Report on the 2nd Workshop on Query Performance Prediction and its Applications in the Era of Large Language Models (QPP++ 2025) at ECIR 2025

Chuan Meng
University of Amsterdam
The Netherlands
c.meng@uva.nl

Guglielmo Faggioli
University of Padova
Italy
guglielmo.faggioli@dei.unipd.it

Mohammad Aliannejadi
University of Amsterdam
The Netherlands
m.aliannejadi@uva.nl

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

Josiane Mothe
Université de Toulouse
France
josiane.mothe@irit.fr

Abstract

Query performance prediction (QPP) is a key and long-standing task in information retrieval (IR). The task of QPP aims to predict ranking quality without human relevance judgments. The rise of large language models (LLMs) has significantly reshaped the IR research landscape. These advancements call for a timely discussion on how we perform, evaluate and apply QPP in the LLM era. To foster this discussion, we have organised the workshop on QPP at the 47th European Conference on Information Retrieval (ECIR 2025). While LLM-related topics were a major focus, we made sure to keep our workshop open to a broad range of QPP-related research areas. The workshop featured five accepted papers and gathered around 20 participants for a half-day of presentations and interactive discussions. These discussions produced several key insights and directions for future QPP research.

Date: 10 April 2025.

Website: https://qppworkshop.github.io/.

1 Motivation

Query performance prediction (QPP) has been studied in the information retrieval (IR) communities for decades [Carmel et al., 2005; Mothe and Tanguy, 2005; Carmel and Yom-Tov, 2010; Arabzadeh et al., 2025]. The task of QPP aims to predict the ranking quality of a search system without relying on human-annotated relevance judgments [Mizzaro et al., 2018; Meng et al., 2025a]. Recently, the rise of large language models (LLMs) has significantly reshaped the IR

research landscape [Zhu et al., 2023; Li et al., 2024], giving rise to new research directions, including LLM-based retrievers/re-rankers [Sun et al., 2023; Ma et al., 2024; Meng et al., 2024], and generative AI systems [Allan et al., 2024; White and Shah, 2025]. These advancements call for a timely discussion on how we perform, evaluate and apply QPP in the LLM era. To foster this discussion, we have organised the workshop *Query Performance Prediction and its Applications in the Era of Large Language Models (QPP++ 2025)* [Meng et al., 2025b], at the 47th European Conference on Information Retrieval (ECIR 2025) in Lucca, Italy. QPP++ 2025 is a continuation of the QPP++ 2023 workshop [Faggioli et al., 2023a,b], held at the 45th European Conference on Information Retrieval (ECIR 2023) in Dublin, Ireland.

2 Workshop organisation

Although our call for papers highlighted LLM-related topics, such as predicting the performance of LLM-based retrievers and re-rankers, assessing generative AI systems, and using LLMs to enhance QPP methods, we also designed the workshop to remain open to a broad range of QPP research directions. These include, but are not limited to, practical applications of QPP [Ganguly and Yilmaz, 2023], multimodal QPP [Tian et al., 2014; Poesina et al., 2023], multilingual QPP, and QPP for conversational search [Faggioli et al., 2023c,d; Meng et al., 2023b,a].

We welcomed both original and previously published work for submission. After review, five submissions were accepted for presentation. The authors of the accepted submissions are affiliated with eleven institutions across six countries, spanning Europe, North America, and Asia. The accepted papers include two original contributions [Parry et al., 2025; Tian et al., 2025b] and three previously published studies [Ebrahimi et al., 2024; Saleminezhad et al., 2025; Poesina et al., 2025]. The proceedings of the QPP++ 2025 Workshop are available on the workshop website.[1]

The accepted papers cover a range of topics. There are three papers focusing on improving QPP modelling in text-based scenarios. Tian et al. [2025b] aim to enhance QPP by leveraging query variants. To avoid the hallucinations or information drifts introduced by query variant generation, Tian et al. [2025b] propose to retrieve similar queries from an external training set. To ensure high recall in retrieving queries with information needs most similar to the target query from the training set, Tian et al. [2025b] propose a two-hop retrieval approach: first, retrieving similar queries from the training set, followed by a second retrieval step using the relevant documents associated with those queries. Saleminezhad et al. [2025] propose a QPP method that measures robustness to query perturbations for automatically predicting the ranking quality of dense retrievers. Ebrahimi et al. [2024] propose a QPP method that learns to map contextual information related to the target query to an IR evaluation measure. Given the target query, the signals include its retrieved documents, and similar queries from a training set with ground-truth ranking performance. Beyond text-based scenarios, Poesina et al. [2025] focus on both text-to-image generation and retrieval, and manually create a joint benchmark for prompt and query performance prediction (PQPP). Instead of focusing on the general QPP task, which aims to predict the ranking quality for a specific query, Parry et al. [2025] propose corpus performance prediction (CPP), which aims to predict the ranking quality of a corpus for a given query class.

