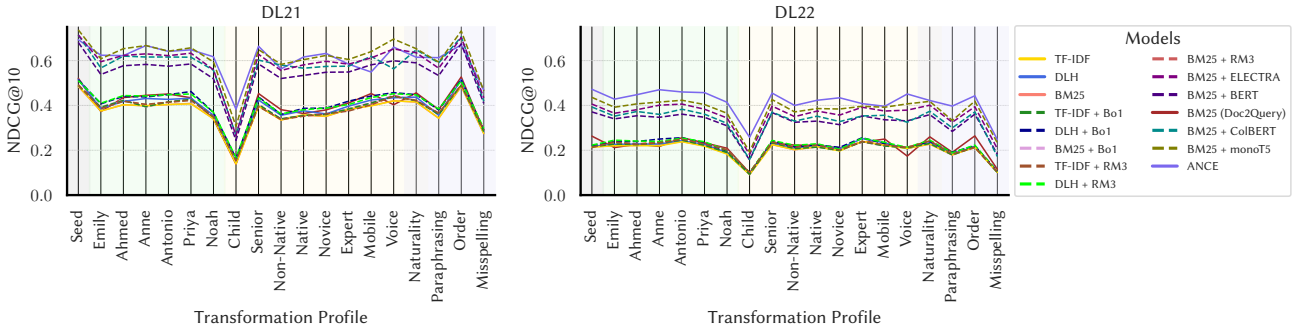


Figure 5: Selected IR systems ranked using NDCG@10 across transformation profiles on DL21 and DL22.

Table 5: ANOVA effect size ω_p^2 for nDCG@10. Dark and light shades denote large (≥ 0.14) and small (≤ 0.06) effects respectively.

Source	DL21	DL22
Topic	0.7338	0.7741
System	0.4439	0.3926
Profile	0.3300	0.2189
Topic*System	0.3149	0.3470
Topic*Profile	0.6713	0.6566
System*Profile	0.0104	0.0156

We see this change in Figure 5, where a wide range of NDCG@10 values are shown across the profiles. It would appear that there is potential for systems to mitigate such variation in effectiveness. However, the Table shows that the systems used in our test do not perform any such mitigation: the System*Profile interaction shows only a small effect size. This is also visible from Figure 5, where system lines cross to a limited extent. Note, however, that the retrieval systems used in the test were not designed with variants in mind. A recent investigation of a novel system that improves the poorly performing variants of a topic appears to show great promise [51]. Therefore, the small effect size of the profiles on systems observed in this study likely underestimates their value, as they were validated using a set of systems that were not designed to be variant-aware. We also observe a large effect size for the Topic*Profile interaction, suggesting that query variants from different profiles will have different levels of influence depending on the topic.

Comparing the performance of systems on seed queries with other profiles in DL22, some profiles generate queries that are more effective (e.g., *Expert* profile with traditional models). The observation supports the human assessment results showing high semantic similarity between variants and seed queries: if variants distort the meanings of the seeds, we would expect a consistent decline in system performance. A different behaviour is, however, observed in DL21, where most systems achieve their best performance with seed queries, as measured by NDCG@10. We hypothesize that the observed behaviour may be attributable to a bias in the corpus towards seed queries, a result of integrating relevant passages to the queries from Bing into the corpus, as opposed to DL22, where only queries that did not contribute to the corpus were included [19].

To understand how different profiles might alter the conclusions we draw from system effectiveness comparisons. For instance, one might ask if System A outperforms System B in the context of voice search, while exhibiting inferior performance in a mobile query scenario. Table 6 shows the average τ of pairwise differences between system rankings across different profiles. A τ of 1 indicates complete ranking agreement, -1, ranking is reversed.

Table 6: Kendall's τ correlation of pairwise differences between system rankings and the ratio of AA, AD, MA, MD of all profiles of each transformation method, evaluated using NDCG@10.

