Generating Effective Health-Related Queries for Promoting Reliable Search Results

Xiana Carrera
Universidade de Santiago de Compostela
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Santiago de Compostela, Spain
xiana.carrera@rai.usc.es

Marcos Fernández-Pichel
Universidade de Santiago de Compostela
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Santiago de Compostela, Spain
marcosfernandez.pichel@usc.es

David E. Losada
Universidade de Santiago de Compostela
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Santiago de Compostela, Spain
david.losada@usc.es

Abstract

Misinformation on the Internet poses significant risks to users seeking health information. This paper addresses the challenge of generating effective health-related queries to promote reliable search results. We propose a method leveraging Large Language Models to generate synthetic narratives that guide the creation of alternative queries. These queries are designed to retrieve more helpful and fewer harmful documents compared to those retrieved by the original user queries. We evaluate the effectiveness of these queries using classic and neural retrieval models across multiple datasets, demonstrating promising improvements in retrieving reputable content.

CCS Concepts

• **Information systems → Query reformulation**.

Keywords

Query Variants, Large Language Models, Health Misinformation

ACM Reference Format:
Xiana Carrera, Marcos Fernández-Pichel, and David E. Losada. 2025. Generating Effective Health-Related Queries for Promoting Reliable Search Results. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730202

1 Introduction

Search engines stand as a widespread tool for seeking health information online [9, 32], but effective information retrieval in the medical domain is still an open challenge [24]. The pervasiveness of misinformation on the Internet [18] poses a risk to users, who can be influenced to make incorrect and harmful decisions [19].

Although significant advancements have been made in neural models for retrieving reliable content [20, 21, 27], the search results of certain queries leave room for improvement. For example, the ten best automatic systems developed under the latest TREC Health Misinformation Track retrieve too many harmful webpages [4]. This suggests that we still need to reduce poor content from search results.

The users' ability to formulate effective health-related queries is limited, and there is potential for designing automated methods to support users in reformulating their health-related information needs. Our results suggest that replacing the original user queries with effective reformulations (e.g., "mental performance factors" by "scientific insight on cognitive abilities") leads to retrieval results that promote reliable and correct information over misinformation.

We show that guiding query generation with TREC-style narratives effectively produces solid health queries. However, web users are known to be reluctant to establish verbose interactions with search engines [14] and, thus, access to an explicit narrative describing the health information need cannot be granted. To address this limitation, we leverage Large Language Models (LLMs) to generate synthetic narratives that are subsequently prompted to generate candidate queries. This requires the careful design of a prompt oriented to generate useful narratives and a prompt oriented to produce effective queries. The first prompt forces the LLM to foresee the user information need, to anticipate criteria for helpful and harmful documents, and to provide additional context. The second prompt instructs the LLM to generate alternative queries, and we test here different prompt templates to understand the effect of system role, narrative, chain of thought, and other elements.

The evaluation of these alternative queries is done with two search models and four different datasets. We found that narrative descriptions are crucial for generating strong queries, and the resulting approach can improve by more than 50% in terms of ranking similarity to ideal rankings. Specifically, the contributions of this paper are:

- We show that query generation using real or synthetic narratives is effective in health search tasks. In particular, synthetic narratives seem particularly promising for demoting harmful results in lexical retrieval models.
- We design and evaluate different prompt templates instructing the LLMs to produce strong queries that guide the search toward reputed content.
- We show that LLMs can automatically generate precise narratives given short (title-like) health topics. The improvements found are generally consistent across multiple datasets and, in particular, they hold in datasets whose search topics do not have

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730202

a narrative field. Thus, successful narrative generation cannot be attributed to data leakage issues.

2 Related Work

Bacciu et al. [2] framed the query recommendation problem as a generative task supported by LLMs and evaluated different few-shot alternatives (e.g., exploiting query logs) to produce recommended queries. Inspired by ensembling methods, Dhole and Agichtein [7] proposed a prompting technique that leverages paraphrases of zero-shot instructions and generates multiple sets of keywords for query reformulation. Wang et al. [30] evaluated fine-tuning (T5) and prompting (FlanT5) methods for query reformulation, experimenting with some alternatives that inject additional context for the query and exploit pseudo-relevance feedback. Motivated by aligning LLMs with human values, PURE [34] is a mechanism that, in the context of conversational agents, converts risky queries into harmless queries before feeding them into LLMs. Alaofi et al. [1] produced query variants from information need statements using a one-shot approach with GPT3.5, showing promise for building test collections. Recently, Ran et al. [22] investigated the effect of fusing search results of LLM-generated query variants and Li et al. [15] showed that query expansion can improve the generalization of strong cross-encoder rankers. A recent study on generative query and document expansions [31] showed a strong negative correlation between the performance of the retrieval model and the gains obtained from expansion.