[1] https://qppworkshop.github.io/

CPP can be used to assess the utility of a corpus for a given query class, for example, to support corpus selection based on query type.

On the day of the workshop, QPP++ 2025 gathered around 20 participants. We began with paper presentation sessions, followed by a keynote, and an interactive discussion session (details of the discussion will be provided in the next section). The keynote, titled *"The Role of Query Performance Prediction in Developing Adaptive Search and RAG Systems"*, was delivered by Debasis Ganguly. The keynote first gave an overview of existing unsupervised and supervised QPP methods, and then introduced the potential of using QPP for adaptive IR and retrieval-augmented generation (RAG) systems. For adaptive IR, Debasis highlighted the importance of query-specific treatment and mentioned that selective relevance feedback [Datta et al., 2024] is a scenario where QPP can be applied. Turning to RAG systems, Debasis pointed out that QPP has the potential to guide the adjustment of RAG hyperparameters, such as the number of retrieved documents used for generation. He also noted that recent work shows the relevance of retrieved documents does not necessarily translate into gains in downstream text generation [Tian et al., 2025a]. This observation shifts the focus from relevance alone to the broader goal of enhancing the *utility* of retrieved documents for generation. As such, QPP might need to go beyond predicting the relevance of retrieved documents and instead aim to estimate their impact on the final generated output. Ultimately, feedback from QPP predictors can be used to improve the effectiveness of RAG systems.

3 Discussion outcomes

QPP++ 2025 featured a breakout discussion session, in which all participants were divided into two groups to discuss the applications and evaluation of QPP, followed by a collective sharing of discussion outcomes. These discussions produced the following insights and directions for future QPP research.

QPP for RAG. QPP has the potential to benefit RAG on multiple aspects. First, predicting whether to perform retrieval or generate directly for a given user query. For example, if the retrieved documents are predicted to be of low quality, the system may choose not to use them and instead respond with an "I don't know" message to the user. Second, exploring the use of QPP to predict how many documents should be fed into generation is an important research direction. Adding more retrieved documents for generation may introduce more noisy information and reduce generation efficiency. Therefore, choosing an adaptive number of documents can help achieve a better effectiveness–efficiency trade-off for RAG. Third, shifting from topical relevance to utility. Given existing findings that the topical relevance of retrieved documents shows limited correlation with downstream text generation quality [Tian et al., 2025a], it implies that predicted ranking quality may not well reflect generation quality. Therefore, adapting QPP methods to predict downstream text generation quality is an important research topic. Fourth, recent work has also explored modelling RAG with reasoning based on chain-of-thought prompts [Li et al., 2025]. An interesting research direction is how to incorporate QPP results into chain-of-thought reasoning to improve the overall reasoning process.

QPP for agentic workflows. Agentic workflows have received attention recently [Zhang et al., 2024; Singh et al., 2025; Zhang et al., 2025]. Unlike static workflows following a fixed pipeline for all user queries, an agentic workflow is a workflow in which one or multiple autonomous agents

dynamically adjust execution paths to each complex user query. QPP has the potential to enhance agentic workflows in several ways. First, QPP can act as an autonomous agent to automatically determine an appropriate execution path given a user request. In the RAG scenarios mentioned in the previous paragraph, QPP can be used to decide whether to use RAG or generation-only. Beyond RAG, QPP can also support iterative refinement of ranking results. For example, it can trigger additional refinement steps such as rewriting the query to expand the original user request, or using a different ranker or corpus. QPP can also decide to return the ranking results to the user or a text generator only when the predicted quality is sufficiently high.

QPP for multi-modality scenarios. Compared to QPP in text-based scenarios, research on QPP in multi-modality settings is still limited. More studies are needed on QPP for tasks such as text-to-image search [Tian et al., 2014], image-to-image search [Poesina et al., 2023], and text-to-image generation [Poesina et al., 2025], as well as in emerging contexts like video search and generation.

QPP for conversational systems. Nowadays, everything is becoming interactive, highlighting the growing importance of conversational systems [Mo et al., 2024b,a]. Given the limited research on QPP for conversational systems, such as conversational search [Faggioli et al., 2023c,d], more studies are expected on QPP for conversational systems.

QPP for other applications. It is important to demonstrate to the broader community how QPP can benefit other applications involving query-specific treatments. For example, QPP has the potential to predict query-specific beam sizes in generative retrieval [Li et al., 2024], or to predict per-query re-ranking depths for re-ranking tasks [Meng et al., 2024].

