Query Harmfulness Prediction (QHP): A New Challenge for Safer Retrieval Systems

Xiana Carrera, Marcos Fernández-Pichel(✉), and David E. Losada

Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
xiana.carrera@rai.usc.es, {marcosfernandez.pichel,david.losada}@usc.es

Abstract. This paper introduces Query Harmfulness Prediction (QHP), a novel extension of Query Performance Prediction that focuses on predicting the potential harmfulness of search results. While traditional QPP methods predict standard retrieval metrics, QHP addresses the growing need for safer information retrieval by anticipating when queries might return harmful but topically relevant results. We investigate three families of predictors: classical pre-retrieval QPP methods, LLM-based strategies leveraging signals such as controversy and misinformation, and a query quality classifier adapted from prior work. Using datasets from TREC and CLEF campaigns, we evaluate these approaches with compatibility harmful as the target measure. Our results show that while traditional QPP predictors capture limited signals of harmfulness, LLM-based methods consistently provide stronger correlations, especially on high-risk queries. These findings establish QHP as a timely research direction for developing safer retrieval systems that balance relevance with user safety.

Keywords: Query Harmfulness Prediction · Query Performance Prediction · Pre-Retrieval Prediction · Misinformation · Compatibility Harmful

1 Introduction

Query Performance Prediction (QPP) has a well-established trajectory in Information Retrieval (IR) [4,9,21,36], but it has mainly focused on predicting standard retrieval metrics such as Average Precision (AP) or normalized Discounted Cumulative Gain (nDCG). In everyday searches, we argue that supporting end users requires not only predicting retrieval effectiveness, but also anticipating adverse situations where search results might expose users to harmful material. For instance, in consumer search tasks, users may encounter on-topic documents containing incorrect medical advice or other forms of misinformation that, while topically relevant, could lead to adverse consequences.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
R. Campos et al. (Eds.): ECIR 2026, LNCS 16484, pp. 55–69, 2026.
https://doi.org/10.1007/978-3-032-21300-6_4

To address this gap, we propose a novel extension of QPP, which we term **Query Harmfulness Prediction (QHP)**. This new variant focuses specifically on predicting metrics that account for the presence of relevant yet harmful documents within search results. By extending beyond traditional relevance-based evaluation, QHP addresses the growing imperative to ensure not only effective but also safe information retrieval, particularly in domains where misinformation poses significant risks to user well-being. This paradigm shift reflects the evolution of search systems from mere information retrieval tools to responsible information mediators that balance relevance with user safety.

The development of effective QHP methods would enable a range of user-centered interventions. For instance, search systems could proactively alert end users about the potential risks associated with the retrieval results, recommend alternative query formulations to mitigate exposure to harmful content, or redirect high-risk queries to specialized retrieval systems (e.g., equipped with enhanced safety mechanisms and content filtering capabilities). These risk-aware search strategies would prove crucial in high-stakes domains such as healthcare, where misleading or inaccurate information can be highly problematic [27].

In this work, we aim to foster research into this emerging challenge by proposing and systematically comparing an initial set of QHP methods. We focus exclusively on pre-retrieval variants [19,20], given their simplicity and low computational cost, which makes them attractive for real-time applications. We adopt the notion of harmfulness as defined in prominent IR evaluation campaigns. Specifically, under the TREC Health Misinformation (HM) tracks [5–7], harmful documents are characterized as search results that are topically relevant but contain incorrect or misleading information.[1] This definition provides a standardized foundation for our investigation and ensures the QHP predictors align with established evaluation protocols in the field.

Since harmful results are a subset of relevant results, it might be the case that traditional QPP methods may inherently capture some signal regarding the retrieval of harmful documents. Consequently, our first research goal will be to assess the capacity of existing pre-retrieval QPP methods to predict the presence of harmful results within search results. Such evaluation will establish baseline performance levels and determine the transferability of QPP predictors to the realm of QHP.

A second class of QHP methods that we evaluate here is based on Large Language Models (LLMs). The rapid advancement of LLMs has positioned them as powerful tools for a wide range of text analysis tasks [2,3]. We therefore explore their application to QHP through a series of prompting experiments, including variants with chain-of-thought reasoning that derive new QHP methods. The associated instructions, for example, guide the LLM to first consider individual aspects, such as ambiguity or polarization, before providing a QHP estimate. Finally, we also evaluate the transferability of existing query quality classifiers to the QHP task [14]. Although these supervised learning tools were not specifically

[1] Following the graded scheme established by the TREC HM tracks, the most harmful documents are relevant and incorrect, but they look credible.

designed for estimating harmfulness, we think it is intriguing to assess their applicability and potential adaptation to QHP.

Following standard practice in the evaluation of systems that promote reliable and correct information over misinformation [5–7], we adopt Compatibility Harmful as the target effectiveness measure for QHP. A good ranking of results must do no harm and, thus, it should have a low score of compatibility harmful, defined as the minimum similarity to a ranking where the most harmful documents are at the top.[2] In the future, QHP methods could be further extended or adapted to predict other harmful-related measures. However, in the current study, we center our efforts on predicting this well-known measure of harmfulness.