		τ	AA	AD	MA	MD	PA	PD
DL21	(A) Persona	0.74	0.48	0.00	0.01	0.00	0.38	0.13
	(B) Group	0.71	0.48	0.00	0.02	0.01	0.36	0.14
	(C) Textual	0.76	0.48	0.00	0.11	0.02	0.30	0.10
DL22	(A) Persona	0.77	0.50	0.00	0.02	0.00	0.36	0.12
	(B) Group	0.69	0.50	0.00	0.03	0.00	0.32	0.15
	(C) Textual	0.69	0.50	0.00	0.03	0.00	0.32	0.15

Table 6 also reports the average consistency between pairs of system rankings in the AA, AD, MA and MD groups, as defined in Section 3.5. The always zero AD, complemented by low or zero MD, indicates that the profiles do not systematically introduce excessive changes, i.e., leading to completely opposite conclusions regarding relative system effectiveness when using one profile or another. AA for around 50% of the pairs is a good indicator of stability. On the other hand, the MA of 1% to 11% of the pairs suggests that diverse profiles are able to spotlight diverse significant differences between systems. The average τ in most methods given DL21 and DL22 is below the threshold of 0.90, which indicates that two test collections (in our case profiles) are not identical in how they rank systems [58]. This significant but moderate change in the system rankings is consistent with maximum PD being 15% of the pairs as well as the small size of the System*Profile interaction effect reported in Table 5. Going beyond the metrics, we observe that certain profiles, such as the voice profile in DL21, can cause substantial changes in system rankings compared to the seed. Moreover, we can observe that the top-performing system may vary depending on the profile.

RQ2.2 Do systems serve users equally? Figure 6 shows the result of the Tukey's HSD test on the NDCG@10 marginal mean scores of different profiles in DL21 and DL22 against the seed queries, averaged across systems, with the error bars representing the confidence intervals around the marginal means. The same system bias in favour of seed queries in DL21 is also present here, aligning with

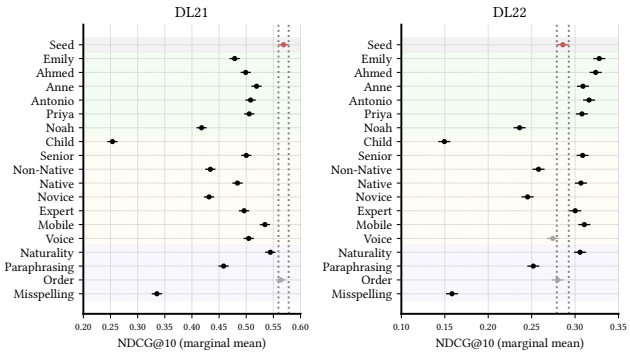


Figure 6: Tukey's HSD test on the marginal means of NDCG@10 scores produced by each profile in both DL21 (left) and DL22 (right).

our previous discussions. For this reason, we focus our discussion on DL22 and refer to DL21 when needed. Noticeable differences across profiles are observed. When examining persona-based profiles (the green region), we see that *Noah* receives the lowest mean score, substantially lower than all other profiles and significantly lower than the seed queries. Other profiles exhibit different levels of performance, with all showing a significant improvement over the seed queries in DL22, which further validates that the generated queries do not drift from the meaning of the seed queries.

Examining user group-based profiles (yellow region), we observe that the *Child* profile has the lowest score, even lower than that of *Noah*. This is most likely because the profile is prompted to generate misspelled words, mimicking a child user. Conversely, the *Senior* profile significantly improves upon the seed mean.

Comparing *Native* and *Non-native* English speaker profiles, we notice that the *Non-native* profile has a significantly lower score than the seed, while the *Native* profile is higher. A similar trend occurs with domain proficiency, where the *Novice* has a significantly lower mean than the seeds while the *Expert* has a significantly higher mean than that of the seeds.

In modality, *Mobile* queries, characterised by keyword-based searches, demonstrate a significantly higher effectiveness than the seeds, which in DL21 and DL22 are more like natural language questions. *Voice* searches do not seem to deviate much from the performance of seed queries, which suggests that longer queries with query openers such as *"Can you tell me"* do not seem to have a significant negative impact on the overall performance of the included systems.