QPP evaluation: from query sorting to downstream tasks. The current common practice for evaluating QPP is to measure correlation coefficients (such as Pearson's ρ and Kendall's τ) between the actual and predicted performance of a set of queries. A QPP method achieving high correlation scores only indicates that it ranks a set of queries well according to their relative difficulty [Raiber and Kurland, 2014]. However, a QPP method that performs well at "query sorting" does not necessarily produce good performance on downstream tasks. Raiber and Kurland [2014] point out that the gap between QPP evaluation and the goals of downstream tasks can limit the practical application of QPP, because the objective of downstream tasks is often not to rank a set of queries by their relative difficulty. Indeed, existing studies have shown that directly applying QPP to downstream tasks often leads to limited benefits. This has been observed in applications such as retriever selection [Khramtsova et al., 2024], rank fusion [Raiber and Kurland, 2014], IR system configuration selection [Deveaud et al., 2018], and clarification need prediction [Arabzadeh et al., 2022]. Even for the task of selecting the best-performing query variant that shares the same information need, essentially a query sorting problem, existing work [Thomas et al., 2017; Scells et al., 2018; Zendel et al., 2021] has found that QPP methods have limited effectiveness. These findings suggest that the evaluation of QPP should be reconsidered. In the future, it is important to assess QPP methods from the perspective of their impact on downstream tasks.

4 Conclusion

The QPP++ 2025 workshop at ECIR 2025 went well. The discussions brought forward useful insights and ideas for future QPP research. One main takeaway is that QPP should be explored in a wider range of applications in IR and natural language processing, such as RAG, agentic

workflows, multi-modal systems, and conversational AI. QPP will attract more attention as more people realise its usefulness. Another important point is to rethink how we evaluate QPP methods: moving beyond simple "query sorting" and focusing more on how these methods help in downstream tasks.

Acknowledgments

We thank ECIR 2025 for hosting the workshop. We are also grateful to the following programme committee members for their time and effort in reviewing and selecting papers: Adrian-Gabriel Chifu, Claudia Hauff, David Carmel, Francesco Luigi De Faveri, Laure Soulier, Laura Menotti, Maik Fröbe, Md Zia Ullah, Oleg Zendel, and Radu Tudor Ionescu. We thank all authors for their submissions, and all participants for their contributions to the discussion. Special thanks to Debasis Ganguly for delivering the keynote talk.

References

James Allan, Eunsol Choi, Daniel P Lopresti, and Hamed Zamani. Future of information retrieval research in the age of generative AI. *arXiv preprint arXiv:2412.02043*, 2024.

Negar Arabzadeh, Mahsa Seifikar, and Charles LA Clarke. Unsupervised question clarity prediction through retrieved item coherency. In *CIKM*, pages 3811–3816, 2022.

Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. Query performance prediction: Theory, techniques and applications. In *WSDM*, pages 991–994, 2025.

David Carmel and Elad Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

David Carmel, Elad Yom-Tov, and Ian Soboroff. Sigir workshop report: Predicting query difficulty - methods and applications. In *ACM SIGIR Forum*, volume 39, pages 25–28. ACM New York, NY, USA, 2005.

Suchana Datta, Debasis Ganguly, Sean MacAvaney, and Derek Greene. A deep learning approach for selective relevance feedback. In *ECIR*, pages 189–204, 2024.

Romain Deveaud, Josiane Mothe, Md Zia Ullah, and Jian-Yun Nie. Learning to adaptively rank document retrieval system configurations. *TOIS*, 37(1):1–41, 2018.

Sajad Ebrahimi, Maryam Khodabakhsh, Negar Arabzadeh, and Ebrahim Bagheri. Estimating query performance through rich contextualized query representations. In *ECIR*, pages 49–58, 2024.

Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, and Fiana Raiber. QPP++ 2023: Query-performance prediction and its evaluation in new tasks. In *ECIR*, pages 388–391. Springer, 2023a.

Guglielmo Faggioli, Nicola Ferro, Josiane Mothe, Fiana Raiber, and Maik Fröbe. Report on the 1st workshop on query performance prediction and its evaluation in new tasks (QPP++ 2023) at ECIR 2023. In *ACM SIGIR Forum*, volume 57, pages 1–7, 2023b.

Guglielmo Faggioli, Nicola Ferro, Cristina Muntean, Raffaele Perego, and Nicola Tonellotto. A spatial approach to predict performance of conversational search systems. In *IIR*, pages 41–46, 2023c.

Guglielmo Faggioli, Nicola Ferro, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto. A geometric framework for query performance prediction in conversational search. In *SIGIR*, page 1355–1365, 2023d.

