QPP++ 2025: Query Performance Prediction and its Applications in the Era of Large Language Models

Chuan Meng[1][0000−0002−1434−7596], Guglielmo Faggioli[2][0000−0002−5070−2049], Mohammad Aliannejadi[1][0000−0002−9447−4172], Nicola Ferro[2][0000−0001−9219−6239], and Josiane Mothe[3][0000−0001−9273−2193]

[1] University of Amsterdam, The Netherlands
{c.meng,m.aliannejadi}@uva.nl
[2] University of Padova, Italy
{guglielmo.faggioli,ferro}@dei.unipd.it
[3] Université de Toulouse, France
josiane.mothe@irit.fr

Abstract. Query performance prediction (QPP) is a key task in information retrieval (IR) and has been studied for over a decade. The task of QPP is defined as estimating search effectiveness without human relevance judgments. In this workshop, we aim to bring together researchers and practitioners from academia and industry to discuss new perspectives on QPP. Amongst the limitations in the existing QPP literature, we can mention little work has focused on (i) predicting the performance of newly emerged large language model (LLM)-based retrievers/re-rankers or of generative AI systems, (ii) leveraging LLM to model QPP, (iii) investigating concrete applications of QPP, (iv) exploring QPP in the context of multi-modal content, and (v) Exploring multilingual QPP. Those are examples of topics that we encourage authors to contribute to in this workshop.

Keywords: Query performance prediction · Query difficulty prediction · Large language models.

1 Background and relevance to ECIR

Query performance prediction (QPP), a.k.a. query difficulty prediction, has attracted the attention of the information retrieval (IR) community for decades [7,27,6,2]. QPP aims to predict the retrieval quality of a search system for a query without relying on human-labeled relevance judgments [26,14,24]. Effective QPP benefits various tasks, e.g., query variant selection [34], selective query expansion [1,8], ranker selection [9,28,17], or query-specific pool depth prediction [15] to reduce human relevance judgment costs in collection construction.

A workshop focused on QPP will shed light on future research directions for QPP, ultimately benefiting the broader IR community. For more details, please visit our official website: https://qppworkshop.github.io/.

2 Motivation, workshop goals and desired outcomes

2.1 Motivation and workshop goals

Given the rapid advancement of LLMs [16,4], we highlight some limitations in the current QPP literature, and present examples of topics we encourage authors to explore in this workshop.

Predicting the performance of LLM-based retrievers/re-rankers, or generative AI systems. Most QPP studies [12,13] still focus on predicting the performance of rankers based on small-scale pre-trained language models (e.g., BERT [10] or T5 [33]). However, little work has explored predicting the performance of newly emerged LLM-based retrievers/re-rankers [19,32,31], or more broadly, LLM-based generative AI systems [29]. Therefore, we aim to address questions including, but not limited to, the following:

(i) To what extent does the performance of existing QPP methods (designed for retrievers/re-rankers using small-scale pre-trained languages) generalise to LLM-based retrievers and re-rankers?
(ii) How can we effectively predict the performance of emerging new re-ranking paradigms based on LLMs, such as pair-wise or list-wise re-rankers [32,31]?
(iii) How can we predict the performance of generative AI systems? E.g., how to predict the text generation quality of an LLM in response to a prompt [5]?

Leveraging the capabilities of LLMs to enhance QPP quality. LLMs have been applied to a wide range of natural language processing (NLP) and IR tasks, achieving numerous state-of-the-art results [19,31]. However, few studies have explored leveraging LLM to model QPP [24]. We aim to address questions including, but not limited to, the following:

(i) What kind of features from LLMs (e.g., embedding) can we use for QPP?
(ii) How well a QPP model based on LLMs performs when predicting the ranking quality of an LLM-based retriever/re-ranker or the text generation quality of an LLM?

Applying QPP to benefit various downstream tasks. Most studies evaluate QPP methods only using metrics not directly correlated to downstream tasks, e.g., linear correlation coefficients. However, it is underexplored whether QPP methods evaluated in such a way can be effectively applied to benefit downstream tasks, especially when it comes to applying them to important tasks in the era of LLMs, e.g., retrieval-augmented generation (RAG).

(i) Which downstream tasks (especially in the era of LLMs) does QPP have the potential to benefit? E.g., in RAG, can QPP be used to determine when to rely on the retrieved documents [20]? In LLM-based re-ranking, can QPP be used to determine query-specific re-ranking depths [25]?
(ii) How exactly to use QPP to benefit those tasks?

