

# Can QPP Choose the Right Query Variant? Evaluating Query Variant Selection for RAG Pipelines

Negar Arabzadeh  
UC Berkeley  
Berkeley, CA, United States

Andrew Drozdov  
Databricks  
San Francisco, CA, United States

Michael Bendersky  
Databricks  
San Francisco, CA, United States

Matei Zaharia  
UC Berkeley  
Berkeley, CA, United States

## Abstract

Large Language Models (LLMs) have made query reformulation ubiquitous in modern retrieval and Retrieval-Augmented Generation (RAG) pipelines, enabling the generation of multiple semantically equivalent query variants. However, executing the full pipeline for every reformulation is computationally expensive, motivating selective execution: can we identify the best query variant before incurring downstream retrieval and generation costs? We investigate Query Performance Prediction (QPP) as a mechanism for variant selection across ad-hoc retrieval, and end-to-end RAG. Unlike traditional QPP, which estimates query difficulty across topics, we study intra-topic discrimination—selecting the optimal reformulation among competing variants of the same information need. Through large-scale experiments on TREC-RAG using both sparse and dense retrievers, we evaluate pre- and post-retrieval predictors under correlation- and decision-based metrics. Our results reveal a systematic divergence between retrieval and generation objectives: variants that maximize ranking metrics such as nDCG often fail to produce the best generated answers, exposing a “utility gap” between retrieval relevance and generation fidelity. Nevertheless, QPP can reliably identify variants that improve end-to-end quality over the original query. Notably, lightweight pre-retrieval predictors frequently match or outperform more expensive post-retrieval methods, offering a latency-efficient approach to robust RAG.

## CCS Concepts

• Information systems → Information retrieval.

## Keywords

Query Performance Prediction, Query Variant Selection, Query Reformulation, Retrieval Augmented Generation

### ACM Reference Format:

Negar Arabzadeh, Andrew Drozdov, Michael Bendersky, and Matei Zaharia. 2026. Can QPP Choose the Right Query Variant? Evaluating Query Variant Selection for RAG Pipelines. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, July 20–24, 2026, Melbourne, VIC, Australia. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3805712.3808571>



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGIR '26, Melbourne, VIC, Australia*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2599-9/2026/07  
<https://doi.org/10.1145/3805712.3808571>

## 1 Introduction

Retrieval-Augmented Generation (RAG) has rapidly become a dominant architectural paradigm for modern information systems [33, 73, 85]. Unlike traditional ad-hoc retrieval, where users directly consume a ranked list, RAG inserts a Large Language Model (LLM) between retrieval and the user, delegating answer synthesis to a generative model conditioned on retrieved evidence [46]. This shift fundamentally alters both the objective and the economics of search [44, 53]. In this setting, *query formulation* plays an amplified role [22, 43, 48]. A user’s original query may fail to retrieve passages that adequately ground generation, exacerbating vocabulary mismatch, intent drift, and underspecification [31, 51, 78]. LLM-based query reformulation has become common practice to mitigate this problem by generating multiple semantically equivalent query variants to improve recall and coverage before answer synthesis [14, 27, 39, 47, 82].

One straightforward strategy is to over-generate query variants, execute the full pipeline for each, and then select the best final answer. A system may generate multiple semantically equivalent reformulations, retrieve and synthesize an answer for each, and use an LLM-as-a-judge, majority voting, or answer-level scoring to choose the most faithful response [13, 57, 81]. While this “generate-all-then-select” paradigm directly optimizes answer quality, it is computationally expensive. Although generating query variants is cheap, executing retrieval, reranking, and long-context generation for every variant requires multiple full LLM inference passes. The cost scales linearly with the number of reformulations, increasing latency and monetary expense. In production settings, such exhaustive execution is often infeasible. This motivates a more efficient alternative: *can we identify the most promising query variant before incurring the downstream generation cost?*

Query Performance Prediction (QPP) offers a natural mechanism for this problem [2, 9, 20, 21]. Traditionally, QPP estimates retrieval effectiveness without relevance judgments and has been used for tasks such as selective query expansion, system routing, and risk-sensitive retrieval [24, 49, 50]. Its evaluation has largely relied on correlation with ranking metrics such as nDCG or Average Precision. While correlation remains the standard criterion, it measures statistical association rather than downstream decision quality. As shown by Ganguly et al. [32], QPP effectiveness varies across evaluation settings—retrieval models, metrics, and rank cut-offs—raising reproducibility concerns and highlighting limitations of correlation-based evaluation.

