

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Chakrabarti, H., & Pera, M. S. (2026). Query Performance Prediction Using a Child-Focused Definition of Relevance. In R. Campos, A. Jatowt, Y. Lan, M. Aliannejadi, C. Bauer, S. MacAvaney, A. Anand, N. Bai, M. Mansoury, Z. Ren, & S. Verberne (Eds.), *Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Proceedings* (pp. 388-399). (Lecture Notes in Computer Science; Vol. 16484 LNCS). Springer Nature. https://doi.org/10.1007/978-3-032-21300-6_29

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.



Query Performance Prediction Using a Child-Focused Definition of Relevance

Hrishita Chakrabarti^(✉)  and Maria Soledad Pera 

Delft University of Technology, Delft, Netherlands
{h.chakrabarti,m.s.pera}@tudelft.nl

Abstract. Query performance prediction (QPP) methods have primarily been tailored to mainstream users, thus relying on the traditional concept of relevance. In the case of children, however, relevance goes beyond content-based resource-query matching, which is why we gauge the performance of existing QPP methods in estimating the *fit* of resources retrieved in response to child-formulated queries. Outcomes from our empirical exploration of various QPP methods using a traditional and a child-focused definition of relevance on 2 datasets reveal the limitations in the adaptability of existing methods to the context of child information retrieval.

Keywords: Query Performance Prediction · Children

1 Introduction

Children regularly turn to popular search engines (**SE**), such as Google and Bing, for information discovery [4, 6, 18]. Prior research indicates that these search tools often retrieve resources that fail to adequately address children’s inquiries [22, 66], and that even the slightest change in a child’s query could lead to overly complex, misleading, or even inappropriate content, impacting a child’s success in completing their search task [55]. Given that the information children encounter can shape their development and worldview [18, 56], it is crucial to understand *when* a system may fail to serve this population effectively.

Query Performance Prediction (**QPP**) estimates the effectiveness of an information retrieval (**IR**) system based on a user’s query [8, 15, 33, 46]. In essence, QPP methods help estimate how difficult it would be for an IR system to retrieve relevant resources for a given query [8]. In the realm of children IR [53], QPP could be used to help identify children’s queries that an IR system is likely to address poorly. Over the years, research efforts have yielded a wide range of methods, from simple lexical techniques to advanced neural-based and Large Language Model (**LLM**)-based approaches [e.g., 15, 19, 46, 61, 63, 68, 71, 73]. However, to the best of our knowledge, existing methods are designed, optimised, and evaluated based on standard IR datasets that reflect the needs of mainstream users—primarily adults—and adopt a traditional definition of relevance based

on topical similarity between queries and retrieved items [12, 13]. QPP, therefore, has generally been approached from a system-centred perspective, focusing on query–system interactions, largely neglecting the user formulating the query. From a user-centred perspective, however, retrieval quality depends on the degree to which the retrieved resources befit the user for satisfying their information needs. This perspective is particularly important for children, whose cognitive abilities, digital knowledge, and information needs differ inherently from those of adults, affecting how they evaluate and interpret the information they are exposed to [38, 53, 67]. This raises questions about the extent to which existing QPP methods can generalise to children’s information-seeking context.

In this work, we empirically explore the performance of QPP methods in estimating the relevance of retrieved resources as it pertains to children. Considering the broad range of cognitive skills and understanding of the world of the target audience, we scope our exploration to children up to grade 5 (i.e., up to 11 years old) and define *child-focused* relevance based on: (1) *topicality*, as a relevant resource must contain content that matches or is related to the information need expressed in a child’s query, (2) *readability*, since children need to understand the resource’s content to be able to use it to address their information need, and (3) *educational alignment*, to ensure that the content befits children’s domain knowledge. We use 2 datasets [28, 42, 44] to analyse the effectiveness and robustness of 18 QPP methods—from simple pre- and post-retrieval methods to state-of-the-art—when using a traditional and child-focused variant of relevance.

From our analysis, none of the studied QPP methods, including state-of-the-art supervised approaches, adequately account for children’s needs when estimating system performance. These findings underscore a critical issue: the lack of non-mainstream user representation in QPP research and its broader impact on information access for underserved and vulnerable populations. To facilitate future work in this direction and for reproducibility purposes, we share our code in a  public GitHub repository: <https://github.com/SOLandChildren/ECIR2026>.

2 Experimental Set Up

In our study, we consider two *definitions of relevance*: 1. **Traditional (TR)**: Following IR standards, the relevance of a resource is judged based on its *topicality* [51], i.e., the extent to which the content in a resource matches or relates to the information need expressed in a user’s query. 2. **Child-focused (CR)**: A resource is relevant to a child if it not only relates to their information need, but also the child can understand the presented content to distil information from it [40]. Thus, along with topicality, we consider *readability* to ascertain if a child could comprehend the content they are exposed to, and *educational alignment* to assess if the resource befits a child’s domain knowledge skill level.