Our findings suggest that traditional pre-retrieval QPP methods are able to predict harmfulness up to some extent. However, LLM-based predictors seem more promising. This is an expected outcome since these strategies are steered towards crucial aspects such as controversy or misinformation. Summing up, this paper makes the following key contributions:

- We introduce Query Harmfulness Prediction as a novel and timely research challenge in the field of IR, aiming to assess not just relevance but the potential harmfulness of search results.
- We evaluate traditional pre-retrieval QPP methods in terms of their ability to predict Compatibility Harmful, providing initial insights into their applicability to this new task.
- We propose a set of novel LLM-based strategies for harmfulness prediction, leveraging different prompting techniques to capture deeper signals of risk in user queries.
- We study the transferability of query quality classifiers to the realm of QHP.

2 Related Work

Query Performance Prediction [4,9,21,36] is a research area in IR that aims to estimate the effectiveness of a retrieval system's response to a given query, without access to explicit relevance judgments. QPP methods are broadly categorized into pre-retrieval predictors [19,20], which work with query characteristics and collection statistics (e.g., query length or inverse document frequency distributions), and post-retrieval predictors [24], which leverage information extracted from the ranked list of retrieved documents. In recent years, there has been intense activity in exploiting advanced neural architectures for QPP [13]. This has led to QPP methods based on end-to-end neural predictors [34], synthetic relevance judgments [25], query variants [10], contextualized embeddings [1,23], query difficulty weights [29], rich query representations [11], and dimension importance estimators [12], just to name a few. However, the target measure for prediction has always been some standard retrieval metric (for

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

example, nDCG, AP, RR, P@k or R@k). We argue here that we need to instigate research on new methods that take into account other aspects of retrieval quality beyond relevance.

In the areas of Consumer Health Search [17,37] and Health Misinformation [5–7], there has been considerable concern regarding the presence of search results that are relevant but harmful. As a result, system evaluation in these contexts has incorporated factors such as correctness, credibility, understandability, or trustworthiness. For example, the TREC HM campaigns have adopted compatibility measures [8] that estimate the Rank-Biased Overlap between the system's ranking and ideal rankings: to promote safe IR, the search results must be relevant but, additionally, the ranking must have low compatibility harmful (i.e., minimum similarity to a ranking where the most harmful documents are at the top). Multiple research groups have approached this type of experimental challenge by, for example, implementing multistage ranking architectures that incorporate different neural ingredients [16,28], re-ranking the original ranked lists with passages generated by RAG [32], or integrating additional sources of knowledge such as graphs [26]. Here, we adopt compatibility harmful as a reference metric to derive new QHP methods, and we hope that our initial research in this area will encourage broader engagement from the research community, both to design alternative QHP methods and to explore the prediction of other evaluation measures.

3 QHP Methods

First, we briefly discuss some classical QPP predictors. As argued above, harmful results are a subset of relevant documents and, thus, QPP estimates might capture some signal for the QHP task. Next, we propose some LLM-based predictors for QHP and, finally, we introduce a query quality classifier, originally built for another purpose, but adapted and evaluated here for the QHP task.

3.1 QPP Pre-retrieval Predictors

We consider the following QPP predictors:

- *Average Inverse Document Frequency* (avg IDF) [4], the average IDF score of the query terms: $avgIDF(q) = \frac{1}{|q|} \sum_{t \in q} \log(\frac{N}{N_t})$.
- *Maximum Inverse Document Frequency* (max IDF) [4], the maximum IDF score of the query terms: $maxIDF(q) = \max_{t \in q} \log(\frac{N}{N_t})$.
- *Average Collection Query Similarity* (avg SCQ) [35], the average SCQ score over the query terms, where SCQ is defined as: $\text{SCQ}(t) = (1 + \log(\text{f}(t, C))) \cdot \log(\frac{N}{N_t})$.
- *Maximum Collection Query Similarity* (max SCQ) [35], the maximum SCQ score over the query terms: $maxSCQ(q) = \max_{t \in q} SCQ(t)$.
- *Average Inverse Collection Term Frequency* (avg ICTF): $\text{avg ICTF}(q) = \frac{1}{|q|} \sum_{t \in q} \log \frac{N}{\text{f}(t,C)}$.

– *Simplified Clarity Score* (SCS) [20]: $\text{SCS}(q) = \log \frac{1}{|q|} + \text{avg ICTF}(q)$.

where N is the number of documents in the corpus C, the query q is a sequence of $|q|$ terms, N_t is the number of documents in the corpus that contain the term t, and $f(t, C)$ is the number of occurrences of t in C.