Recent work broadens QPP evaluation beyond the single-query, single-ranker setting. Zendel et al. [83] shows that standard formulations conflate queries with the underlying information need and that prediction quality depends on how well a query represents that need, motivating multiple query representations. Building on this, Santra et al. [67, 68] proposes generalized, downstream-aware frameworks that evaluate QPP by its ability to support ranker selection and improve retrieval in fusion pipelines, rather than solely by correlation. Collectively, these studies suggest that statistical agreement alone is insufficient to capture QPP’s practical utility. More directly related to query reformulation, Scells et al. [69] introduced Query Variation Performance Prediction (QVPP) in the CLEF 2017 TAR task [41], aiming to select the most effective manually crafted Boolean query variant in a high-recall medical setting. While this demonstrates the feasibility of variant-level prediction, it is restricted to Boolean retrieval in a specific domain and does not consider neural retrievers, LLM-generated reformulations, or generation-based tasks [18]. In contrast, recent RAG studies show that retrieval effectiveness and downstream utility are weakly aligned [3, 15, 77], and that predicted retrieval quality can guide adaptive retrieval in agentic RAG systems [76]. Building on this view of QPP as a decision-support mechanism, we investigate whether QPP can select among over-generated query variants prior to retrieval or generation.

Applying QPP to RAG, however, introduces two fundamental shifts. First, the objective changes: classical QPP predicts *retrieval relevance*, whereas RAG ultimately optimizes *generation utility*, i.e., the ability of retrieved evidence to support a faithful and complete answer. Recent work highlights a “utility gap,” where documents that score highly under ranking metrics do not necessarily improve generation quality [1, 5, 63, 77]. Second, the decision setting differs. Traditional QPP primarily estimates difficulty across different information needs (*inter-topic prediction*), whereas query reformulation produces multiple variants of the same information need. The task, therefore, becomes one of *intra-topic discrimination* in selecting the single most promising variant among competing reformulations.

Despite these advances, existing work remains largely retrieval-centric, focusing either on estimating query difficulty or selecting among rankers. In contrast, we study QPP in the context of variant selection over LLM-generated reformulations, where the objective is not merely to predict retrieval effectiveness but to identify the variant that yields the strongest end-to-end performance. To this end, we consider two information access paradigms: (i) *retrieval-only*, where output quality is measured using ranking metrics; and (ii) *retrieval-augmented generation*, where retrieved documents condition the generation process, and performance is measured using nugget-based utility. We evaluate QPP from a utility-oriented perspective across these two paradigms. This framing allows us to examine whether retrieval-optimal variants align with RAG-optimal ones and to quantify the extent of divergence between ranking-based and generation-based objectives. To structure our investigation, we address the following research questions:

- **RQ1:** Can pre-retrieval QPP methods reliably identify the best-performing query variant for retrieval and RAG?
- **RQ2:** Do post-retrieval QPP methods provide additional gains in variant selection accuracy for retrieval and RAG?

- **RQ3:** Does strong retrieval performance prediction translate into strong end-to-end RAG performance prediction?
- **RQ4:** Does correlation-based evaluation capture QPP’s utility in query variant selection, or should a more utility-oriented evaluation be preferred?

Through a large-scale replicability study on the TREC-RAG 2024 benchmark, we re-implement and evaluate a comprehensive suite of established pre- and post-retrieval QPP baselines within a unified experimental framework [55, 56, 75]. For each information need, we generate 30 LLM-based query variants and evaluate them using a broad spectrum of QPP methods, ranging from traditional lexical and statistical predictors to more recent supervised, neural, and transformer-based approaches. We systematically apply these predictors across all query variants, evaluate both sparse and dense retrievers, and assess performance under two information-seeking paradigms: retrieval-only and RAG.

Our results provide a systematic comparison of QPP behavior across these settings. While QPP methods can identify variants that improve over the original query in certain cases, we observe that strong retrieval prediction does not consistently translate into improved end-to-end generation quality. By releasing all code, query variants, retriever configurations, and evaluation scripts, we aim to establish a transparent and extensible benchmark for studying QPP in modern retrieval and generation pipelines. We release all data and code at <https://github.com/Narabzad/QPP-4-RAG>.

## 2 Related Work

**QPP Methods.** QPP estimates query effectiveness without relevance judgments and is commonly categorized into pre-retrieval and post-retrieval methods [9, 10, 20].

*Pre-retrieval Methods.* Pre-retrieval predictors rely solely on query and collection statistics. Classical approaches leverage term specificity and distributional signals such as IDF variants, SCS, SCQ, and ICTF [35, 45, 86]. More recent methods incorporate semantic representations, including embedding-based predictors and supervised transformer models that regress effectiveness directly from query text [12, 42, 61, 62, 64]. State-of-the-art approaches such as Query Space Distance (QSD) assume a smooth performance landscape in semantic space, estimating effectiveness by interpolating from neighboring queries in embedding space [16].