Exploring QPP in the context of multi-modal content. Most studies focus on QPP in the context of text. However, research on QPP in the context of multi-modal content, such as images or even video, remains limited [30]. We aim to address questions including, but not limited to, the following:

(i) QPP for text-to-image search [35], image-to-image search [30], or image generation.
(ii) QPP in the context of video search/generation.

Exploring multilingual QPP. There is limited research on QPP for languages other than English. How well do current QPP methods generalise to non-English languages?

Other QPP-related research. Beyond the above aspects, there is a need for deeper exploration in other QPP-related areas, including, but not limited to, the following: QPP for conversational search [15,23,21,22,18], recommendation systems, question answering, and fairness. We aim to stimulate discussions on these topics.

2.2 Desired outcomes

We anticipate two primary outcomes from the workshop:

(i) We plan to compile the workshop proceedings from the papers submitted by participants, which will be published in the CEUR-WS.org proceedings.
(ii) We aim to draft a position paper outlining the roadmap identified during the workshop discussions and submit it to the SIGIR Forum.

3 Format and structure

3.1 Paper selection

Paper types. We welcome manuscripts about any QPP-related subjects. We welcome a diverse range of submission types:
– New papers. We welcome paper submissions in various types, including research papers, position papers, reproducibility papers, survey papers, and data collection papers. All submissions should be ranged from 4 to 10 pages in length, including references.
– Published papers. We welcome papers already accepted at top-tier conferences or in journals, e.g., SIGIR, CIKM, WWW, ECIR, ACL, EMNLP, TOIS, IP&M. Authors should submit 1-page abstract for a published paper, including the full reference where the paper was accepted.

To facilitate workshop discussions, we encourage (though it is not mandatory) authors to use common, publicly available benchmarks. This will help authors share their experimental experiences and insights more effectively.

Selection process and type of presentation. Each manuscript will be peer-reviewed by at least three programme committee (PC) members. Authors of accepted papers will be invited to give oral presentations at the workshop. The accepted papers will be published in the CEUR-WS.org proceedings series. Since the proceedings only accept papers with a minimum of 4 pages, we will combine 1-page abstract submissions into one or more papers for publication.

3.2 Planned activities

We plan to organise a full-day workshop, with the tentative schedule presented at our website: https://qppworkshop.github.io/.

4 Intended audience

Targeted groups. Our intended audience for this workshop consists of two main groups: (i) Researchers and practitioners from academia and industry with a background in IR and experience in QPP. We believe they will gain valuable insights by discussing their recent work and potential research directions. (ii) Individuals working in IR, NLP, or general artificial intelligence who are new to QPP and interested in exploring this area. We aim to guide them towards important research directions/questions, and provide valuable resources on QPP.

5 List of organisers

Chuan Meng is a final-year Ph.D. candidate at the University of Amsterdam, supervised by Prof. dr. Maarten de Rijke and dr. Mohammad Aliannejadi. He is currently an applied scientist intern at Amazon. He works on IR and NLP, focusing on QPP, conversational search, neural ranking and LLM-based relevance judgement prediction. He has published relevant papers in proceedings such as SIGIR, EMNLP, CIKM, and AAAI. He has proposed novel QPP frameworks using LLMs [24] and for conversational search [23,20,21]. He serves as a committee member for various conferences including SIGIR, WWW, ACL, EMNLP, WSDM, CIKM. He also serves as a journal reviewer for TOIS and IP&M. He co-organised a tutorial on QPP at ECIR 2024 [2] and SIGIR-AP 2024 [3].

Guglielmo Faggioli is a postdoctoral researcher at the University of Padua (UNIPD), Italy. His main research interests regard IR focusing on evaluation, performance modelling, QPP, conversational search, and privacy-preserving IR. He published several papers on the QPP topic. In terms of organisational activities, he has been a co-organiser of the QPP++ workshop, colocated with ECIR 2023, the intelligent Disease Progression Prediction (iDPP) CLEF Lab since 2022 (iDPP@CLEF 2022, iDPP@CLEF 2023, iDPP@CLEF 2024), and program co-chair of IIR 2023.