Claveau [6] exploited GPT2 to generate documents from the original queries and used these synthetic documents for query expansion. Query2Doc [28] improved standard sparse and dense search models by generating pseudo-documents (using few-shot prompts) and then expanding the queries with these pseudo-documents. In a similar vein, Jagerman et al. [11] compared different prompts to expand queries using FLAN-T5 or FLAN-UL2. In [16], the authors generated a (pseudo) relevance feedback (RF) model that is combined with the original query's model. The RF model is obtained by prompting the initial query to an LLM and asking to generate multiple related elements, such as keywords, entities, or facts.

Thomas et al. [26] tested several prompt templates to generate query-document relevance labels. In their evaluation of LLMs as webpage raters, these authors also exploited different parts of TREC topics, including the narrative field (from TREC Robust). Human-produced narratives are available in standard IR benchmarks, but end users hardly provide these lengthy descriptions. Thus, we design here specific prompts to generate synthetic narratives.

None of the previous studies specifically focused on health misinformation. We contribute to the existing literature by designing methods that not only generate superior queries but also steer the results toward reputable content. The construction of synthetic narratives, which inform about potential helpful and harmful results, represents an innovation with respect to the research done in the literature.

3 Methodology

For a given user query q, we wish to generate n alternative queries, $q'_1, \ldots, q'_n$, that are prone to retrieving more helpful documents and fewer harmful documents compared to those retrieved by q. To that end, we perform an intermediate step in which we create a synthetic narrative sn by feeding q into an LLM with a custom prompt, denoted as $prompt_{narr}$: $sn = LLM(prompt_{narr}(q))$.

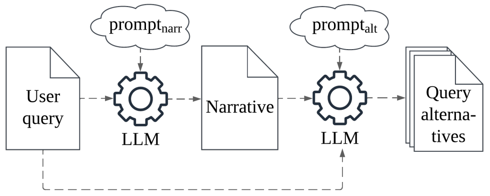

Figure 1: Two-stage process to generate query variants

Role (R)

> You are a search engineer trying to improve the relevance, correctness and credibility of search results for health-related queries.
> Given a query, you must provide a list of *n* alternative queries that express the same information need as the original one, but that are phrased in such a way that they are more likely to retrieve relevant, correct and credible documents.

Narrative (N)

> **Query:** A person has typed [*keyword_query*] into a search engine.
> They were looking for: *[query_narrative]*

Chain of thought (C)

> **Instructions:** Let's think step by step: Consider the underlying intent of the search.
> Measure how prone the original query is to retrieve useful documents.
> Measure how prone the original query is to retrieve supportive documents for the correct treatment of the query's question.
> Measure how prone the original query is to retrieve credible documents.
> Consider the aspects above and the relative importance of each, and produce an array of variant queries without providing any reasoning. Example: ["query variant 1", "query variant 2", ...]

Figure 2: Template used for generating alternative queries ($prompt_{altq}$). CoT prompt ($C = 1$) is the text shaded in green. CoT prompt ($C = 2$) includes both textual parts (green & blue).

To experiment with a variety of narrative styles, we made preliminary tests with the following versions of $prompt_{narr}$: i) a concise prompt that asks to generate a single paragraph narrative with the only indications of "detailing the information need" and "describing the characteristics of helpful and harmful documents", ii) a TREC-style prompt, which is analogous to the basic prompt, but also requests to use the "standard TREC format for narratives", and, iii) a more elaborated prompt whose goal is to produce narratives with a specific structure, voice, tone, language style, and intention of obtaining objective, neutral, informative, and factual results. The second variant tended to produce the most explanatory narrative descriptions and, thus, we adopted it for the subsequent experiments.

Next, we take the resulting synthetic narrative sn and the original user query q and feed them again into the LLM, using a prompt, $prompt_{altq}$, that asks for the generation of alternative queries taking sn into account: $q'_1, \ldots, q'_n = LLM(prompt_{altq}(q, sn))$.

The configuration with two independent steps allows us to generate alternative queries with the synthetic narrative but also to test query generation with real narratives from the TREC topics (if available). To do so, we just need to replace the synthetic narrative sn by a real narrative in the equation above. This will help compare the relative performance of real and generated narratives.

