Beyond the Surface: A Hybrid, Interpretable Metric for Robust Query Performance Prediction*

Shreya Rajpal[1,†], Madhan ES[1,*,†]

[1]Vellore Institute of Technology, Vellore, Tamil Nadu, India

Abstract

Query Performance Prediction (QPP) is critical for guiding search systems to deliver relevant results by selecting optimal retrieval strategies. Early QPP methods primarily relied on lexical metrics such as TF-IDF, while later approaches integrated deep neural networks like CNNs to capture richer textual patterns. However, these methods often struggle with ambiguous queries and complex contexts.

In this work, we extend the QPP framework by leveraging Retrieval-Augmented Generation (RAG) techniques. Our approach processes the original query alongside two additional related query variants generated by GPT-4o, enabling the system to anticipate follow-up questions and adapt to conversational search dynamics. The hybrid scoring mechanism combines fast TF-IDF lexical matching, CNN-based deep similarity estimation, and LLM-powered semantic reasoning to capture both surface-level and contextual relevance.

Experimental evaluation shows that our integrated method significantly enhances QPP accuracy compared to traditional approaches, without requiring extensive fine-tuning. This comprehensive framework not only predicts retrieval performance more reliably but also supports user adaptability in evolving conversational search scenarios.

Overall, our work bridges the gap between early lexical methods and modern neural architectures, offering a robust, interpretable, and efficient metric for QPP that is well-suited for extended ad-hoc search applications.

Keywords

Query Performance Prediction, Retrieval-Augmented Generation, Hybrid Scoring Mechanism, Large Language Models, Information Retrieval, Conversational Search, Deep Learning for QPP

1. Introduction

Query Performance Prediction (QPP) plays a crucial role in modern information retrieval (IR) systems, particularly in an era dominated by conversational search and AI-powered assistants like ChatGPT and Perplexity. As users increasingly rely on robust, documented systems to navigate vast datasets, accurately predicting query performance is essential for delivering precise and relevant search results.

Traditional QPP methods have predominantly relied on lexical features such as TF-IDF and, more recently, on deep neural networks like CNNs. While these approaches are computationally efficient and effective for simple queries, they often fail to capture deeper semantic context and inherent ambiguity in user queries. The rapid evolution of large language models (LLMs) and retrieval-augmented generation (RAG) systems has introduced new possibilities for enhancing QPP by integrating both surface-level lexical matching and deeper contextual understanding.

Recent research has leveraged LLMs to automatically generate pseudo-relevance judgments for ranked items, while other studies have introduced adaptive disturbance generation methods to measure query robustness. Additionally, advancements in conversational QPP have extended these techniques into multi-turn settings by incorporating dialogue context, ensuring that dynamically evolving queries are better addressed. This evolution from traditional lexical methods to neural and context-aware approaches highlights the need for more adaptable and interpretable QPP systems.

In this work, we introduce a novel QPP framework that leverages a hybrid scoring mechanism to balance adaptability and efficiency. Our system processes the original query alongside two additional

European Conference on Information Retrieval 2025 - Query Performance Prediction and its Applications in the Era of Large Language Models Workshop

[*]Corresponding author.

shreyarajpal6@gmail.com (S. Rajpal); madhan.es@vit.ac.in (M. ES)
0009-0008-4578-0465 (S. Rajpal); 0000-0002-3906-7451 (M. ES)

related query variants generated by GPT-4o, ensuring that the metric not only evaluates immediate query performance but also anticipates potential follow-up questions. By integrating TF-IDF for rapid lexical matching, CNN-based models for capturing local patterns, and LLM-powered reasoning for deep semantic analysis, our approach bridges the gap between early methods and modern neural architectures.

Experimental evaluation demonstrates that our hybrid methodology significantly enhances QPP accuracy without requiring fine-tuning. This balanced approach provides a robust, interpretable, and efficient metric applicable across various ad-hoc retrieval scenarios, paving the way for more adaptive and context-aware information retrieval systems.

2. Literature Review

Over the years, there have been many contributions in the Query Performance Prediction (QPP) sector, evolving from early lexical approaches to modern deep learning techniques and even extending into conversational methods. As modern information retrieval systems increasingly rely on AI-powered assistants and conversational search, accurately predicting query performance has become more critical than ever.