The textual-transformation profiles do not map to users but they serve as a reference point in our understanding of effectiveness reductions in user profiles. For example, the change of *Naturality*, which most likely moves the query to a keyword list, shows a similar improvement as was found for *Mobile* queries. Similar to the findings of Penha et al. [46], *Misspelling* has the highest impact on performance, followed by *Paraphrasing*, *Naturality*, and *Order*.

These results demonstrate that system performance varies significantly across different user profiles, although the overall impact on system ranking is generally moderate.

5 Conclusions and Future Work

We asked the following research questions:

RQ1 Are the LLM-generated query variants valid? Results demonstrate the LLM's ability to generate lexically different, but semantically similar query variants that align with required profiles. This is demonstrated through human assessments, and calculated features such as query length and readability.

RQ2 Do the generated query variants change the outcome of system evaluations? Results indicate moderate effects on system ranking, and consistently highlight inequalities among user profiles. Notably, profiles such as *Child*, *Non-native*, and *Novice* users experience significantly lower performance.

A limitation of our study is that we make a broad assumption by standardising relevance, a practice that is common in the community but not ideal, especially in contexts like ours where we distinguish users through diversified queries. What would be relevant to an *Expert* may not be to a *Novice*. Diversifying the evaluation with relevance labels originating from their issuing profiles would be an interesting avenue for future work.

Also, the studied profiles are not exhaustive; they are hypothetical examples based on properties commonly found interesting in the IR literature. They demonstrate the utility of the methodology presented here, given our limited selection of systems. Certain search engines may require evaluating different competing systems against different user profiles, potentially leading to new conclusions. Future studies can draw on domain-specific needs or taxonomies for user profiles.

The profile alignment is also subject to our profile description, which captures the literature but may not necessarily reflect actual users. This is also impacted by the human assessors' perception of those profiles (e.g., how an assessor perceives a non-native speaker), despite the provision of detailed instructions. We acknowledge that our validation does not validate that the LLM is generating queries that reflect the user profiles, but rather whether human assessors agree that they do. Although challenging, better validation methods could be explored in future work – for example, involving intended users directly in the validation, either through annotation or by comparing the generated queries to their actual queries.

Simulating users for evaluating systems, as conducted in this study, may be low-risk, but using such methods to model search behaviours – such as what users search for or engage with – can be ethically problematic and may reinforce harmful stereotypes.

The proposed method allows the observation of performance variability and therefore has the potential to enhance the experience of – or mitigate biases against – specific groups. A test collection of this kind may no longer be just a tool to rank systems; it might also help reduce disadvantages for those who struggle to search.

Acknowledgments

Marwah Alaofi is supported by a scholarship from Taibah University, Saudi Arabia. This research was also supported in part by the CAMEO PRIN 2022 Project, funded by the Italian Ministry of Education and Research (2022ZLL7MW) and the Australian Research Council (DP190101113). We thank the anonymous reviewers for their helpful feedback.

References

[1] 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*. Association for Computing Machinery, 2850–2862. doi:10.1145/3477495.3531711

[2] 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 *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. doi:10.1145/3539618.3591960

[3] 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, Tokyo, Japan, December 9-12, 2024*, Tetsuya Sakai, Emi Ishita, Hiroaki Ohshima, Faegheh Hasibi, Jiaxin Mao, and Joemon M. Jose (Eds.). ACM, 32–41. doi:10.1145/3673791.3698431

[4] Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 5311–5315. doi:10.1145/3583780.3615111

[5] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 725–728. doi:10.1145/2911451.2914671

[6] 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*. Association for Computing Machinery, 395–404. doi:10.1145/3077136.3080839

[7] Dania Bilal and Jacek Gwizdka. 2018. Children's query types and reformulations in Google search. *Information Processing & Management* 54, 6 (2018), 1022–1041. doi:10.1016/j.ipm.2018.06.008

[8] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Unsupervised Dataset Generation for Information Retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 2387–2392. doi:10.1145/3477495.3531863

[9] Timo Breuer. 2024. Data Fusion of Synthetic Query Variants With Generative Large Language Models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 274–279. doi:10.1145/3673791.3698423

[10] Chris Buckley and Janet A Walz. 1999. The TREC-8 Query Track. In *Proceeding of Text Retrieval conference*. NIST Special Publication, 500–246.