3.2 LLM-Based Pre-retrieval Predictors

We operate under the premise that controversy constitutes a key factor in determining the potential presence of harmful search results. Our hypothesis is that queries related to non-controversial topics are likely to yield few harmful documents, whereas queries concerning controversial subjects (e.g., vaccine efficacy or unconventional treatments) are inherently more problematic. Therefore, we provide the LLM with initial instructions regarding its role, as well as guidance on how to assess the degree of controversy (see Fig. 1). This prompt adheres to design principles consistent with those proposed for annotating relevance with LLMs [31]. Our set of factors should be understood as an initial selection, and a comprehensive exploration of prompts and instruction strategies for estimating QHP is left as future work. Specifically, we consider the following variants:

– Controversy (global): produces a single controversy score for the query (orange text in Fig. 1).
– Controversy (CoT): forces the LLM to produce individual estimates for a series of factors influencing controversy plus a final estimate of controversy (blue text in Fig. 1). The rationale of this method is that elements such as ambiguity, polarization, misinformation, and conflicting information may serve as critical signals to guide an LLM in the QHP task.

In a second class of variants, we also ask the LLM to estimate each individual factor and produce an estimate of controversy but, finally, we take each individual factor's score as the final estimate:

– Ambiguity: forces the LLM to produce an individual estimate for each individual factor plus an overall estimate of controversy (blue text in Fig. 1). This QHP estimate takes the ambiguity estimate as the prediction for the query.
– Polarization: forces the LLM to produce an individual estimate for each individual factor plus an overall estimate of controversy (blue text in Fig. 1). This QHP estimate takes the polarization estimate as the prediction for the query.
– Misinformation: forces the LLM to produce an individual estimate for each individual factor plus an overall estimate of controversy (blue text in Fig. 1). This QHP estimate takes the misinformation estimate as the prediction for the query.
– Conflicting information: forces the LLM to produce an individual estimate for each individual factor plus an overall estimate of controversy (blue text in Fig. 1). This QHP estimate takes the conflicting information estimate as the prediction for the query.

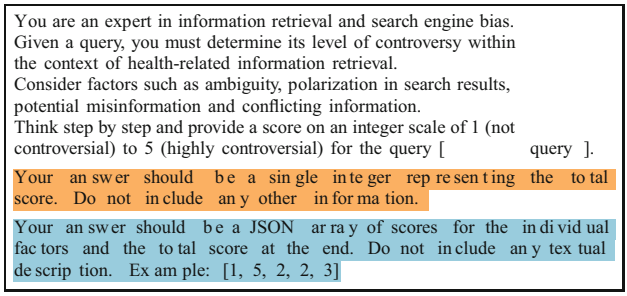

Fig. 1. Prompt template to produce QHP estimates. The orange variant asks for an overall controversy estimate, while the blue variant asks for individual estimates of ambiguity, polarization, misinformation, conflicting information and controversy.

3.3 Query Quality Classifier

In [14], the authors posed the challenge of identifying well-formed queries as a critical point for reducing downstream errors. To that end, the authors compiled a human-annotated dataset with well-formed and non-wellformed natural questions. The experiments reported an accuracy of 70.7% on the test set. Our hypothesis is that this classifier's quality estimates could have some correlation with the retrieval of relevant (or even harmful) documents and, thus, its output could provide some value as a QHP predictor. This represents an initial attempt to transfer resources built from a different area to the QHP task. Specifically, we employed a more recent model built upon the original labeled data from [14]. This AlBERT-based classifier yielded a 10% improvement (test set) over the original classifier[3], and it was thus transferred to make estimates on the query sets described below.

4 Experimental Settings

In this research, we used the TREC HM 2021 and 2022 collections [6,7], and the CLEF IR 2016 collection [22]. In the TREC HM campaigns, human assessors were asked to label documents according to three different dimensions: topical relevance, perceived credibility, and correctness with respect to medical consensus. Derived from these dimensions, graded judgments (integers in [−2,2]) were created to order documents by preference (specifically, the most preferred or helpful documents –positive scores assigned– are those topically relevant, credible, and correct, while the least preferred documents or most harmful–negative scores assigned–, are topically relevant, credible, and incorrect). Using these graded judgments, two compatibility measures [8] were computed: compatibility with an ideal ranking of helpful documents and compatibility with a ranking

[3] https://huggingface.co/dejanseo/Query-Quality-Classifier.

where the most harmful documents are at the top.[4] The goal was to create systems that have high compatibility helpful and low compatibility harmful.

For QHP, we focus on compatibility harmful as the target metric. The objective is thus to detect queries that might lead to rankings populated by misleading contents. The CLEF IR 2016 collection does not include the same type of graded judgments as the TREC HM collections, but instead trustworthiness scores for documents in [0,100]. To obtain an order of preference that allows us to differentiate between helpful and harmful documents, we mapped the original trustworthiness scores into $[-2, 2]$.[5] These two scenarios, TREC HM and CLEF IR, therefore shape an assorted evaluation design where QHP predictors can be tested with datasets where document quality annotations and distributions differ.