Inspired by Thomas et al. [26], we set several template variants for $prompt_{altq}$ based on two binary parameters (R for the presence or absence of system role and N for the presence or absence of the narrative), and a chain of thought (CoT) variable C ($C = 0$ means no CoT text, $C = 1$ is a basic CoT instruction [35], and $C = 2$ is a more elaborated CoT that asks the LLM to reason about the credibility and correctness of the documents potentially retrieved by the query). Figure 2 shows this prompt template and Figure 1 the two-stage process for generating queries.

4 Experimental Setup

We evaluated the query generation process using four datasets: three of them from the TREC Health Misinformation (HM) track, editions 2020, 2021 and 2022 [3–5], and one from the IR track at the CLEF eHealth Evaluation Lab 2016 (CLEF IR) [13]. These datasets supply large web corpora crawled from the web (C4 or ClueWeb12 B13), a set of health-related search topics (50 topics for each TREC HM collection and 300 topics for CLEF IR), and query-document judgments. The TREC HM search topics contain a keyword query (e.g., "ibuprofen COVID-19"), a description field in the form of a question (e.g., "Can ibuprofen worsen COVID-19?"), and a narrative field (paragraph explaining the information need). Keyword queries are representative of the type of searches submitted by web users and, thus, we assume that this is the only information available to the query generation process. The CLEF IR collection (ad-hoc search task) contains only keyword queries for each search topic.

In the TREC HM collections, judgments were done at three levels (usefulness –i.e., on topic–, correctness, and credibility) and then converted into graded relevance labels, where the most helpful documents are those useful, correct, and credible while the most harmful documents are those useful, incorrect, and credible. We follow the standard evaluation of the TREC HM track based on computing the compatibility of search rankings with helpful and harmful results. A good ranking is helpful and not harmful and, thus, it should have high compatibility helpful (maximum similarity to an ideal ranking where the most helpful documents are at the top) and low compatibility harmful (minimum similarity to a ranking where the most harmful documents are at the top).[1]

The CLEF IR collection contains assessments on topical relevance, understandability, and trustworthiness (range [0-100]). We uniformly mapped the trustworthiness scores to integers in the range $[-2, 2]$, and the resulting preference order was used to compute the compatibility helpful and harmful metrics.

For narrative and query generation, we employed OpenAI's GPT-4, with a temperature of 0.2 and a frequency penalty of 0. We generated five alternative queries for each original query and report here the average performance obtained with these five queries. Further improvements (e.g., fusing the rankings of the query variants or making query-dependent selection of the number of variants) were left to future work. The generated queries, narratives, prompts and code needed to run our experiments are publicly available.[2] We also experimented with LLaMA3[3] finding similar trends (due to space constraints, the experimental report contains only GPT-4's results, but we offer below some comments about LLaMA3's performance).

4.1 Search Models

To evaluate the effect of query generation on both traditional and neural models, we first conducted a comparison of retrieval baselines in the context of BEIR [25]. Specifically, we compared BM25 [23], sparse models (SPARTA [36], SPLADE [8], DocT5query [17]), dense models (DPR [12], ANCE [33], TAS-B [10]) and multiple cross-attentional models.[4] Following this comparison, we selected BM25 as a traditional (weaker but lighter) retrieval approach and the cross-attentional MiniLM-L-12-v2 model [29] (re-ranking the top 100 BM25 results), which had the most solid search performance with the original queries (but it is computationally demanding).

5 Results

The results (see Table 1) show the potential of the query generation approach.[5] For each model, the table compares the performance of the original queries (first row in each block) against the performance of i) queries generated with no narrative, ii) queries generated with the real narrative from the collection (if available), and iii) queries generated with synthetic narratives. For both BM25 and miniLM-12, searching with the new queries produced significant advantages. In general, the effects were stronger with queries produced by prompting the LLM with the real narratives. In any case, the variants with synthetic narratives also led to promising results. The column Help-Harm reports the difference between compatibility helpful and compatibility harmful, which was an official measure in the TREC HM tracks and represents an aggregation of both criteria.[6] For the strongest model, miniLM-12, nearly all query generation variants led to improvements in Help-Harm. For BM25, the effect of the new queries is more modest, mainly because they tended to reduce the retrieval of harmful documents at the cost of retrieving fewer helpful ones. The system role (R) does not seem crucial; the experiments do not conclusively demonstrate any significant benefit (or drawback) from integrating role instructions into the template for query generation. Regarding equivalent experiments with LLaMA3, we observed similar trends and, for example, in all collections miniLM-12's Comp. Help-Harm improved with queries produced from synthetic narratives (e.g., R=1, N=1, C=1 led to .179 vs baseline's .148 in TREC HM 2020).