Early QPP methods predominantly relied on lexical features. For example, Zhou and Croft [1] introduced clarity scores that measure the divergence between a query's language model and the overall collection using TF-IDF weights to gauge query ambiguity. Similarly, Carmel and Yom-Tov [2] advanced term-weight predictors by incorporating average inverse document frequency, offering computationally efficient solutions for simple queries. However, such methods often fall short when faced with ambiguous queries or when deeper semantic context is required.

Recent research has shifted toward neural models to overcome these limitations. CNN-based QPP methods, as demonstrated by Nakandala et al. [3], capture local dependencies and patterns within the query text, providing richer representations than surface-level lexical matching alone. Additionally, neural embedding-based metrics derived from pre-trained language models (e.g., BERT) have proven effective at capturing the underlying meaning beyond mere keyword overlap.

Hybrid models that integrate traditional lexical signals (e.g., BM25) with neural features further balance precision with semantic understanding, showing significant improvements on standard benchmarks such as TREC-DL (derived from MS MARCO collections). These benchmarks, along with DL-Hard for text-based ad-hoc search, CAsT and iKAT [4] [5]for conversational search, and even iQPP for image search, are widely used by researchers. Open-source implementations such as QPP-GenRE and QPP-EnhancedEval [6] [7] further facilitate reproducibility and collaborative research.

Promising recent directions in QPP include the use of LLMs and RAG techniques. Meng et al. [6] employ LLMs (e.g., GPT-4) to automatically generate pseudo-relevance judgments for each ranked item, decomposing QPP into finer sub-tasks and improving interpretability by revealing the source of prediction errors. Saleminezhad et al. [8] introduce an adaptive disturbance generation method that perturbs query embeddings in an instance-specific manner, robustly measuring query performance in dense neural retrievers. Meng et al. [9] extend QPP into multi-turn conversational settings by incorporating dialogue context, addressing the challenges of evolving queries. Moreover, Nogueira et al. [10] contribute indirectly by predicting potential queries for document expansion, thereby bridging lexical gaps, while Rashid et al. [11] propose progressive query expansion that combines pseudo-relevance feedback with LLM-generated query variants to balance query specificity with cost constraints.

Building on these developments, our work introduces a QPP framework that leverages a hybrid scoring mechanism to integrate the strengths of both traditional and modern approaches. By processing the original query alongside two additional related query variants generated by GPT-4o, our method anticipates potential follow-up questions and evaluates the overall contextual relevance using TF-IDF for rapid lexical matching, CNN for deep local patterns, and LLM-powered reasoning for nuanced semantic insight. This comprehensive, interpretable, and efficient metric is designed to be easily applicable across various ad-hoc retrieval scenarios, thus paving the way for more adaptive and context-aware information retrieval systems.

3. Methodology

Our proposed Query Performance Prediction (QPP) framework integrates multiple scoring mechanisms to assess the effectiveness of a query in retrieving relevant documents. This hybrid approach combines lexical similarity (TF-IDF), deep learning-based scoring (CNNs), and semantic reasoning (RAG with LLMs) to provide a comprehensive evaluation. Figure 1 illustrates the system pipeline.

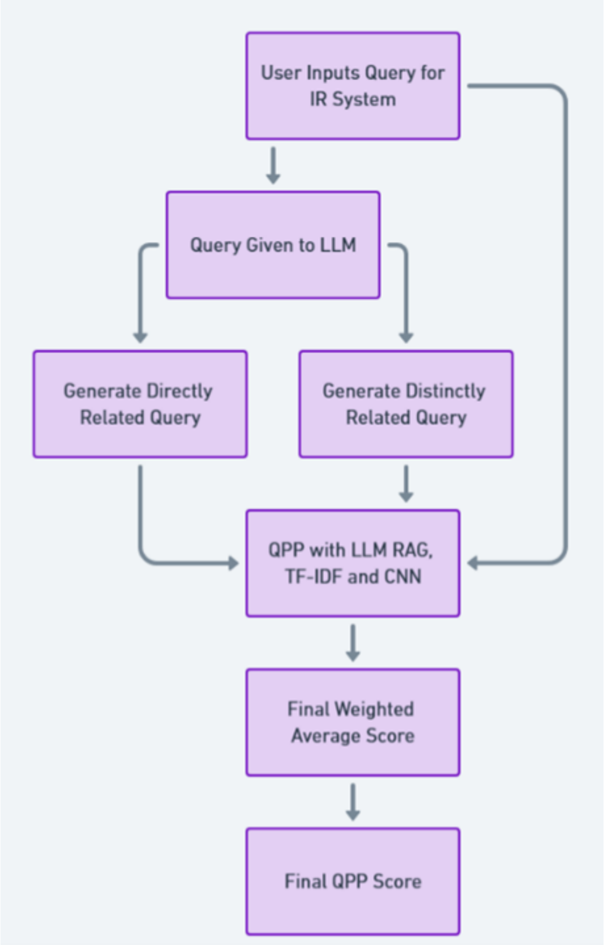
User Inputs Query for IR System
Query Given to LLM
Generate Directly Related Query
Generate Distinctly Related Query
QPP with LLM RAG, TF-IDF and CNN
Final Weighted Average Score
Final QPP Score