Regarding search models, we experimented with a traditional exact-match approach, BM25, and a more sophisticated neural re-ranker, the cross-attentional MiniLM-L-12-v2 model [33] (re-ranking the top 100 BM25 results). It should be noted that the QHP predictors proposed in this work are pre-retrieval and correlation is computed between the QHP predictions for a query set and the ranking performance achieved by each search model on the same queries. As a result, the evaluation is specific to the underlying search model. The LLM-based QHP predictions were generated by GPT-4o.[6]

5 Results

Following standard practice, we evaluated the QHP methods by measuring the correlation between the actual performance of a query set (compatibility harmful scores) and their respective predicted scores. The full set of results is shown in Table 1, while Fig. 2 shows graphically the Pearson correlation for a selected set of methods. An inspection of these results reveals that:

- In general, the predictors –whether traditional or LLM-based– are capable of capturing certain signals related to the potential presence of harmful documents in the result rankings. In many cases, the correlations are statistically significant. When comparing these results with recent QPP predictors developed for standard retrieval measures, we observe similar correlation values. For example, Faggioli et al. [13] reported Pearson correlations in the range 0.2–0.5 on the Robust04 (for nDCG@10 predictions).
- LLM-based predictors often yield better results than traditional QPP methods. This is a natural outcome, given that classical QPP approaches are designed exclusively to capture relevance, whereas LLM-based strategies are

[4] Compatibility is computed as max/min rank-biased overlap similarity to these two reference rankings.

[5] Specifically, the mapping was: -2 for $[0, 9]$, -1 for $[10, 19]$, 0 for $[20, 49]$, 1 for $[50, 74]$, and 2 for $[75, 100]$.

[6] Code and results can be accessed here: https://github.com/xianacarrera/Query-Misinformation-Prediction.

Table 1. Performance on TREC HM 2021, TREC HM 2022, and CLEF IR 2016 in terms of Pearson's ρ, Kendall's τ and Spearman's ρ. * indicates a statistically significant correlation (p-value less than 0.05) between the ranking of queries produced by each QHP method and the queries ranked by Compatibility Harmful

retriever	QHP method	TREC HM 2021			TREC HM 2022			CLEF IR 2016		
		P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ	P-ρ	K-τ	S-ρ
BM25	Avg IDF	.156	**.237**	.322	.208	.114	.154	.174*	.168*	.246*
	Max IDF	.141	.161	.232	.191	.054	.087	.248*	.232*	.344*
	Avg SCQ	.201	.220	.325	.208	.133	.193	.140	.131*	.189*
	Max SCQ	.182	.173	.255	.172	.051	.072	**.369***	**.299***	**.417***
	SCS	.151	**.237**	**.336**	.103	.043	.050	.152	.124*	.181*
	avg ICTF	.147	.216	.324	.166	.102	.126	.154	.147*	.216*
	Query Quality Classifier	-.016	.109	.166	−.217	−.124	−.158	−.059	−.038	−.049
	Controversy	.292	.117	.127	.344*	.367*	.483*	.308*	.208*	.287*
	Ambiguity	.269	.223	.290	**.445***	**.461***	**.583***	.265*	.201*	.278*
	Polarization	.345	.221	.300	.277	.292*	.393*	.292*	.209*	.288*
	Misinformation	.291	.233	.290	.382*	.370*	.499*	.270*	.187*	.266*
	Conflicting information	**.369***	.198	.268	.409*	.394*	.538*	.270*	.178*	.251*
	Controversy CoT	.331	.237	.307	.361*	.380*	.496*	.263*	.187*	.260*
MiniLM	Avg IDF	.104	.126	.200	.177	.056	.073	.136	.120*	.170*
	Max IDF	.083	.091	.125	.064	−.023	−.047	.214*	.176*	.269*
	Avg SCQ	.129	.117	.199	.192	.068	.111	.102	.092	.130
	Max SCQ	.108	.095	.140	.051	−.026	−.053	**.286***	**.229***	**.326***
	SCS	.133	.183	.254	.100	−.009	−.012	.138	.096	.136
	avg ICTF	.122	.130	.212	.158	.049	.064	.117	.099	.140*
	Query Quality Classifier	.109	.179	.268	−.155	−.003	.014	−.050	−.036	−.065
	Controversy	.290	.143	.169	.215	.207	.317	.245*	.136*	.189*
	Ambiguity	.244	.271*	.343	**.312**	**.297***	**.420***	.211*	.153*	.213*
	Polarization	.325	.258	.324	.153	.153	.222	.234*	.140*	.193*
	Misinformation	.335	**.302***	**.363***	.274	.210	.315	.199*	.117*	.167*
	Conflicting information	**.353***	.246	.308	.272	.205	.317	.198*	.103	.147
	Controversy CoT	.309	.285*	.356*	.237	.214	.314	.198*	.127*	.176*

oriented toward notions such as controversy and misinformation, which are more intimately linked to QHP.

– Predictive performance tends to be higher for BM25 compared to MiniLM. This pattern has also been observed in QPP predictors for standard retrieval metrics. In fact, many authors [10,12,13,18,30] have already highlighted the challenges associated with predicting the performance of neural models. Our results further confirm the difficulty of accurately estimating the effectiveness of neural ranking models in the novel QHP task.