We use two *datasets*. *kid-friend* [28] includes 50 adult-formulated queries about personal, political, educational, and entertainment information needs of 4th grade children (up to 10 years old), 2385 web resources—each represented by

title, snippet, and full text—and 2303 query-resource pairs labelled by human annotators for topicality (1 relevant, 0 otherwise). We extend this labelling into a binary “*relevance vector*” $< t, r, e >$ representing topicality, readability, and educational alignment. For t , we retain the original annotation; r is 1 if the resource’s readability as per the Spache-Allen formula [1] is ≤ 4 (i.e., for grade 4), 0 otherwise; e is 1 if the resource’s educational value computed using the **fine-web-edu** classifier [52] is ≥ 3 on a 0–5 scale (corresponding to primary/grade school content [36, 52]), 0 otherwise. Resources with $t = 0$ are treated as non-relevant ($< 0, 0, 0 >$). In this extension, 114 query-resource pairs are labelled as relevant across all three components, 405 based on topicality and either readability or educational alignment, 1011 only by topicality, and 773 non-relevant. **requik** [42, 43] consists of 82 queries formulated by children from grades 3–5 (ages 9–11) on classroom-related topics, 317 web resources (title and snippet), 413 human-annotated query-resource pairs. As these annotations indicate suitability for classroom use, relevant resources (92 out of 413) are labelled as $< 1, 1, 1 >$; non-relevant ones as $< 0, 0, 0 >$.

We examine baseline and state-of-the-art **QPP methods** categorized as: (1) *pre-retrieval*—QS [34], Average Pointwise Mutual Information (PMIavg) [33], Query Length (QL) [48], VARavg [71], Inverse Document Frequency (IDFavg) [15], SCQavg [71], Average Inverse Corpus Term Frequency (ICTFavg) [34], Simplified Clarity Score (SCS) [34]; (2) *post-retrieval*—Clarity [15], Weighted Information Gain (WIG) [61], Normalized Query Commitment (NQC) [61], Score Magnitude and Variance (SMV) [63], σ_{max} [57], and $n(\sigma_x\%)$ [17], and (3) *neural-based*—NQAQPP [31], BERTQPP [3], and qppBERT-PL [20] (3 BERT-based methods reported to outperform other neural approaches [3, 11, 20, 31]), as well as an LLM-based method, QPP-GenRE [46]. Method descriptions are in the original works; parameters and implementation details are in our [repository](#). To train neural-based methods, due to scarce datasets labelled for child relevance, we created **clear_qpp**, using the CLEAR corpus [16] (queries and web resources for grades 3–12, ages 8–18) and the SQuAD dataset [60] (queries and web resources from mainstream users) **clear_qpp** includes 1500 queries (1000 randomly sampled from CLEAR and 500 from SQuAD), and 3010 query-sample pairs labelled with relevance vectors. A pair is labelled 1 for t if present in the original dataset, 0 otherwise; r and e are assigned as done in **kid-friend**.

For evaluation purposes, following common practice [3, 8, 32], the considered QPP methods predict query performance when retrieving 1000 documents using a standard **first-stage retriever** BM25 [37]. To assess the effectiveness of a QPP method, we compute the correlation between a set of queries ranked by their predicted and actual performance [8]; the latter measured using the retriever’s NDCG@10 score per query¹. We use Pearson’s ρ for linear correlation and Kendall’s τ for ranking correlation. (We omitted Spearman’s ρ as it showed similar trends). Higher coefficients indicate more accurate predictions. As corre-

¹ To compute NDCG@10 based on **CR**, the relevance score for a query-resource pair is the sum of its vector components. For instance, for a pair labelled $< 1, 0, 1 >$, its relevance score is 2. For **TR**, the relevance score is the topicality label.

lation does not capture error magnitudes across queries, we use the scaled Mean Absolute Ranking Error (**sMARE**) [27], which is the average difference in the ranks of a query, based on its predicted and actual performance; lower sMARE indicates higher accuracy. For consistency in our analysis, we use the inverted version from [49]. To measure robustness across queries, we use risk sensitivity: the extent to which a QPP method can predict performance for most queries without predicting significantly worse performance for others, compared to one (or more) baseline methods [49]. To avoid bias from selecting a single baseline [21], we treat all methods as baselines for each other and use the **GeoRisk** metric ($\alpha = 5$), due to its effectiveness in discerning a QPP method’s robustness when compared to multiple baselines using different datasets [49]; a higher GeoRisk score signals higher risk sensitivity. Since these metrics return a single point value per method, limiting statistical analysis, we follow [49] and average 1000 bootstrap samples across all queries in a dataset.