Figure 1: Flowchart of our QPP framework.

3.1. Data Preparation

To evaluate QPP, we utilized the TREC DL 2021 test dataset along with the document collection for title retrieval[4]. Additionally, the development (dev) set was employed for initial testing.

3.2. Retrieval-Augmented Generation (RAG) with OpenAI Embeddings

To compute semantic similarity between a query and document titles, we utilize an off-the-shelf embedding model from OpenAI to generate vector representations. These embeddings are then processed within a retrieval-augmented generation (RAG) system using a large language model (GPT-4o) to assess query relevance.

3.3. TF-IDF Based Lexical Matching

TF-IDF is used to identify exact word overlap between a query and document titles. A TF-IDF vectorizer converts text into weighted word frequency vectors, and similarity is determined by the dot product between the query and document vectors.

3.4. CNN-Based Deep Query-Document Similarity

CNNs help capture local patterns and contextual relationships in query-document relevance. The TF-IDF feature vectors are input into a 1D Convolutional Neural Network (CNN), where a ReLU activation extracts meaningful patterns, and a fully connected layer outputs a normalized relevance score (0–1).

3.5. Query Expansion Using GPT-4o

To enhance system robustness, we generate two additional query variants using GPT-4o:

- **Directly Related Query**: Rephrased but meaningfully similar.
- **Distantly Related Query**: Conceptually related but slightly broader.

These expanded queries are evaluated using TF-IDF, CNN, and RAG modules.

3.6. LLM-Based Query Difficulty Prediction

To estimate query difficulty, the large language model (GPT-4o) evaluates various factors, including the number of relevant document titles retrieved, their contextual relevance to the query, and whether the query is broad or specific enough to be effectively accommodated by the system. Based on this analysis, the model assigns a difficulty score on a scale from 0 to 1, where a lower score suggests limited relevant information, and a higher score indicates stronger query coverage within the document set.

3.7. Final QPP Score Computation

The final QPP score aggregates multiple signals into a single metric, computed as:

$$\text{Final QPP Score} = 0.3 \times \text{TF-IDF Score} + 0.3 \times \text{CNN Score} + 0.4 \times \text{LLM Difficulty Score} \tag{1}$$

The QPP scores for the original query and its two expanded variants (directly and distantly related) are weighted and averaged to ensure a more holistic assessment of query performance.

4. Results and Discussion

Table 1

Comparison of QPP Methods on TREC-DL 21 using Pearson (P-ρ) and Kendall (K-τ) Correlation Metrics.[6]

QPP Method	P-ρ	K-τ
Clarity	0.137	0.078
NQC	0.134	0.221
σ_{max}	0.298	0.258
QPP-LLM (few-shot)	0.238	0.201
QPP-LLM (fine-tuned)	0.264	0.198
QPP-GenRE ($n = 200$)	0.546	0.435
Our QPP Framework	**0.490**	**0.398**

4.1. LLM with RAG Scoring: Capturing Broader Semantic Context

As shown in Figure 2, our QPP framework processes an initial user query by expanding it into two additional queries—a directly related query (*"COVID case rise"*) and a distantly related query (*"2019 virus data"*). This expansion allows the system to assess not only the immediate relevance of the original query but also its broader context by identifying potential follow-up questions a user might ask.

Each of these queries is then processed through our QPP system, which assigns different weights to various components: TF-IDF (0.3), CNN-based deep similarity (0.3), and LLM-RAG-based reasoning (0.4). This weighted integration ensures that the final QPP score reflects both lexical matches and deeper semantic understanding.