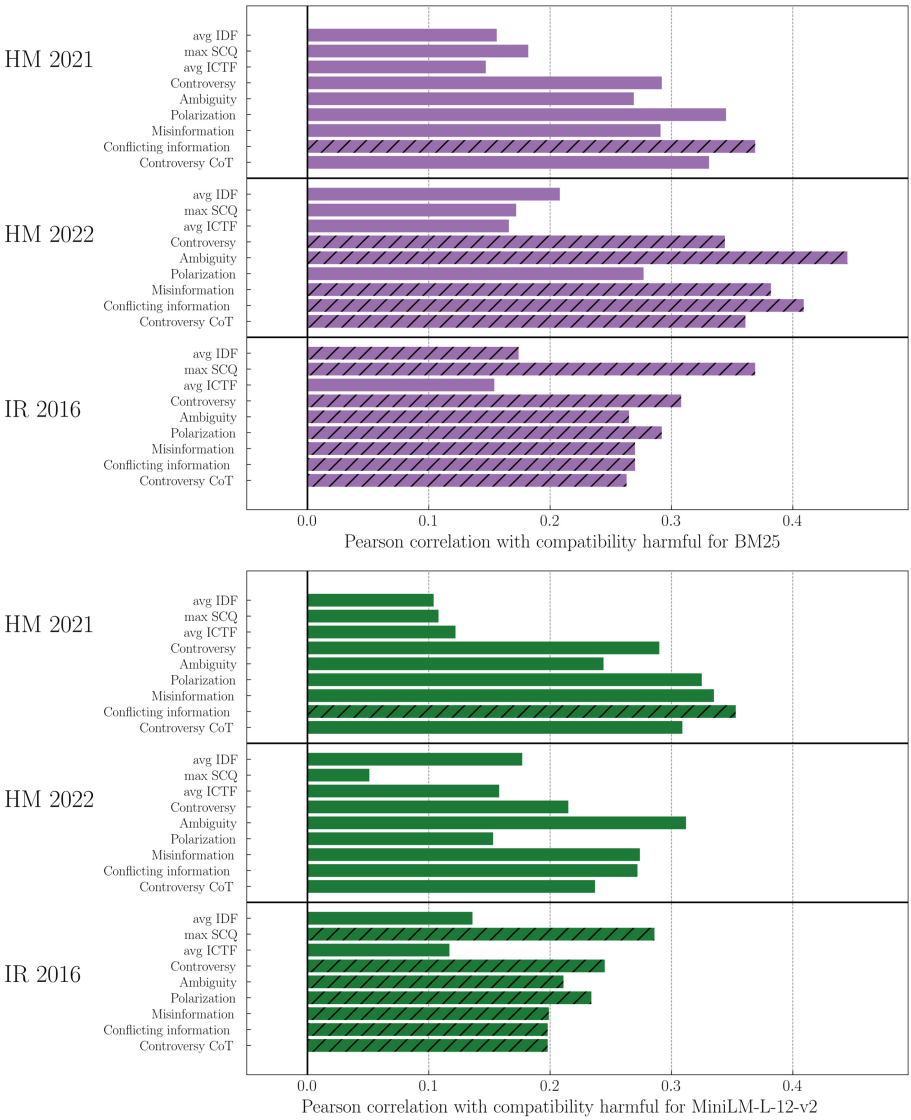

Fig. 2. Pearson correlation with compatibility harmful for selected QHP methods. Statistically significant results are hatched.

- The query quality classifier does not transfer well to this task, suggesting that its training data is not well-suited to contribute effectively to QHP.
- The classical predictors are only competitive on the CLEF IR 2016 collection. This can be attributed to the high prevalence of harmful documents in this dataset. CLEF IR 2016 induces a broader definition of harmfulness, resulting in queries with a much higher number of harmful documents (see Fig. 3, in

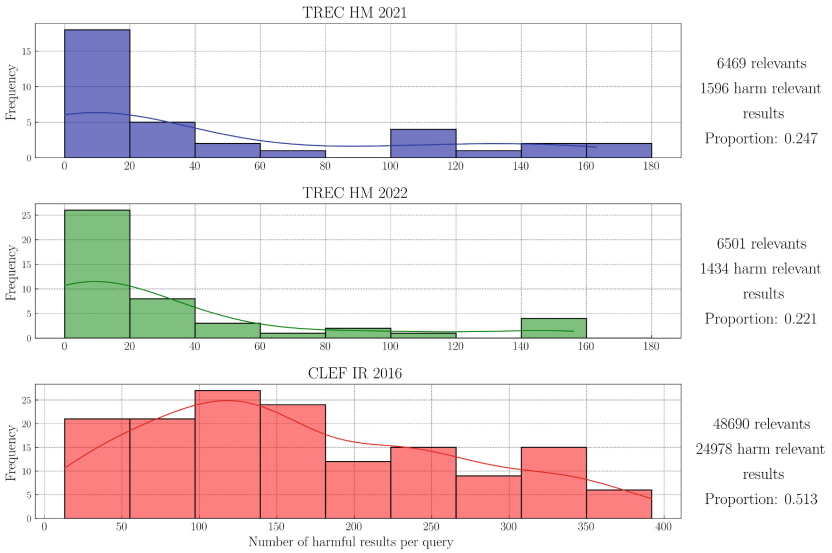

Fig. 3. Distribution of harmful search results.