To further analyze the results, Figure 3 plots the effect of the synthetic variant (R=1, N=1, C=1) with respect to the original performance (X axis: Comp. Helpful, Y axis: Comp. Harmful). For each model and collection, we draw an arrow from the point representing the baseline performance (original queries) to the point representing the synthetic variant. For BM25, in all collections, the synthetic variants produced better retrieval of harmful results (i.e., lower comp. harmful) but poorer retrieval of helpful results. For miniLM-12, all synthetic variants led to better compatibility helpful scores, and in two collections (TREC HM 2021, CLEF IR 2016) they also yielded better (lower) compatibility harmful. Overall, the effect of the synthetic queries is positive, but we need to further study

[1] Rank similarities are computed using Rank-Biased Overlap (RBO).

[2] https://github.com/xianacarrera/Generating-Effective-Health-Queries

[3] Llama 3.1, instruct version, 8B params, quantization Q8_0: https://ollama.com/library/llama3.1:8b-instruct-q8_0

[4] Electra-base, MiniLM-L-4-v2, MiniLM-L-6-v2, MiniLM-L-12-v2, TinyBERT-L-2-v2, TinyBERT-L-4, TinyBERT-L-6 & MonoT5 (base, base-med & large).

[5] Due to space constraints, we only report the results achieved with C=1. In any case, we found no significant differences among the three chain-of-thought variants.

[6] This is a derived measure from compatibility helpful and compatibility harmful and, thus, the table only reports significance tests for the original harmful and helpful metrics.

Table 1: Compatibility Helpful, Harmful, and Helpful-Harmful. The R, N, and C columns reflect the prompt configurations for generating queries. For each block, collection, and measure, the best performance is bolded. The helpful and harmful scores are marked with * when the improvement over the baseline is statistically significant (Wilcoxon test, $\alpha = .05$)

Model	Queries			TREC HM 2020			TREC HM 2021			TREC HM 2022			CLEF IR 2016		
	R	**N**	**C**	Help	Harm	Help - Harm	Help	Harm	Help - Harm	Help	Harm	Help - Harm	Help	Harm	Help - Harm
BM25	original qs			.214	.047	.167	**.129**	.145	-.016	**.173**	.144	.029	**.101**	.272	-.172
no narrative	0	0	1	.236	.058	.178	.109	.108	.001	.100	.077*	.023	.086	.179*	-.093
	1	0	1	.232	.058	.174	.105	.112	-.007	.112	.077*	.035	.085	.182*	-.097
real narrative	0	1	1	.256	.052	.204	.103	**.098***	.005	.086	.080*	.006	—	—	—
	1	1	1	**.259**	.048	**.211**	.118	.100*	**.018**	.100	.085*	.015	—	—	—
synthetic narrative	0	1	1	.191	.046	.145	.098	.102	-.004	.072	**.049***	.023	.089	**.155***	**-.066**
	1	1	1	.197	**.045**	.152	.106	.103	.003	.095	.056*	**.039**	.087	.157*	-.070
MiniLM-12	original qs			.226	**.078**	.148	.132	.136	-.004	.179	**.131**	.048	.095	.211	-.116
no narrative	0	0	1	.289*	.089	.200	.135	**.129**	.006	.190	.140	.050	.096	.198*	-.102
	1	0	1	.288*	.092	.196	.137	.134	.003	.191*	.138	.053	.096	.200*	-.104
real narrative	0	1	1	.307*	.092	.215	.142*	.134	**.008**	**.199***	.137	**.062**	—	—	—
	1	1	1	**.312***	.089	**.223**	**.144**	.138	.006	.195*	.144	.051	—	—	—
synthetic narrative	0	1	1	.274*	.087	.187	.136	.132	.004	.188	.134	.054	**.100***	**.196**	**-.096**
	1	1	1	.274*	.085	.189	.138	.134	.004	.189*	.135	.054	.099*	.196*	-.097