3 Results and Discussion

To probe QPP methods in the context of children’s information seeking, we examine their performance *across relevance definitions*. For this, we turn to the scores yielded by each method for **kid-friend** using **TR** and **CR**, reported in Table 1 (Student’s T-test, $p < 0.05$). From both effectiveness and robustness perspectives, **QS**, **PMIavg**, **VARavg**, **SCQavg**, **ICTFavg**, **IDFavg**, **SCS**, **Clarity**, and $n(\sigma_x\%)$ perform significantly better for **CR** than **TR**; vice versa for **NQAQPP** and **BERTQPP**. The remaining methods exhibit inconsistent trends. For instance, **WIG**, **NQC**, and **SMV** are more effective but less robust for **CR** than **TR** (**NQC**’s sMARE scores not significantly different), and σ_{max} is significantly more effective for **CR** than for **TR**, but equally robust for both definitions. We also examine QPP methods as a whole, seeking to identify which ones are best-suited for a particular relevance definition. Comparing all methods per metric (ANOVA $p < 0.05$ for statistical significance), we note that post-retrieval methods—except **Clarity**—perform significantly better for both **TR** and **CR** compared to the others. Yet, Figs. 1a and 1b, showing score-based rankings per metric, reveal that no single post-retrieval method consistently outperforms across all metrics.

Akin to an ablation study, we assess the contribution of each relevance factor using: (1) **TR+r**, which considers topicality and resource readability, and (2) **TR+e**, focusing on topicality and educational alignment. Pairwise comparisons of the scores of each method for **TR**, **TR+r**, **TR+e**, and **CR** reported in Table 1 (Student’s T-test $p < 0.05$ with Bonferroni’s correction, $\alpha = 0.05$, $N = 6$), indicate that difference across the relevance variations is most prominent for Pearson’s correlation, which is why we use that metric to illustrate salient trends in Fig. 1d which depicts QPP methods ranked per relevance variation based on Pearson’s correlation scores. We use colour-coding for visualising the ranks per column, with dark green denoting the highest-ranked method, and dark red the lowest-ranked. From the figure, we observe that pre- and post-retrieval methods perform significantly better for **TR+e** and significantly worse for **TR+r** relative to **TR**, except for **QL**, **NQC**, and **SMV**, which exhibit the opposite trend.

Table 1. Performance across TR and CR on **kid-friend**. ^{a,b,c} for significance (Student’s T-test, $p < 0.05$) w.r.t. corresponding metric score for TR, TR+r, and TR+e, resp.; * denotes significance w.r.t. to all three.