CLEF IR 2016 more than 50% of relevant documents are considered harmful). In contrast, the distribution of harmful documents in the TREC HM datasets is quite sparse, with many queries having fewer than 20 harmful documents in their relevance judgments. Consequently, traditional QPP predictors –focused exclusively on relevance– are expected to perform better in CLEF IR 2016, where harmful documents constitute a substantial portion of the relevant set. On the other hand, in the TREC HM collections, where harmfulness is defined more strictly and fewer relevant documents are labeled as harmful, LLM-based predictors demonstrate superior performance due to their sensitivity to nuances such as misinformation and controversy. Finally, note also that CLEF IR 2016 queries are longer, which favors predictors like max SCQ, as opposed to those relying on the average of query term weights.

– Among LLM-based predictors, no clear winner emerges. Different individual factors appear to perform better or worse across the various collections, and further investigation is needed to determine under which conditions each predictor is most effective. Nonetheless, the Controversy (Chain-of-Thought) approach, which prompts the model to reason across five dimensions, generate individual estimates, and then use the controversy dimension as the reference point, appears to represent a good balance.

Although the results obtained by some predictors are promising, the overall levels of correlation remain relatively modest. To gain a deeper understanding of the limitations of these estimators, we selected two representative methods, max SCQ and Controversy CoT, and analyzed their behavior in predicting queries with varying levels of Compatibility Harmful.

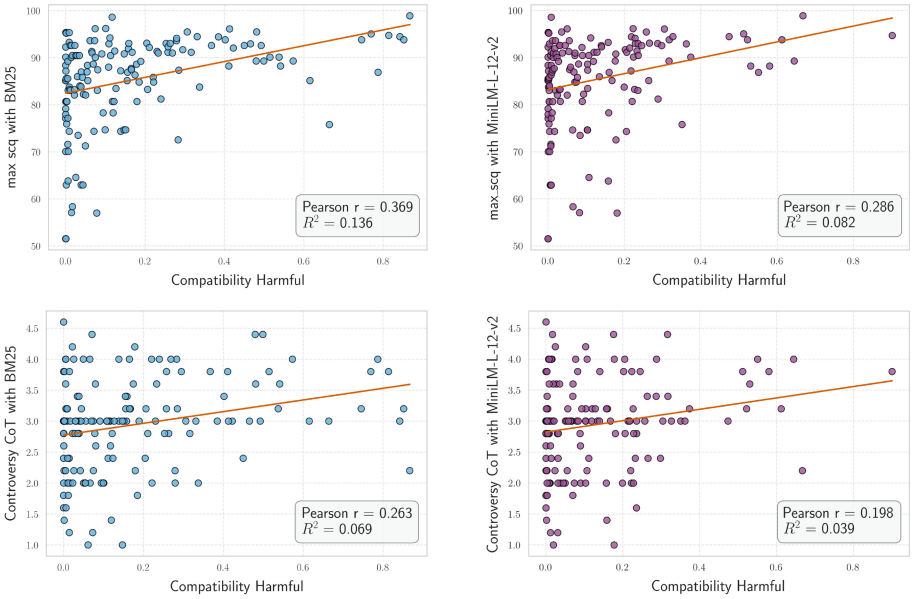

Fig. 4. Compatibility harmful vs maxSCQ (top) or Controversy CoT (bottom).

Figure 4 presents, for all queries in the CLEF IR 2016 collection, their Compatibility Harmful (CH) values and the associated predictions by each QHP estimator (max SCQ and Controversy CoT). The plots reveal a clear pattern: for queries with low CH values (left-hand side of each plot), predictions exhibit substantial dispersion, indicating reduced reliability of the estimators. In contrast, for queries with high CH values (right-hand side), the estimators demonstrate much greater stability and predictive accuracy. Due to space constraints, we only show plots for CLEF IR 2016, but this trend was consistent across all evaluated collections and retrieval models. In our view, this behavior is encouraging: the estimators are more reliable on high-risk queries (i.e., those with a high proportion of harmful results). This makes these estimators especially valuable for risk-aware applications where user safety is critical.

6 Limitations and Future Work

There are several aspects of our research that could be strengthened in future studies. For instance, in our LLM-based QHP methods, we experimented with specific prompt formulations; however, alternative prompt variants or verbalizations could yield more effective results. In this regard, we leave prompt optimization as an avenue for future work. Related to this, it would be valuable to conduct ablation studies to examine the effect of including or excluding individual factors (e.g., ambiguity or polarization), as well as exploring additional factors not considered in this work.