Mohammad Aliannejadi is an assistant professor at the University of Amsterdam. His research interests include conversational information access, recommender systems, and QPP. Mohammad has co-organised various evaluation campaigns such as TREC CAsT, TREC iKAT, ConvAI3, and IGLU, focusing on different aspects of user interaction with conversational agents. Moreover, Mohammad has held multiple tutorials and lectures on conversational search, such as CHIIR, SIKS, and ASIRF.

Nicola Ferro is a full professor in Computer Science at the Department of Information Engineering at the University of Padua, Italy. His main research interests are information retrieval, data management, and representation, and their evaluation. He chairs the Steering Committee of CLEF, the European evaluation initiative on multimodal and multilingual information access systems, and the Steering Committee of ESSIR, the European Summer School on Information Retrieval. He is a senior PC Member in top-tier conferences, like ECIR, ACM SIGIR, ACM CIKM, and WSDM. He is the General Co-chair of SIGIR 2025. He was General Chair of ESSIR 2016 and Associate Editor for ACM TOIS. He was awarded the SIGIR Academy in 2023.

Josiane Mothe is a full professor in Computer Science at Université de Toulouse and researcher at Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS. Her research focuses on information systems and IR, including selective query processing and QPP. She also works on low-resource languages and deep learning on images for different applications. She participates to several EU funded projects as well as national ANR ones and is leading a 2M € regional project. She serves as a senior PC member in major conferences in IR (SIGIR, ECIR, CIKM,..), is co-editor of SIGIR Forum, and Associate Editor for ACM TOIS. She was General Co-chair of SIGIR 2023 and keynote speaker at ECIR 2024.

6 Relation to other workshops

Our workshop (QPP++ 2025) is the continuation of the QPP++ 2023 workshop held in ECIR 2023 [11]. QPP++ 2023 saw strong attendance and outcomes: it received over 10 submissions, and selected 6 of them for publication in the CEUR-WS.org proceeding; the workshop attracted approximately 25 participants. However, at the time QPP++ 2023 was held, LLMs had not yet fully captured widespread attention, and the workshop did not emphasize the application of QPP to downstream tasks. We believe that QPP++ 2025 is well-timed to advance research into QPP and its applications in the era of LLMs.

References

1. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: ECIR. pp. 127–137. Springer (2004), https://link.springer.com/chapter/10.1007/978-3-540-24752-4_10
2. Arabzadeh, N., Meng, C., Aliannejadi, M., Bagheri, E.: Query Performance Prediction: From Fundamentals to Advanced Techniques. In: ECIR. pp. 381–388. Springer (2024), https://link.springer.com/chapter/10.1007/978-3-031-56069-9_51
3. Arabzadeh, N., Meng, C., Aliannejadi, M., Bagheri, E.: Query Performance Prediction: Techniques and Applications in Modern Information Retrieval. In: SIGIR-AP. pp. 291–294 (2024)
4. Askari, A., Meng, C., Aliannejadi, M., Ren, Z., Kanoulas, E., Verberne, S.: Generative Retrieval with Few-shot Indexing. arXiv preprint arXiv:2408.02152 (2024)
5. Bizzozzero, N., Bendidi, I., Risser-Maroix, O.: Prompt Performance Prediction for Generative IR. arXiv:2306.08915 (2023)
6. Carmel, D., Yom-Tov, E.: Estimating the Query Difficulty for Information Retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services **2**(1), 1–89 (2010)
7. Carmel, D., Yom-Tov, E., Soboroff, I.: SIGIR Workshop Report: Predicting Query Difficulty - Methods and Applications. SIGIR Forum **39**(2), 25–28 (Dec 2005). https://doi.org/10.1145/1113343.1113349, https://doi.org/10.1145/1113343.1113349
8. Datta, S., Ganguly, D., MacAvaney, S., Greene, D.: A Deep Learning Approach for Selective Relevance Feedback. In: ECIR. pp. 189–204. Springer (2024)
9. Deveaud, R., Mothe, J., Ullah, M.Z., Nie, J.Y.: Learning to Adaptively Rank Document Retrieval System Configurations. ACM Transactions on Information Systems (TOIS) **37**(1), 1–41 (2018), https://hal.science/hal-02092955/document
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL. pp. 4171–4186 (2019), https://aclanthology.org/N19-1423.pdf
11. Faggioli, G., Ferro, N., Mothe, J., Raiber, F.: QPP++ 2023: Query-Performance Prediction and its Evaluation in New Tasks. In: ECIR. pp. 388–391. Springer (2023)
12. Faggioli, G., Formal, T., Lupart, S., Marchesin, S., Clinchant, S., Ferro, N., Piwowarski, B.: Towards Query Performance Prediction for Neural Information Retrieval: Challenges and Opportunities. In: ICTIR. pp. 51–63 (2023)
13. Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., Piwowarski, B.: Query Performance Prediction for Neural IR: Are We There Yet? In: ECIR. pp. 232–248. Springer (2023)
14. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: An Enhanced Evaluation Framework for Query Performance Prediction. In: ECIR. pp. 115–129. Springer (2021)
15. Ganguly, D., Yilmaz, E.: Query-specific Variable Depth Pooling via Query Performance Prediction. In: SIGIR. pp. 2303–2307 (2023)
16. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv:2310.06825 (2023)
17. Khramtsova, E., Zhuang, S., Baktashmotlagh, M., Zuccon, G.: Leveraging LLMs for Unsupervised Dense Retriever Ranking. arXiv:2402.04853 (2024)
18. Lu, L., Meng, C., Ravenda, F., Aliannejadi, M., Crestani, F.: Zero-Shot and Efficient Clarification Need Prediction in Conversational Search. In: ECIR. Springer (2025)