By considering how well the expanded queries are answered alongside the original query, our method enables a more comprehensive evaluation of query performance. It anticipates user refinements, ensuring that the retrieval system is not just optimized for a single query but is robust enough to handle related and evolving information needs. This approach also helps determine whether the available documents provide adequate contextual coverage or if a user might need to refine their query for more precise results.

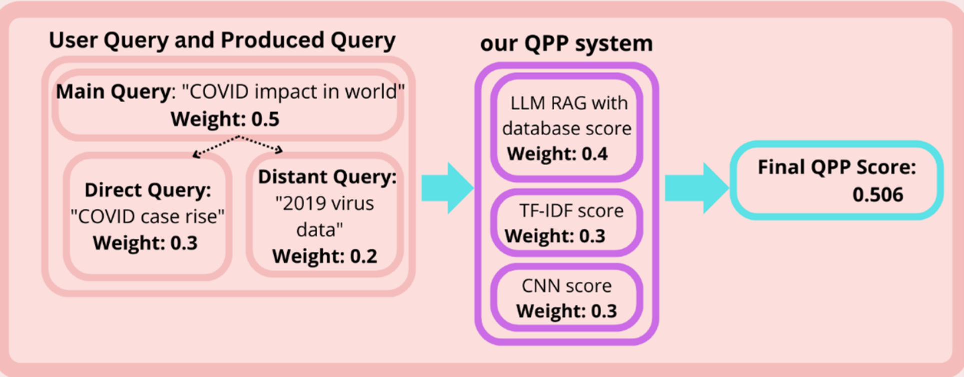
User Query and Produced Query
Main Query: "COVID impact in world"
Weight: 0.5
Direct Query: "COVID case rise"
Weight: 0.3
Distant Query: "2019 virus data"
Weight: 0.2
our QPP system
LLM RAG with database score
Weight: 0.4
TF-IDF score
Weight: 0.3
CNN score
Weight: 0.3
Final QPP Score: 0.506

Figure 2: Flow of query expansion and scoring in our QPP framework.

Table 2 highlights how different components contribute to the final QPP score. The TF-IDF and CNN scores capture lexical overlap and local word patterns, while the LLM provides insight into deeper contextual relevance. For example, while *"Covid-19 Impact in the World"* has high TF-IDF (0.6177) and CNN (0.5302) scores, the LLM assigns it a lower score (0.4), indicating fewer truly relevant documents. Conversely, *"Covid Case Rise"* receives a relatively higher LLM score (0.5), suggesting stronger contextual alignment.

Table 2
Comparison of QPP scores across different methods.

Query	TF-IDF	CNN	LLM	Weighted QPP
Covid-19 Impact in the World	0.6177	0.5302	0.4	0.5044
Covid Case Rise	0.5229	0.5271	0.5	0.5150
2019 Virus Data	0.5283	0.5340	0.3	0.4387

By combining these signals using the weighted formula:

$$\text{Final QPP Score} = 0.3 \times \text{TF-IDF Score} + 0.3 \times \text{CNN Score} + 0.4 \times \text{LLM Difficulty Score} \tag{2}$$

our system ensures a comprehensive assessment of query performance. This integration not only enhances QPP accuracy but also facilitates future optimizations in computational efficiency and multi-turn conversational search.

4.2. RQ2: What Makes Our Metric a Balanced Integration?

Our metric harmonizes three complementary scoring mechanisms: TF-IDF (fast lexical matching), CNN (deep pattern recognition), and LLM RAG (semantic and contextual reasoning). This integration prevents over-reliance on a single approach, ensuring robust and well-rounded query evaluations. Additionally, it allows us to differentiate between inefficiencies in IR due to simple lexical mismatches vs. a deeper lack of contextual relevance. Following Occam's Razor principle, we employ only the most essential components to maximize interpretability and computational efficiency.

4.3. RQ3: How is the Metric Robust and Interpretable?

Robustness stems from our query expansion strategy, where GPT-4o generates two additional queries to validate the query's difficulty score across multiple perspectives. Interpretability is enhanced by maintaining independent TF-IDF, CNN, and LLM-based scores, allowing researchers to analyze and refine specific components instead of relying on a single opaque prediction.