QPP	TR				TR+r				TR+e				CR			
	P- ρ	K- τ	sMARE	GeoRisk	P- ρ	K- τ	sMARE	GeoRisk	P- ρ	K- τ	sMARE	GeoRisk	P- ρ	K- τ	sMARE	GeoRisk
QL	0.143	0.099	0.686	0.458	0.386*	0.267*	0.74*	0.479*	0.178 ^{a,b}	0.099*	0.696 ^{a,b}	0.455 ^{a,b}	0.115*	0.1 ^{a,b}	0.694*	0.457*
QS	-0.193	-0.078	0.63	0.453	-0.393*	-0.242*	0.602*	0.428*	-0.13 ^{a,b}	-0.055 ^{a,b}	0.652 ^{a,b}	0.454 ^b	-0.065*	-0.035*	0.655*	0.459*
PMIavg	0.094	0.128	0.708	0.486	-0.046*	-0.003*	0.671*	0.478*	0.182 ^{a,b}	0.152 ^{a,b}	0.72 ^{a,b}	0.495 ^{a,b}	0.2*	0.155*	0.718*	0.495 ^{a,b}
VARavg	-0.087	-0.048	0.636	0.468	-0.269*	-0.191*	0.611*	0.44*	-0.041 ^{a,b}	-0.016 ^{a,b}	0.658 ^{a,b}	0.466 ^{a,b}	0.038*	0.012*	0.662*	0.471*
SCQavg	-0.125	-0.068	0.627	0.463	-0.339*	-0.215*	0.604*	0.434*	-0.085 ^{a,b}	-0.045 ^{a,b}	0.651 ^{a,b}	0.46 ^{a,b}	0.007*	-0.015*	0.652*	0.466*
ICTFavg	-0.054	0.028	0.685	0.47	-0.196*	-0.084*	0.659*	0.449*	-0.006 ^{a,b}	0.026*	0.694 ^{a,b}	0.468 ^{a,b}	0.119*	0.11*	0.714*	0.481*
IDFavg	-0.121	-0.061	0.64	0.472	-0.314*	-0.202*	0.612*	0.444*	-0.078 ^{a,b}	-0.048 ^{a,b}	0.654 ^{a,b}	0.466 ^{a,b}	0.036*	0.001*	0.664*	0.476*
SCS	-0.137	-0.065	0.646	0.449	-0.348*	-0.205*	0.621*	0.421*	-0.116 ^{a,b}	-0.059 ^{a,b}	0.652 ^{a,b}	0.446 ^{a,b}	0.004*	-0.006*	0.668*	0.459*
Clarity	-0.061	-0.041	0.643	0.451	-0.283*	-0.212*	0.613*	0.425*	0.036 ^{a,b}	0*	0.661 ^{a,b}	0.457 ^{a,b}	0.095*	0.02*	0.662*	0.462*
WIG	0.22	0.185	0.711	0.517	0.208*	0.135*	0.707*	0.512*	0.225 ^{a,b}	0.147*	0.699 ^{a,b}	0.506 ^{a,b}	0.311*	0.205*	0.719*	0.507*
NQC	0.291	0.205	0.723	0.515	0.367*	0.218*	0.724	0.508*	0.257 ^{a,b}	0.156 ^{a,b}	0.708 ^{a,b}	0.5 ^{a,b}	0.362 ^{a,c}	0.232*	0.724*	0.508 ^{a,c}
SMV	0.265	0.187	0.719	0.515	0.351*	0.2*	0.717	0.507*	0.247 ^{a,b}	0.133 ^{a,b}	0.704 ^{a,b}	0.499 ^{a,b}	0.353*	0.223*	0.72*	0.508*
σ_{max}	0.227	0.228	0.729	0.521	0.217*	0.165*	0.712*	0.519*	0.288 ^{a,b}	0.22*	0.732 ^{a,b}	0.523 ^{a,b}	0.337*	0.245*	0.739*	0.521*
$n(\sigma_{x\%})$	0.12	0.115	0.695	0.496	-0.036*	0.003*	0.666*	0.47*	0.192 ^{a,b}	0.125 ^{a,b}	0.704 ^{a,b}	0.498 ^{a,b}	0.268*	0.159*	0.712*	0.505*
NQAQPP	-0.055	-0.008	0.676	0.442	-0.136*	-0.084*	0.641*	0.433*	-0.045 ^{a,b}	-0.033 ^{a,b}	0.663 ^{a,b}	0.438 ^{a,b}	-0.07*	-0.02*	0.662 ^{a,b}	0.438 ^{a,b}
BERTQPP	-0.012	-0.001	0.681	0.446	-0.117*	-0.048*	0.671*	0.443*	-0.042 ^{a,b}	-0.039 ^{a,b}	0.667 ^{a,b}	0.439 ^{a,b}	-0.076*	-0.041 ^{a,b}	0.662*	0.434*
qppBERT-PL	-0.338	-0.16	0.64	0.433	-0.255*	-0.153	0.639*	0.437*	-0.345 ^{a,b}	-0.221 ^{a,b}	0.632 ^{a,b}	0.421 ^{a,b}	-0.242*	-0.175*	0.635*	0.427*
QPP-GenRE	0.000	0.000	0.743	0.531	0.000	0.000	0.729*	0.517*	0.000	0.000	0.726 ^{a,b}	0.515 ^{a,b}	-0.083	0.025	0.685*	0.444*

Recall that queries in **kid-friend** are formulated by adults, which may not capture children’s distinct formulation patterns [29,65]. With queries playing a pivotal role in QPP performance, we gauge the applicability of QPP for children in a realistic scenario by using child-formulated queries from **requik**. Comparing results reported in Table 2 with **kid-friend** for estimating performance based on CR (Student’s T-test $p < 0.05$), we note that regardless of the metric considered, QL and σ_{max} perform better on **kid-friend** than **requik**. VARavg, SCQavg, IDFavg, SCS, NQAQPP, and qppBERT-PL score higher for **requik** than **kid-friend**, although the differences in correlation scores for qppBERT-PL are not significant. The remaining methods exhibit inconsistent trends across metrics. For example, QS and ICTFavg are significantly more effective but less robust for **requik** than **kid-friend**. Conversely, Clarity is significantly less effective but more robust for **kid-friend**. We investigate whether a method or category is effective and robust on both datasets by comparing the scores for a metric yielded for each method per dataset (Table 2, ANOVA, $p < 0.05$). All methods significantly differ between datasets on all metrics; post-retrieval methods, except Clarity, perform the best across all metrics for **kid-friend**, whereas for **requik**, two pre-retrieval methods—VARavg and SCQavg—outperform the post-retrieval methods. Note that differences in performance are more prominent in **requik** than **kid-friend**. This is evident in Figs. 1b and 1c in which we rank the methods based on their scores for all metrics, with ranking positions per column colour-coded from dark green (highest) to dark red (lowest).