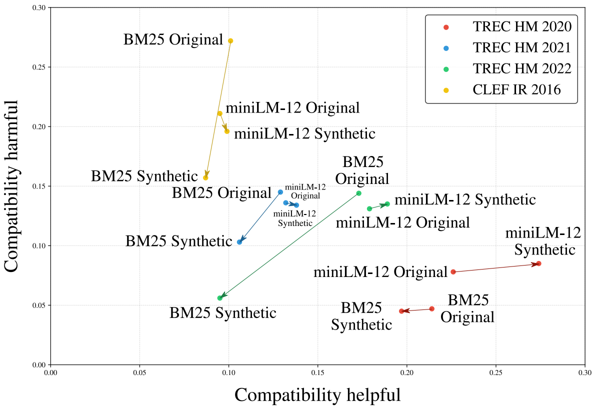

Figure 3: Effect on performance of the synthetic variants. The optimal point is at the bottom-right corner (compatibility helpful equal to 1 and compatibility harmful equal to 0).

the situations in which fewer helpful or more harmful documents are retrieved. Related to this, the fusion of rankings from different variants and a per-query selection of the number of variants are promising avenues to trade between helpfulness and harmfulness.

We manually inspected the query variants that produced the highest impact on effectiveness. From this qualitative analysis, we observed that strong improvements in the retrieval of helpful documents were often associated with queries that clarified the intent of the search. For example, “Hib vaccine COVID-19” transformed to “Does the hib vaccine provide protection against COVID-19?”, “Inhalers COVID-19” to “Effectiveness of inhalers for COVID-19 symptoms”, or “birth control pill ovarian cysts treatment” to “How do birth control pill treat ovarian cysts?”. On the other hand, the variants that were more effective at demoting harmful documents often introduced search terms and phrases about scientific evidence or safety. For example, “High temperatures and humidity COVID-19” transformed to “Scientific studies on climate factors and COVID-19 transmission”, “Breast milk COVID-19” to “Is it safe to breastfeed if I have COVID-19?”, or “baking soda cancer” to “baking soda cancer prevention evidence from health organizations”. The variants that clearly improved helpfulness and harmfulness usually contributed by clarifying the intent and guiding the search toward reputed results. For example, “tylenol osteoarthritis” to “Tylenol dosage and side effects for osteoarthritis” or “magnetic wrist straps arthritis” to “scientific studies on magnetic wristbands for arthritis treatment”. But there is no free lunch, as some apparently good reformulations (e.g., “Breast milk COVID-19” to “Breastfeeding guidelines for mothers with COVID-19” or “vitamin d asthma attacks” to “Scientific studies on vitamin D and asthma prevention”) led to poorer results. This suggests that further research is still required to understand the conditions under which a given query effectively retrieves reliable content.

6 Conclusions and Future Work

This paper proposed and evaluated a two-stage process that generates query variants for health queries. Our results suggest that TREC-style narratives play a crucial role in promoting helpful search results and demoting harmful ones, and we have shown that LLMs can be exploited to produce synthetic narratives. These narratives can effectively guide the creation of alternative queries. Although LLMs are prone to hallucination and can produce wrong completions, our results suggest that the resulting queries effectively retrieve reputable documents.

In our future work, we intend to study the features that make a query effective and design strategies for combining results from multiple queries. We will also keep in mind the computational load of the approach since the incorporation of LLMs at query time results in delays in the search process. On the other hand, automatic query suggestions shown to the users can also be highly informative and educational, helping them understand how to construct their own queries effectively.

Acknowledgments

Funded by MICIU/AEI/10.13039/501100011033 (PID2022-137061OB-C22, supported by ERDF) and Xunta de Galicia-Consellería de Cultura, Educación, Formación Profesional e Universidades (ED431G 2023/04, ED431C 2022/19, supported by ERDF).

References

[1] 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* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1869–1873. https://doi.org/10.1145/3539618.3591960

[2] Andrea Bacciu, Enrico Palumbo, Andreas Damianou, Nicola Tonellotto, and Fabrizio Silvestri. 2024. Generating Query Recommendations via LLMs. arXiv:2405.19749 [cs.IR] https://arxiv.org/abs/2405.19749

[3] Charles L. A. Clarke, Maria Maistro, Saira Rizvi, Mark D. Smucker, and Guido Zuccon. 2020. Overview of the TREC 2020 Health Misinformation Track.

[4] Charles L. A. Clarke, Maria Maistro, Mahsa Seifikar, and Mark D. Smucker. 2022. Overview of the TREC 2022 Health Misinformation Track (Notebook).

[5] Charles L. A. Clarke, Maria Maistro, and Mark D. Smucker. 2021. Overview of the TREC 2021 Health Misinformation Track.