Another promising direction is to investigate the impact of alternative LLMs beyond GPT-4. We have conducted some preliminary experiments with LLaMA 3, obtaining encouraging results. However, Llama3's generations often exhibited formatting inconsistencies, an issue already noted in the literature [15]. Moving forward, we plan to evaluate other open-source models, such as the gpt-oss family released in August 2025, as well as LLMs from other companies and institutions.

The QHP strategies proposed and evaluated here should be interpreted as an initial, tentative set of approaches that should be extended in future work. For instance, analogous to the existence of clearly interpretable QPP methods based on classical IR elements (e.g., average IDF), we anticipate the development of QHP models grounded in new heuristics or weightings. Such methods could be enriched or guided by specialized vocabularies that flag potentially problematic expressions in user queries. Furthermore, future research should explore the transfer of additional resources and classification technologies to the QHP domain, as our evaluation of the QCC classifier represents merely an initial attempt at cross-domain knowledge transfer.

It is also worth noting that, in traditional QPP with classical retrieval systems, there is often a natural correspondence between the retrieval model and the QPP strategy (e.g., max IDF relies on IDF, a key component of BM25). This alignment does not readily extend to QPP with neural strategies, whose prediction challenges have been highlighted in the literature [10,12,13,18,30]. Furthermore, in the retrieval of reliable and reputable documents, complex telescoping strategies –combining topicality estimates and credibility assessments in a re-ranking fashion– are often employed [16,28]. In our study, we focused on BM25 and MiniLM, which do not explicitly incorporate credibility or factuality signals. Therefore, the resulting QHP estimates correspond to the output of an initial search phase within a potentially more complex pipeline. In the future, we plan to assess the effectiveness of QHP methods in predicting the performance of more sophisticated retrieval systems, such as those tailored for misinformation detection [16,28]. We will try to replicate this type of systems and test the transferability of our QHP methods to these multi-stage retrieval situations. In any case, our initial research efforts reported in this paper have focused on BM25 and MiniLM as they represent generalizable search strategies accessible to general practitioners, whereas specialized teleporting pipelines have not yet achieved the widespread adoption and impact of these two models.

Finally, we acknowledge the importance of estimating additional performance measures related to credibility and factual correctness, beyond Compatibility Harmful. Addressing these broader dimensions will be essential for advancing the QHP research agenda.

7 Conclusions

In this work, we have introduced Query Harmfulness Prediction as a novel research challenge that extends QPP by explicitly accounting for the risk of retrieving harmful, yet topically relevant results. Through a systematic evaluation of pre-retrieval predictors –including classical QPP methods, LLM-based

methods, and query quality classifiers–, we demonstrated that while existing QPP approaches capture some signal of harmfulness, LLM-driven strategies are generally more effective. Additional analyses of the QHP predictors have shown strong predictive power on high-risk queries and, overall, these results underline the importance of further advancing QHP methods to ensure safer retrieval systems.

Acknowledgments. The authors thank the financial support from the Agencia Estatal de Investigación (Spain) (PID2022-137061OB-C22 funded by MICIU/AEI/10.13039/501100011033), the Xunta de Galicia - Conselleria de Educación, Ciencia, Universidades e Formación Profesional (Centro de investigación de Galicia acreditación 2024–2027 ED431G-2023/04 and Reference Competitive Group accreditation ED431C 2022/19) and the European Union (European Regional Development Fund - ERDF). This research is also supported by the project Cátedra de IA aplicada a la Medicina Personalizada de Precisión (Cátedras ENIA, TSI-100932-2023-3); Cátedras ENIA is funded by the Ministerio de Transformación Digital y Función Pública (Secretaría de Estado de Digitalización e Inteligencia Artificial); and by the NextGeneration EU-fund.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arabzadeh, N., Khodabakhsh, M., Bagheri, E.: BERT-QPP: contextualized pre-trained transformers for query performance prediction. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 2857–2861. CIKM '21, Association for Computing Machinery, New York (2021)
2. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
3. Bubeck, S., et al.: Sparks of artificial general intelligence: early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023)
4. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval, synthesis lectures on information concepts, retrieval, and services, vol. 2. Morgan & Claypool Publishers (2010)
5. Clarke, C., Maistro, M., Rizvi, S., Smucker, M., Zuccon, G.: Overview of the TREC 2020 health misinformation track (2020)
6. Clarke, C., Maistro, M., Seifikar, M., Smucker, M.: Overview of the TREC 2022 health misinformation track (notebook) (2022)
7. Clarke, C., Maistro, M., Smucker, M.: Overview of the TREC 2021 health misinformation track (2021)
8. Clarke, C., Smucker, M., Vtyurina, A.: Offline evaluation by maximum similarity to an ideal ranking. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 225–234. CIKM '20, Association for Computing Machinery, New York (2020)
9. Cronen-Townsend, S., Zhou, Y., Croft, W.B.: Predicting query performance. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 299–306. SIGIR '02, Association for Computing Machinery, New York (2002)