Overall, based on our analysis of QPP performance across definitions and datasets, we find that post-retrieval methods outperform remaining categories for TR and CR when applied to queries formulated by adult proxies reflecting trends reported in previous literature [e.g., 26,70]; but for child-formulated queries, we observe the opposite, with two simple pre-retrieval strategies surpassing

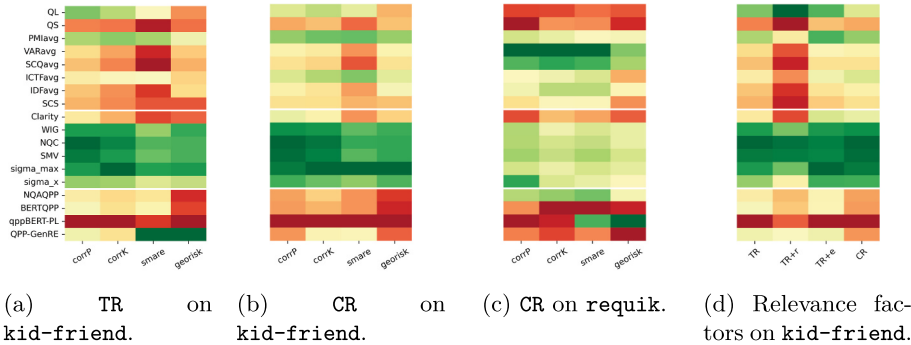


Fig. 1. QPP method-wise ranking across relevance definitions, relevance factors, and datasets; highest ranked per column is dark green, lowest dark red. (Color figure online)

post-retrieval counterparts. Depending on the dataset, some methods demonstrate higher effectiveness, whereas others show more robustness, highlighting a trade-off between these perspectives. Several methods that on the surface exhibit better performance for child-focused relevance, struggle to leverage readability when estimating retrieval quality—a key factor in children’s ability to distil the information from online content [2, 44]. Interestingly, none of the state-of-the-art supervised methods outperformed the simpler unsupervised counterparts across any metric, relevance definition, or dataset. This suggests that neural-based models are unable to learn lexical patterns that signal retrieval quality for children’s queries. Similar shortcomings have been observed on BERT-based methods when applied to the conversational search scenario [45], where model performance dropped for complex natural-language queries—a formulation commonly employed by children [29, 65]. Although QPP-GenRE (LLM-based) can adapt to conversational scenarios [46], it does not generalize to our case. We attribute the poor performance of BERT and LLM methods on child-specific datasets to the underlying language models that are trained on data from mainstream users [7, 14]. Consequently, these models encode text representations biased towards mainstream perspectives, ignoring topics and terminology unique to underserved groups such as children.

QPP has been extensively studied for several decades [8, 10, 23], with research spanning strategies for different types of rankers [26], QPP for conversational search [24, 25], multi-modal search scenarios [59, 64], and optimising retrieval augmented generation (RAG) [47]. Nonetheless, our findings expose a critical gap: existing methods consistently underperform in scenarios involving children, raising broader questions about the applicability of QPP to underserved groups [53, 54]. Advancing research in children’s IR, we moved beyond topicality to consider two factors influencing child-focused relevance. However, additional aspects warrant attention. For instance, children tend to be less critical than adults [30, 58], making them more vulnerable to misinformation [39, 41]. Emotions conveyed in search results also affect children’s engagement and compre-

Table 2. QPP performance across datasets based on CR. * denotes significance (Student’s T-test, $p < 0.05$) w.r.t. corresponding metric score using **kid-friend**.

QPP	kid-friend				requik			
	P- ρ	K- τ	sMARE	GeoRisk	P- ρ	K- τ	sMARE	GeoRisk
QL	0.115	0.1	0.694	0.457	0.055*	0.026*	0.679*	0.452*
QS	-0.065	-0.035	0.655	0.459	0*	0.068*	0.684*	0.444*
PMIavg	0.2	0.155	0.718	0.495	0.259*	0.17*	0.711*	0.488*
VARavg	0.038	0.012	0.662	0.471	0.42*	0.324*	0.762*	0.512*
SCQavg	0.007	-0.015	0.652	0.466	0.347*	0.279*	0.746*	0.506*
ICTFavg	0.119	0.11	0.714	0.481	0.208*	0.165*	0.72*	0.466*
IDFavg	0.036	0.001	0.664	0.476	0.224*	0.207*	0.727*	0.484*
SCS	0.004	-0.006	0.668	0.459	0.17*	0.146*	0.708*	0.461*
Clarity	0.095	0.02	0.662	0.462	0.062*	0.092*	0.688*	0.453*
WIG	0.311	0.205	0.719	0.507	0.283*	0.165*	0.72	0.492*
NQC	0.362	0.232	0.724	0.508	0.285*	0.189*	0.725	0.492*
SMV	0.353	0.223	0.72	0.508	0.296*	0.199*	0.73*	0.497*
σ_{max}	0.337	0.245	0.739	0.521	0.258*	0.169*	0.721*	0.493*
$n(\sigma_x\%)$	0.268	0.159	0.712	0.505	0.364*	0.186*	0.716*	0.485*
NQAQPP	-0.07	-0.02	0.662	0.438	0.282*	0.231*	0.739*	0.489*
BERTQPP	-0.076	-0.041	0.662	0.434	0.108*	-0.022*	0.658*	0.442*
qppBERT-PL	-0.242	-0.175	0.635	0.427	0.000	0.000	0.745*	0.537*
QPP-GenRE	-0.083	0.025	0.685	0.444	0.096*	0.025	0.683*	0.435*