19. Ma, X., Wang, L., Yang, N., Wei, F., Lin, J.: Fine-Tuning LLaMA for Multi-Stage Text Retrieval. arXiv:2310.08319 (2023), https://arxiv.org/abs/2310.08319
20. Meng, C.: Query Performance Prediction for Conversational Search and Beyond. In: SIGIR (2024)
21. Meng, C., Aliannejadi, M., de Rijke, M.: Performance Prediction for Conversational Search Using Perplexities of Query Rewrites. In: QPP++2023. pp. 25–28 (2023)
22. Meng, C., Aliannejadi, M., de Rijke, M.: System Initiative Prediction for Multi-turn Conversational Information Seeking. In: CIKM. pp. 1807–1817 (2023)
23. Meng, C., Arabzadeh, N., Aliannejadi, M., de Rijke, M.: Query Performance Prediction: From Ad-hoc to Conversational Search. In: SIGIR. p. 2583–2593 (2023)
24. Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., de Rijke, M.: Query Performance Prediction using Relevance Judgments Generated by Large Language Models. arXiv:2404.01012 (2024), https://arxiv.org/abs/2404.01012
25. Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., de Rijke, M.: Ranked List Truncation for Large Language Model-based Re-Ranking. In: SIGIR (2024)
26. Mizzaro, S., Mothe, J., Roitero, K., Ullah, M.Z.: Query Performance Prediction and Effectiveness Evaluation Without Relevance Judgments: Two Sides of the Same Coin. In: SIGIR. pp. 1233–1236 (2018)
27. Mothe, J., Tanguy, L.: Linguistic Features to Predict Query Difficulty. In: ACM Conference on Research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop. pp. 7–10 (2005)
28. Mothe, J., Ullah, M.Z.: Selective Query Processing: A Risk-Sensitive Selection of Search Configurations. TOIS **42**(1), 1–35 (2023), https://ut3-toulouseinp.hal.science/hal-03853742/file/Mothe_2022_CIKM.pdf
29. Poesina, E., Costache, A.V., Chifu, A.G., Mothe, J., Ionescu, R.T.: PQPP: A Joint Benchmark for Text-to-Image Prompt and Query Performance Prediction. arXiv:2406.04746 (2024), https://arxiv.org/html/2406.04746v1
30. Poesina, E., Ionescu, R.T., Mothe, J.: IQPP: A Benchmark for Image Query Performance Prediction. In: SIGIR. pp. 2953–2963 (2023), https://hal.science/hal-04346953v1/file/2302.10126.pdf
31. Pradeep, R., Sharifymoghaddam, S., Lin, J.: RankVicuna: Zero-Shot Listwise Document Reranking with Open-Source Large Language Models. arXiv:2309.15088 (2023)
32. Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., et al.: Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. arXiv:2306.17563 (2023)
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. JMLR **21**(140), 1–67 (2020)
34. Scells, H., Azzopardi, L., Zuccon, G., Koopman, B.: Query variation performance prediction for systematic reviews. In: SIGIR. pp. 1089–1092 (2018)
35. Tian, X., Jia, Q., Mei, T.: Query Difficulty Estimation for Image Search with Query Reconstruction Error. IEEE Transactions on Multimedia **17**(1), 79–91 (2014)