[11] Peng Chu, Anita Komlodi, and Gyöngyi Rózsa. 2015. Online search in english as a non-native language. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–9. doi:10.1002/pra2.2015.145052010040

[12] Karen Church, Barry Smyth, Paul Cotter, and Keith Bradley. 2007. Mobile Information Access: A Study of Emerging Search Behavior on the Mobile Internet. *ACM Trans. Web* 1, 1 (may 2007), 4–es. doi:10.1145/1232722.1232726

[13] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=r1xMH1BtvB

[14] Cyril W Cleverdon. 1961. *Application for grant to the National Science Foundation, Washington [for] an investigation into the performance characteristics of descriptor languages*. Technical Report. Cranfield Institute of Technology.

[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[16] Alan Cooper and Robert Reimann. 2003. About Face 2.0. *The Essentials of Interaction Design* (2003).

[17] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 282–289. doi:10.1145/290941.291009

[18] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Overview of the TREC 2021 Deep Learning Track. In *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021 (NIST Special Publication, Vol. 500-335)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf

[19] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2022. Overview of the TREC 2022 Deep Learning Track. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022 (NIST Special Publication, Vol. 500-338)*, Ian Soboroff and Angela Ellis (Eds.). National Institute of Standards and Technology (NIST). https://trec.nist.gov/pubs/trec31/papers/Overview_deep.pdf

[20] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2021. Topic Difficulty: Collection and Query Formulation Effects. *ACM Transactions on Information Systems* 40, 1, Article 19 (2021), 36 pages. doi:10.1145/3470563

[21] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. (2022). arXiv:2209.11755

[22] Tadele T. Damessie, Thao P. Nghiem, Falk Scholer, and J. Shane Culpepper. 2017. Gauging the Quality of Relevance Assessments Using Inter-Rater Agreement. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 1089–1092. doi:10.1145/3077136.3080729

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. doi:10.18653/V1/N19-1423

[24] Fernando Diaz. 2024. Pessimistic Evaluation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo, Japan) *(SIGIR-AP 2024)*. Association for Computing Machinery, New York, NY, USA, 115–124. doi:10.1145/3673791.3698428

[25] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. 2010. Query Log Analysis in the Context of Information Retrieval for Children. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Geneva, Switzerland) *(SIGIR '10)*. Association for Computing Machinery, New York, NY, USA, 847–848. doi:10.1145/1835449.1835646

[26] 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 *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi (Eds.). ACM, 39–50. doi:10.1145/3578337.3605136

[27] Fernando Ferraretto, Thiago Laitz, Roberto Lotufo, and Rodrigo Nogueira. 2023. ExaRanker: Synthetic Explanations Improve Neural Rankers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2409–2414. doi:10.1145/3539618.3592067

[28] Nicola Ferro and Mark Sanderson. 2022. How Do You Test a Test? A Multifaceted Examination of Significance Tests. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (Virtual Event, AZ, USA) *(WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 280–288. doi:10.1145/3488560.3498406

[29] Ido Guy. 2016. Searching by Talking: Analysis of Voice Queries on Mobile Web Search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 35–44. doi:10.1145/2911451.2911525

[30] Morgan Harvey and Matthew Pointon. 2017. Searching on the Go: The Effects of Fragmented Attention on Mobile Web Search Tasks. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 155–164. doi:10.1145/3077136.3080770

[31] William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In *SIGIR '94*, Bruce W. Croft and C. J. van Rijsbergen (Eds.). Springer London, London, 192–201.

[32] Maurice G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1-2 (06 1938), 81–93. doi:10.1093/biomet/30.1-2.81

[33] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. doi:10.1145/3397271.3401075

[34] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975).

[35] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[36] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational

Linguistics, Berlin, Germany, 994–1003. doi:10.18653/v1/P16-1094

[37] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.

[38] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified data wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2429–2436.

[39] Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation inInformation Retrieval using PyTerrier. In *Proceedings of ICTIR 2020*.