hension [39, 42]. Building from this initial exploration, future work should assess QPP using a broader relevance definition, one balancing children’s knowledge and skills while protecting them from harm. Child-aware QPP methods could, in turn, guide retrieval pipelines to enhance children’s search experiences, as done in other QPP-based IR applications [e.g., 5, 9, 35]. For instance, such methods could prompt an IR system to reformulate a child’s query for more suitable results, or, in cases of ambiguity, employ clarification strategies as in conversational search [e.g., 50, 62, 69, 72, 74].

4 Concluding Remarks

We empirically investigated QPP methods in estimating retriever performance using a child-focused definition of relevance. Our findings reveal inconsistent trends across relevance definitions, metrics, and datasets, demonstrating that the studied methods are ill-suited for children’s information seeking. These results highlight the need for further research investment in adapting IR tasks and algorithms to non-mainstream users.

Acknowledgments. This work is supported by SNSF Award #[IC00I0-227887 project n.10000973 SOL] and used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-16917.

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

References

1. Allen, G., Milton, A., Wright, K.L., Fails, J.A., Kennington, C., Pera, M.S.: Supercalifragilisticexpialidocious: why using the “right” readability formula in children’s web search matters. In: European Conference on Information Retrieval, pp. 3–18. Springer (2022)
2. Amendum, S.J., Conradi, K., Hiebert, E.: Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students’ reading fluency and comprehension. *Educ. Psychol. Rev.* **30**(1), 121–151 (2018)
3. 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 (2021)
4. Azpiazu, I.M., Dragovic, N., Pera, M.S., Fails, J.A.: Online searching and learning: yum and other search tools for children and teachers. *Inf. Retrieval J.* **20**(5), 524–545 (2017)
5. Bellogín, A., Castells, P.: Predicting neighbor goodness in collaborative filtering. In: International Conference on Flexible Query Answering Systems, pp. 605–616. Springer (2009)
6. Bilal, D., Huang, L.M.: Readability and word complexity of serps snippets and web pages on children’s search queries: Google vs bing. *Aslib J. Inf. Manag.* **71**(2), 241–259 (2019)
7. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
8. Carmel, D., Yom-Tov, E.: Estimating the query difficulty for information retrieval. Morgan & Claypool Publishers (2010)
9. Carmel, D., Yom-Tov, E., Roitman, H.: Enhancing digital libraries using missing content analysis. In: Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1–10 (2008)
10. Carmel, D., Yom-Tov, E., Soboroff, I.: Sigir workshop report: predicting query difficulty-methods and applications. In: ACM SIGIR Forum, vol. 39, pp. 25–28. ACM, New York (2005)
11. Chen, X., He, B., Sun, L.: Groupwise query performance prediction with bert. In: European Conference on Information Retrieval, pp. 64–74. Springer (2022)
12. Clarke, C.L., Craswell, N., Voorhees, E.M.: Overview of the trec 2012 web track (2011)
13. Collins-Thompson, K., Macdonald, C., Bennett, P., Diaz, F., Voorhees, E.M.: Trec 2014 web track overview (2015)
14. Costa-juss, C.B.M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. *GeBNLP* **2019**, 33 (2019)