Debasis Ganguly and Emine Yilmaz. Query-specific variable depth pooling via query performance prediction. In *SIGIR*, pages 2303–2307, 2023.

Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, and Guido Zuccon. Leveraging LLMs for unsupervised dense retriever ranking. In *SIGIR*, pages 1307–1317, 2024.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025.

Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. A survey of generative search and recommendation in the era of large language models. *arXiv preprint arXiv:2404.16924*, 2024.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning LLaMA for multi-stage text retrieval. In *SIGIR*, pages 2421–2425, 2024.

Chuan Meng, Mohammad Aliannejadi, and Maarten de Rijke. Performance prediction for conversational search using perplexities of query rewrites. In *QPP++2023*, pages 25–28, 2023a.

Chuan Meng, Negar Arabzadeh, Mohammad Aliannejadi, and Maarten de Rijke. Query performance prediction: From ad-hoc to conversational search. In *SIGIR*, page 2583–2593, 2023b.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Ranked list truncation for large language model-based re-ranking. In *SIGIR*, 2024.

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Query performance prediction using relevance judgments generated by large language models. *TOIS*, 2025a.

Chuan Meng, Guglielmo Faggioli, Mohammad Aliannejadi, Nicola Ferro, and Josiane Mothe. QPP++ 2025: Query performance prediction and its applications in the era of large language models. In *ECIR*, pages 319–325, 2025b.

Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. Query performance prediction and effectiveness evaluation without relevance judgments: Two sides of the same coin. In *SIGIR*, pages 1233–1236, 2018.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. A survey of conversational search. *arXiv preprint arXiv:2410.15576*, 2024a.

Fengran Mo, Chen Qu, Kelong Mao, Yihong Wu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. Aligning query representation with rewritten query and relevance judgments in conversational search. In *CIKM*, pages 1700–1710, 2024b.

Josiane Mothe and Ludovic Tanguy. Linguistic features to predict query difficulty. In *ACM Conference on Research and Development in Information Retrieval, SIGIR, Predicting query difficulty-methods and applications workshop*, pages 7–10, 2005.

Andrew Parry, Jan Heinrich Merker, Simon Ruth, Maik Fröbe, and Harrisen Scells. Corpora performance prediction. In *QPP++ 2025*, 2025.

Eduard Poesina, Radu Tudor Ionescu, and Josiane Mothe. IQPP: A benchmark for image query performance prediction. In *SIGIR*, pages 2953–2963, 2023.

Eduard Poesina, Adriana Valentina Costache, Adrian-Gabriel Chifu, Josiane Mothe, and Radu Tudor Ionescu. PQPP: A joint benchmark for text-to-image prompt and query performance prediction. In *CVPR*, 2025.

Fiana Raiber and Oren Kurland. Query-performance prediction: Setting the expectations straight. In *SIGIR*, pages 13–22, 2014.

Abbas Saleminezhad, Negar Arabzadeh, Radin Hamidi Rad, Soosan Beheshti, and Ebrahim Bagheri. Robust query performance prediction for dense retrievers via adaptive disturbance generation. *Machine Learning*, 114(3):1–23, 2025.

Harrisen Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. Query variation performance prediction for systematic reviews. In *SIGIR*, pages 1089–1092, 2018.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *EMNLP*, pages 14918–14937, 2023.

Paul Thomas, Falk Scholer, Peter Bailey, and Alistair Moffat. Tasks, queries, and rankers in pre-retrieval performance prediction. In *ADCS*, pages 1–4, 2017.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. Is relevance propagated from retriever to generator in RAG? In *ECIR*, pages 32–48, 2025a.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. Revisiting query variants: The advantage of retrieval over generation of query variants for effective QPP. In *QPP++ 2025*, 2025b.

Xinmei Tian, Qianghuai Jia, and Tao Mei. Query difficulty estimation for image search with query reconstruction error. *IEEE Transactions on Multimedia*, 17(1):79–91, 2014.

Ryen W White and Chirag Shah. *Information Access in the Era of Generative AI*. Springer, 2025.

Oleg Zendel, J Shane Culpepper, and Falk Scholer. Is query performance prediction with multiple query variations harder than topic performance prediction? In *SIGIR*, pages 1713–1717, 2021.

Guibin Zhang, Kaijie Chen, Guancheng Wan, Heng Chang, Hong Cheng, Kun Wang, Shuyue Hu, and Lei Bai. EvoFlow: Evolving diverse agentic workflows on the fly. *arXiv preprint arXiv:2502.07373*, 2025.

Weinan Zhang, Junwei Liao, Ning Li, and Kounianhua Du. Agentic information retrieval. *arXiv preprint arXiv:2410.09713*, 2024.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.