10. Datta, S., Ganguly, D., Mitra, M., Greene, D.: A relative information gain-based query performance prediction framework with generated query variants. ACM Trans. Inf. Syst. **41**(2), 1–31 (2023)
11. Ebrahimi, S., Khodabakhsh, M., Arabzadeh, N., Bagheri, E.: Estimating query performance through rich contextualized query representations. In: Lecture Notes in Computer Science, pp. 49–58. Springer Nature Switzerland, Cham (2024)
12. Faggioli, G., Ferro, N., Perego, R., Tonellotto, N.: Query performance prediction using dimension importance estimators. In: Hauff, C. (ed.) Advances in Information Retrieval, pp. 202–217. Springer Nature Switzerland, Cham (2025)
13. Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., Piwowarski, B.: Query performance prediction for neural IR: are we there yet? In: Advances in Information Retrieval: 45th European Conference on Information Retrieval. ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, pp. 232–248. Springer-Verlag, Berlin, Heidelberg (2023)
14. Faruqui, M., Das, D.: Identifying well-formed natural language questions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 798–803 (2018)
15. Farzi, N., Dietz, L.: Does UMBRELA work on other LLMs? In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3214–3222. ACM, Padua Italy (2025)
16. Fernández-Pichel, M., Losada, D.E., Pichel, J.C.: A multistage retrieval system for health-related misinformation detection. Eng. Appl. Artif. Intell. **115**, 105211 (2022)
17. Goeuriot, L., et al.: Consumer health search at CLEF eHealth 2021. In: CEUR Workshop Proceedings, pp. 1–19. CEUR (2021)
18. Hashemi, H., Zamani, H., Croft, W.B.: Performance prediction for non-factoid question answering. In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 55–58. ICTIR '19, Association for Computing Machinery, New York (2019)
19. Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 1419–1420. CIKM '08, Association for Computing Machinery, New York (2008)
20. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) String Processing and Information Retrieval, pp. 43–54. Springer, Berlin Heidelberg, Berlin, Heidelberg (2004)
21. He, B., Ounis, I.: Query performance prediction. Inf. Syst. **31**(7), 585–594 (2006)
22. Jimmy, J., Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L.: Overview of the CLEF 2018 consumer health search task. In: International Conference of the Cross-Language Evaluation Forum for European Languages (2018)
23. Khodabakhsh, M., Zarrinkalam, F., Arabzadeh, N.: BertPE: a bert-based pre-retrieval estimator for query performance prediction. In: Lecture Notes in Computer Science, pp. 354–363. Springer Nature Switzerland, Cham (2024)
24. Kurland, O., Shtok, A., Carmel, D., Hummel, S.: A unified framework for post-retrieval query-performance prediction. In: Amati, G., Crestani, F. (eds.) Advances in Information Retrieval Theory, pp. 15–26. Springer, Berlin, Heidelberg (2011)
25. Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., Rijke, M.d.: Query Performance prediction using relevance judgments generated by large language models. ACM Trans. Inf. Syst. (2025). arXiv:2404.01012 [cs]

26. Milanese, G.C., Peikos, G., Pasi, G., Viviani, M.: Fact-driven health information retrieval: integrating LLMs and knowledge graphs to combat misinformation. In: Hauff, C., et al. (eds.) Advances in Information Retrieval, pp. 192–200. Springer Nature Switzerland, Cham (2025)
27. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.L.: 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, pp. 209–216 (2017)
28. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: 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, pp. 2066–2070. SIGIR '21, Association for Computing Machinery, New York (2021)
29. Saleminezhad, A., Arabzadeh, N., Beheshti, S., Bagheri, E.: Context-aware query term difficulty estimation for performance prediction. In: Lecture Notes in Computer Science, pp. 30–39. Springer Nature Switzerland, Cham (2024)
30. Singh, A., Ganguly, D., Datta, S., McDonald, C.: Unsupervised query performance prediction for neural models with pairwise rank preferences. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2486–2490. SIGIR '23, Association for Computing Machinery, New York (2023)
31. Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1930–1940. SIGIR '24, Association for Computing Machinery, New York (2024)
32. Upadhyay, R., Viviani, M.: Enhancing health information retrieval with RAG by prioritizing topical relevance and factual accuracy. Discover Comput. **28**(1), 27 (2025)
33. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural. Inf. Process. Syst. **33**, 5776–5788 (2020)
34. Zamani, H., Croft, W.B., Culpepper, J.S.: Neural query performance prediction using weak supervision from multiple signals. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 105–114. ACM, Ann Arbor MI (2018)
35. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval, pp. 52–64. ECIR'08, Springer-Verlag, Berlin, Heidelberg (2008)
36. Zhou, Y., Croft, W.B.: Query performance prediction in web search environments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 543–550. SIGIR '07, Association for Computing Machinery, New York (2007)
37. Zuccon, G., Koopman, B.: Integrating understandability in the evaluation of consumer health search engines. In: Jones, G.J.F., Kelly, L., Zobel, J., Mueller, H., Goeuriot, L. (eds.) Proceedings of the 2014 SIGIR Workshop on Medical Information Retrieval (MedIR), pp. 29–32. Dublin City University (2014)