15. 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 (2002)
16. Crossley, S., Heintz, A., Choi, J.S., Batchelor, J., Karimi, M., Malatinszky, A.: A large-scaled corpus for assessing text readability. *Behav. Res. Methods* **55**(2), 491–507 (2023)
17. Cummins, R., Jose, J., O’Riordan, C.: Improved query performance prediction using standard deviation. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1089–1090 (2011)
18. Danovitch, J.H.: Growing up with google: how children’s understanding and use of internet-based devices relates to cognitive development. *Hum. Behav. Emerg. Technol.* **1**(2), 81–90 (2019)
19. Datta, S., Ganguly, D., Greene, D., Mitra, M.: Deep-QPP: a pairwise interaction-based deep learning model for supervised query performance prediction. In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pp. 201–209 (2022)
20. Datta, S., MacAvaney, S., Ganguly, D., Greene, D.: A ‘pointwise-query, listwise-document’ based query performance prediction approach. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2148–2153 (2022)
21. Dinçer, B.T., Ounis, I., Macdonald, C.: Tackling biased baselines in the risk-sensitive evaluation of retrieval systems. In: *European Conference on Information Retrieval*, pp. 26–38. Springer (2014)
22. Duarte Torres, S., Hiemstra, D., Weber, I., Serdyukov, P.: Query recommendation for children. In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2010–2014 (2012)
23. Faggioli, G., Ferro, N., Mothe, J., Raiber, F.: QPP++ 2023: query-performance prediction and its evaluation in new tasks. In: *European Conference on Information Retrieval*, pp. 388–391. Springer (2023)
24. Faggioli, G., Ferro, N., Muntean, C., Perego, R., Tonellotto, N., et al.: A spatial approach to predict performance of conversational search systems. In: *CEUR Workshop Proceedings*, vol. 3448, pp. 41–46. CEUR-WS (2023)
25. Faggioli, G., Ferro, N., Muntean, C.I., Perego, R., Tonellotto, N.: A geometric framework for query performance prediction in conversational search. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1355–1365 (2023)
26. Faggioli, G., Formal, T., Marchesin, S., Clinchant, S., Ferro, N., Piwowarski, B.: Query performance prediction for neural IR: are we there yet? In: *European Conference on Information Retrieval*, pp. 232–248. Springer (2023)
27. Faggioli, G., Zendel, O., Culpepper, J.S., Ferro, N., Scholer, F.: sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Inf. Retrieval J.* **25**(2), 94–122 (2022)
28. Fröbe, M., Barthololy, S.C., Hagen, M.: How child-friendly is web search? An evaluation of relevance vs. harm. In: *European Conference on Information Retrieval*, pp. 214–222. Springer (2025)
29. Gwizdka, J., Bilal, D.: Analysis of children’s queries and click behavior on ranked results and their thought processes in google search. In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, pp. 377–380 (2017)

30. Hämäläinen, E.K., Kiili, C., Marttunen, M., Räikkönen, E., González-Ibáñez, R., Leppänen, P.H.: Promoting sixth graders' credibility evaluation of web pages: an intervention study. *Comput. Hum. Behav.* **110**, 106372 (2020)
31. 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 (2019)
32. Hauff, C.: Predicting the effectiveness of queries and retrieval systems (2010)
33. 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 (2008)
34. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: *International Symposium on String Processing and Information Retrieval*, pp. 43–54. Springer (2004)
35. Hosseini, S.M., Arabzadeh, N., Zihayat, M., Bagheri, E.: Enhanced retrieval effectiveness through selective query generation. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3792–3796 (2024)
36. HuggingFace: google-bert/bert-base-uncased · hugging face (2024). <https://huggingface.co/google-bert/bert-base-uncased>
37. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Inf. Process. Manag.* **36**(6), 809–840 (2000)
38. Kuhlthau, C.: Learning in digital libraries: an information search process approach. *Library Trends* 45 (1997)
39. Landoni, M., Huibers, T., Murgia, E., Pera, M.S.: Good for children, good for all? In: *European Conference on Information Retrieval*, pp. 302–313. Springer (2024)
40. Landoni, M., Huibers, T.W., Aliannejadi, M., Murgia, E., Pera, M.S.: Getting to know you: search logs and expert grading to define children's search roles in the classroom. In: *CEUR Workshop Proceedings*, vol. 2950, pp. 44–52. Rheinisch Westfälische Technische Hochschule (2021)
41. Landoni, M., Murgia, E., Huibers, T., Pera, M.S.: How does information pollution challenge children's right to information access? In: *3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval, ROMCIR 2023*, pp. 17–29 (2023)
42. Landoni, M., Pera, M.S., Murgia, E., Huibers, T.: Inside out: exploring the emotional side of search engines in the classroom. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 136–144 (2020)
43. Madrazo Azpiaz, I., Dragovic, N., Anuyah, O., Pera, M.S.: Looking for the movie seven or sven from the movie frozen? A multi-perspective strategy for recommending queries for children. In: *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, pp. 92–101 (2018)
44. Madrazo Azpiaz, I., Pera, M.S.: An analysis of transfer learning methods for multilingual readability assessment. In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 95–100 (2020)
45. Meng, C., Arabzadeh, N., Aliannejadi, M., De Rijke, M.: 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*, pp. 2583–2593 (2023)
46. 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.* **43**(4), 1–35 (2025)