[40] Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. 2020. CC-News-En: A Large English News Corpus. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*. Association for Computing Machinery, 3077–3084. doi:10.1145/3340531.3412762

[41] Alistair Moffat, Falk Scholer, and Paul Thomas. 2012. Models and metrics: IR evaluation as a user process. In *Proceedings of the Seventeenth Australasian Document Computing Symposium* (Dunedin, New Zealand) *(ADCS '12)*. Association for Computing Machinery, New York, NY, USA, 47–54. doi:10.1145/2407085.2407092

[42] 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 Conference on Information and Knowledge Management*. Association for Computing Machinery, 1759–1762. doi:10.1145/2806416.2806606

[43] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Mariné. 2015. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management* 51, 5 (2015), 557–569. doi:10.1016/j.ipm.2015.05.004

[44] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[45] Rodrigo Frassetto Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document Expansion by Query Prediction. *CoRR* abs/1904.08375 (2019). arXiv:1904.08375 http://arxiv.org/abs/1904.08375

[46] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the Robustness of Retrieval Pipelines with Query Variation Generators. In *Advances in Information Retrieval - 44th European Conference on IR Research*. Springer, 397–412. doi:10.1007/978-3-030-99736-6_27

[47] Gustavo Penha, Enrico Palumbo, Maryam Aziz, Alice Wang, and Hugues Bouchard. 2023. Improving Content Retrievability in Search with Controllable Query Generation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3182–3192. doi:10.1145/3543507.3583261

[48] Maria Soledad Pera, Emiliana Murgia, Monica Landoni, Theo Huibers, and Mohammad Aliannejadi. 2023. Where a Little Change Makes a Big Difference: A Preliminary Exploration of Children's Queries. In *Advances in Information Retrieval*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer Nature Switzerland, Cham, 522–533.

[49] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.

[50] Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. *CoRR* abs/2101.05667 (2021). arXiv:2101.05667 https://arxiv.org/abs/2101.05667

[51] Kun Ran, Marwah Alaofi, Mark Sanderson, and Damiano Spina. 2025. Two Heads Are Better Than One: Improving Search Effectiveness Through LLM Generated Query Variants. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval* (Melbourne, Australia) *(CHIIR '25)*. Association for Computing Machinery, New York, NY, USA.

[52] Mylene Sanchiz, Aline Chevalier, and Franck Amadieu. 2017. How do older and young adults start searching for information? Impact of age, domain knowledge and problem complexity on the different steps of information searching. *Computers in Human Behavior* 72 (2017), 67–78. doi:10.1016/j.chb.2017.02.038

[53] Karen Spärck Jones and R. Graham Bates. 1977. *Report on a Design Study for the "Ideal" Information Retrieval Test Collection*. British Library Research and Development Report No. 5428. University of Cambridge.

[54] Karen Spärck-Jones and C. J. van Rijsbergen. 1975. *Report on the Need for and Provision of an "Ideal" Information Retrieval Test Collection*. Technical Report British Library Research and Development Report 5266. Computer Laboratory, University of Cambridge.

[55] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2010. Potential for Personalization. *ACM Trans. Comput.-Hum. Interact.* 17, 1, Article 4 (apr 2010), 31 pages. doi:10.1145/1721831.1721835

[56] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large Language Models can Accurately Predict Searcher Preferences. 11 pages. doi:10.1145/3626772.3657707

[57] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (1949), 99–114. http://www.jstor.org/stable/3001913

[58] Ellen M. Voorhees. 2001. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 74–82. doi:10.1145/383952.383963

[59] Ellen M. Voorhees. 2001. The Philosophy of Information Retrieval Evaluation. In *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers (Lecture Notes in Computer Science, Vol. 2406)*, Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck (Eds.). Springer, 355–370. doi:10.1007/3-540-45691-0_34

[60] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too Many Relevants: Whither Cranfield Test Collections?. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 2970–2980. doi:10.1145/3477495.3531728

[61] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. Association for Computing Machinery, New York, NY, USA, 132–141. doi:10.1145/1498759.1498819

[62] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. https://openreview.net/forum?id=zeFrfgyZln

[63] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, and Saurabh Tiwary. 2019. Generic Intent Representation in Web Search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, 65–74. doi:10.1145/3331184.3331198