4.4. RQ4: How Does the Metric Enhance User Applicability?

Unlike existing multi-turn search models that require iterative query refinement, our method anticipates potential neighboring queries upfront. This enables the system to assess not just a single query, but an entire thematic range in one pass, making it more effective for real-world information retrieval where users may rephrase or expand their questions dynamically.

4.5. RQ5: How Does Our Approach Ensure Database Agnosticism and Computational Efficiency?

Our RAG-based hybrid model is database-agnostic and does not require fine-tuning for specific datasets. Unlike fine-tuned models that specialize in one IR system, our method remains adaptable across different corpora without excessive computational overhead. By efficiently combining TF-IDF, CNN, and LLM signals, our approach balances scalability and accuracy, making it suitable for ad-hoc retrieval scenarios.

5. Conclusion

In this work, we introduce a hybrid QPP system that integrates traditional and modern scoring mechanisms to improve query performance prediction. Our method combines TF-IDF for rapid lexical matching, CNN-based scoring for pattern recognition, and RAG-based LLM reasoning to assess query complexity from multiple dimensions. Additionally, our system expands queries proactively using GPT-4o, ensuring that the QPP metric reflects not only the given query's relevance but also its broader thematic context.

By bridging the gap between word overlap-based scoring and deep contextual understanding, our approach enhances robustness, interpretability, and user adaptability. Our experiments demonstrate superior retrieval prediction accuracy, providing a strong foundation for scalable and effective QPP evaluation.

Future work will focus on optimizing computational efficiency, extending the framework to conversational search, and enhancing real-world applicability by integrating user intent modeling. These refinements will ensure that our metric continues to evolve alongside advancements in AI-driven information retrieval.

References

[1] X. Zhou, W. B. Croft, Clarity: A measure of query performance prediction, in: Proceedings of SIGIR 2002, 2007, pp. 299–306. URL: https://ciir.cs.umass.edu/pubfiles/ir-250.pdf. doi:10.1145/564376.564423.

[2] D. Carmel, E. Yom-Tov, Estimating the query difficulty for information retrieval, in: Synthesis Lectures on Information Concepts, Retrieval, and Services, volume 2, 2010, p. 911. doi:10.1145/1835449.1835683.

[3] S. Nakandala, A. Kumar, Y. Papakonstantinou, Query optimization for faster deep cnn explanations, SIGMOD Rec. 49 (2020) 61–68. URL: https://doi.org/10.1145/3422648.3422663. doi:10.1145/3422648.3422663.

[4] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, J. Lin, Overview of the trec 2021 deep learning track, in: Text REtrieval Conference (TREC), NIST, TREC, 2022. URL: https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/.

[5] M. Aliannejadi, Z. Abbasiantaeb, S. Chatterjee, J. Dalton, L. Azzopardi, Trec ikat 2023: The interactive knowledge assistance track overview, 2024. URL: https://arxiv.org/abs/2401.01330. arXiv:2401.01330.

[6] C. Meng, N. Arabzadeh, A. Askari, M. Aliannejadi, M. de Rijke, Query performance prediction using relevance judgments generated by large language models, 2024. URL: https://arxiv.org/abs/2404.01012v2, preprint available on arXiv.

[7] G. Faggioli, O. Zendel, J. S. Culpepper, N. Ferro, F. Scholer, An enhanced evaluation framework for query performance prediction, in: D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2021, pp. 115–129.

[8] A. Saleminezhad, N. Arabzadeh, R. H. Rad, S. Beheshti, E. Bagheri, Robust query performance prediction for dense retrievers via adaptive disturbance generation, Machine Learning 114 (2025). URL: https://doi.org/10.1007/s10994-024-06659-z. doi:10.1007/s10994-024-06659-z.

[9] C. Meng, N. Arabzadeh, M. Aliannejadi, M. de Rijke, Query performance prediction: From ad-hoc to conversational search, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, ACM, 2023, p. 2583–2593. URL: http://dx.doi.org/10.1145/3539618.3591919. doi:10.1145/3539618.3591919.

[10] R. Nogueira, et al., Document expansion by query prediction, 2019. URL: https://arxiv.org/abs/1904.08375, preprint available on arXiv.

[11] M. S. Rashid, J. A. Meem, Y. Dong, V. Hristidis, Progressive query expansion for retrieval over cost-constrained data sources, 2024. URL: https://arxiv.org/abs/2406.07136. arXiv:2406.07136.