47. Meng, C., Faggioli, G., Aliannejadi, M., Ferro, N., Mothe, J.: QPP++ 2025: query performance prediction and its applications in the era of large language models. In: European Conference on Information Retrieval, pp. 319–325. Springer (2025)
48. 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)
49. Nascimento, R.M.D.A., Sousa, D.X.D., Faggioli, G., Alvarenga, P.J.L., Ferro, N., Gonçalves, M.A.: A robustness assessment of query performance prediction (QPP) methods based on risk-sensitive analysis. In: Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), pp. 416–429 (2025)
50. Ortega, D., Söhnle, S., Vu, N.T.: Improving and understanding clarifying question generation in conversational search. In: International Conference on Text, Speech, and Dialogue, pp. 222–235. Springer (2024)
51. Peikos, G., Pasi, G.: A systematic review of multidimensional relevance estimation in information retrieval. *Wiley Interdisc. Rev. Data Mining Knowl. Discov.* **14**(5), e1541 (2024)
52. Penedo, G., et al.: The fineweb datasets: decanting the web for the finest text data at scale. *Adv. Neural. Inf. Process. Syst.* **37**, 30811–30849 (2024)
53. Pera, M.S., Huibers, T., Murgia, E., Landoni, M.: Some things never change: overcoming persistent challenges in children IR. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, pp. 3107–3109. Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3726302.3730270>
54. Pera, M.S., Murgia, E., Landoni, M., Cena, F., Mauro, N., Huibers, T.: Report on the 1st workshop on information retrieval for understudied users (ir4u2 2024) at ecir 2024. In: ACM SIGIR Forum, vol. 58, pp. 1–5. ACM, New York, NY, USA (2024)
55. Pera, M.S., Murgia, E., Landoni, M., Huibers, T., Aliannejadi, M.: Where a little change makes a big difference: a preliminary exploration of children’s queries. In: European Conference on Information Retrieval, pp. 522–533. Springer (2023)
56. Pera, M.S., Wright, K.L., Kennington, C., Fails, J.A.: Children and information access: fostering a sense of belonging. In: Smith-Renner, A., Taele, P. (eds.) Joint Proceedings of the IUI 2023 Workshops: HAI-GEN, ITAH, MILC, SHAI, SketchRec, SOCIALIZE (CEUR Workshop Proceedings, vol. 3359), CEUR, Sydney, Australia, pp. 254–257 (2023)
57. Pérez-Iglesias, J., Araujo, L.: Standard deviation as a query hardness estimator. In: International Symposium on String Processing and Information Retrieval, pp. 207–212. Springer (2010)
58. Pilgrim, J., Vasinda, S.: Fake news and the “wild wide web”: a study of elementary students’ reliability reasoning. *Societies* **11**(4), 121 (2021)
59. Poesina, E., Ionescu, R.T., Mothe, J.: iQPP: a benchmark for image query performance prediction. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2953–2963 (2023)
60. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392 (2016)
61. Shtok, A., Kurland, O., Carmel, D., Raiber, F., Markovits, G.: Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst. (TOIS)* **30**(2), 1–35 (2012)

62. Siro, C., Abbasiantaeb, Z., Yuan, Y., Aliannejadi, M., de Rijke, M.: Do images clarify? A study on the effect of images on clarifying questions in conversational search. In: Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2025, pp. 273–291. Association for Computing Machinery, New York, NY, USA (2025). <https://doi.org/10.1145/3698204.3716464>
63. Tao, Y., Wu, S.: Query performance prediction by considering score magnitude and variance together. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 1891–1894 (2014)
64. Tian, X., Jia, Q., Mei, T.: Query difficulty estimation for image search with query reconstruction error. *IEEE Trans. Multimedia* **17**(1), 79–91 (2014)
65. Vanderschantz, N., Hinze, A.: Children’s query formulation and search result exploration. *Int. J. Digit. Libr.* **22**(4), 385–410 (2021)
66. Vanderschantz, N., Hinze, A., Cunningham, S.J.: “Sometimes the internet reads the question wrong”: children’s search strategies & difficulties. *Proc. Am. Soc. Inf. Sci. Technol.* **51**(1), 1–10 (2014)
67. Walter, V.A.: The information needs of children. In: *Advances in Librarianship*. Emerald Group Publishing Limited (1994). [https://doi.org/10.1108/S0065-2830\(1994\)0000018006](https://doi.org/10.1108/S0065-2830(1994)0000018006)
68. 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 (2018)
69. Zamani, H., et al.: Analyzing and learning from user interactions for search clarification. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, pp. 1181–1190. Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3397271.3401160>
70. Zendel, O., Shtok, A., Raiber, F., Kurland, O., Culpepper, J.S.: Information needs, queries, and query performance prediction. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 395–404. SIGIR’19, Association for Computing Machinery, New York, NY, USA (2019). <https://doi.org/10.1145/3331184.3331253>
71. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In: *European Conference on Information Retrieval*, pp. 52–64. Springer (2008)
72. Zhao, Z., Dou, Z., Zhou, Y.: Generating intent-aware clarifying questions in conversational information retrieval systems. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, pp. 3384–3394. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3627673.3679851>
73. 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 (2007)
74. Zou, J., Aliannejadi, M., Kanoulas, E., Pera, M.S., Liu, Y.: Users meet clarifying questions: toward a better understanding of user interactions for search clarification. *ACM Trans. Inf. Syst.* **41**(1) (2023). <https://doi.org/10.1145/3524110>