[6] Vincent Claveau. 2022. Neural text generation for query expansion in information retrieval. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Melbourne, VIC, Australia) *(WI-IAT '21)*. Association for Computing Machinery, New York, NY, USA, 202–209. https://doi.org/10.1145/3486622.3493957

[7] Kaustubh D. Dhole and Eugene Agichtein. 2024. *GenQREnsemble: Zero-Shot LLM Ensemble Prompting for Generative Query Reformulation.* Springer Nature Switzerland, 326–335. https://doi.org/10.1007/978-3-031-56063-7_24

[8] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2288–2292.

[9] William Hersh. 2024. Search still matters: information retrieval in the era of generative AI. *Journal of the American Medical Informatics Association* (2024), ocae014.

[10] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 113–122.

[11] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. arXiv:2305.03653 [cs.IR] https://arxiv.org/abs/2305.03653

[12] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).

[13] Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurélie Névéol, João Palotti, and Guido Zuccon. 2016. Overview of the CLEF eHealth Evaluation Lab 2016.

[14] Gondy Leroy. 2009. Persuading consumers to form precise search engine queries. In *AMIA Annual Symposium Proceedings*, Vol. 2009. 354.

[15] Minghan Li, Honglei Zhuang, Kai Hui, Zhen Qin, Jimmy Lin, Rolf Jagerman, Xuanhui Wang, and Michael Bendersky. 2024. Can Query Expansion Improve Generalization of Strong Cross-Encoder Rankers?. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2321–2326. https://doi.org/10.1145/3626772.3657979

[16] Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative Relevance Feedback with Large Language Models. 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, 2026–2031. https://doi.org/10.1145/3539618.3591992

[17] Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery. https://cs.uwaterloo.ca/~jimmylin/publications/Nogueira_Lin_2019_docTTTTTquery-latest.pdf

[18] Wei Peng, Sue Lim, and Jingbo Meng. 2023. Persuasive strategies in online health misinformation: a systematic review. *Information, Communication & Society* 26, 11 (2023), 2131–2148.

[19] Frances A. Pogacar, Amira Ghenai, Mark D. Smucker, and Charles L.A. Clarke. 2017. The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* (Amsterdam, The Netherlands) *(ICTIR '17)*. Association for Computing Machinery, New York, NY, USA, 209–216. https://doi.org/10.1145/3121050.3121074

[20] Ronak Pradeep and Jimmy Lin. 2024. *Towards Automated End-to-End Health Misinformation Free Search with a Large Language Model.* 78–86. https://doi.org/10.1007/978-3-031-56066-8_9

[21] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2066–2070.

[22] 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.

[23] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.

[24] Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical information retrieval: A literature review. *Journal of Healthcare Informatics Research* (2024), 1–40.

[25] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663* (2021).

[26] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. Large language models can accurately predict searcher preferences. arXiv:2309.10621 [cs.IR] https://arxiv.org/abs/2309.10621

[27] Rishabh Upadhyay, Arian Askari, Gabriella Pasi, and Marco Viviani. 2024. *Beyond Topicality: Including Multidimensional Relevance in Cross-encoder Re-ranking: The Health Misinformation Case Study.* 262–277. https://doi.org/10.1007/978-3-031-56027-9_16

[28] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9414–9423. https://doi.org/10.18653/v1/2023.emnlp-main.585

[29] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33 (2020), 5776–5788.

[30] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023. Generative Query Reformulation for Effective Adhoc Search. In *The First Workshop on Generative Information Retrieval, Gen-IR@SIGIR2023.*

[31] Orion Weller, Kyle Lo, David Wadden, Dawn Lawrie, Benjamin Van Durme, Arman Cohan, and Luca Soldaini. 2024. When do Generative Query and Document Expansions Fail? A Comprehensive Study Across Methods, Retrievers, and Datasets. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1987–2003. https://aclanthology.org/2024.findings-eacl.134

[32] Bangan Wu, Qianqian Ben Liu, Xitong Guo, and Chen Yang. 2024. Investigating patients' adoption of online medical advice. *Decision Support Systems* 176 (2024), 114050.

[33] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[34] Wenjin Yao, Yidong Wang, Zhuohao Yu, Rui Xie, Shikun Zhang, and Wei Ye. 2024. PURE: Aligning LLM via Pluggable Query Reformulation for Enhanced Helpfulness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8721–8744. https://doi.org/10.18653/v1/2024.findings-emnlp.509

[35] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic Chain of Thought Prompting in Large Language Models. arXiv:2210.03493 [cs.CL] https://arxiv.org/abs/2210.03493

[36] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2020. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. *arXiv preprint arXiv:2009.13013* (2020).