*Post-retrieval Methods.* Post-retrieval predictors analyze properties of the retrieved documents and often achieve higher predictive power at the higher computational cost [4, 6, 61]. Classical score-distribution methods, including Clarity, WIG, NQC, and SMV and more, remain strong baselines [25, 71, 72, 87]. These approaches measure signals such as the KL-divergence between the query-induced language model and the retrieved documents, or the variance of top retrieval scores, under the intuition that well-separated score distributions indicate easier queries. More recently, supervised QPP models directly learn to predict performance by fine-tuning neural models conditioned on retrieval outputs, often achieving SOTA results for specific retrievers [7, 26, 28, 34]. Generative relevance estimation methods further extend this paradigm by using LLMs to estimate document utility and approximate ranking metrics [50, 65, 66]. While effective, these approaches are computationally expensive.

**Evaluation and Applications of QPP.** Traditional QPP evaluation relies on correlation with ranking metrics. However, correlation alone may not reflect practical utility. Recent work advocates decision-oriented evaluation, assessing how well QPP supports downstream tasks such as ranker selection, routing, or risk-sensitive retrieval [59, 67, 68]. Ganguly et al. [32] further highlights the instability and context sensitivity of correlation-based assessment, emphasizing the need to evaluate QPP with respect to task-level outcomes—especially when retrieval is part of a larger pipeline. Consistent with this decision perspective, QPP has been widely used as a control signal in retrieval systems, informing selective query expansion, fusion weighting, and per-query routing between sparse and dense retrievers [11, 23, 29]. More recently, QPP-inspired signals have been explored in RAG pipelines to determine when retrieval is necessary, while parallel work in prompt performance prediction highlights performance variance across semantically equivalent inputs [76, 77].

**Our Position** Prior work largely frames QPP as a retrieval-centric difficulty estimator or system selection signal. We instead study QPP as a generation-aware mechanism for selecting among LLM-generated query variants, explicitly evaluating its impact on retrieval and RAG. This reframing enables us to examine the divergence between retrieval-optimal and generation-optimal variants within a unified framework.

## 3 Experimental Setup

### 3.1 Dataset

To study whether QPP can reliably select the best query variant for both retrieval and RAG pipelines, we require a benchmark that supports evaluation at multiple stages of the pipeline. Specifically, the dataset must allow separate assessment of (i) retrieval effectiveness and (ii) end-to-end RAG performance. For this reason, we conduct our experiments on the TREC-RAG 2024 benchmark. TREC-RAG is designed explicitly to evaluate RAG systems and provides evaluation protocols for retrieval and RAG tasks separately. The benchmark consists of 56 queries constructed over the MS MARCO v2.1 corpus, which contains over 138 Million passages. Importantly, these queries have been carefully and thoroughly judged across retrieval and generative dimensions by both human assessors and LLM-based judges, enabling a fair comparison of performance under different pipeline configurations. In this work, we specifically utilize human annotations for both retrieval and nugget-based evaluations, as detailed further in Section 3.4.

### 3.2 Query Reformulation

To generate diverse query variants corresponding to the same underlying information need, we adopt a range of state-of-the-art LLM-based reformulation methods implemented through the QueryGym framework [19] including:

- **Generative Query Reformulation (GenQR)** [80]: A zero-shot LLM-based method that directly generates enriched query variants to improve ad-hoc search.
- **GenQR-Ensemble** [27]: A GenQR extension that applies an ensemble prompting approach that combines multiple LLM-generated reformulations to create a more robust query variant.

- **MuGI** [84]: Generates auxiliary questions and multi-step reasoning expansions to capture latent user intent more effectively.
- **QA-Expand** [70]: An expansion strategy that enhances the query by generating multi-question answer pairs to provide broader semantic context.
- **Query2Doc** [79]: A generative document-centric strategy where the LLM generates a synthetic pseudo-document (answer) that serves to provide a rich set of semantically related terms likely to appear in relevant documents.
- **Query2Exp** [40]: An expansion-based approach where the LLM generates explanatory notes or definitions to clarify the underlying concepts and append this context to the original prompt.

All query variants are generated using GPT-4o following the original prompts implemented in the QueryGym framework<sup>1</sup>. We set the temperature to 0.6 and generated 5 samples per method to capture semantic diversity while managing variance. This results in 30 generated variants per information need plus the original query, totaling 31 variations.

### 3.3 Answer Synthesis

*Retrieval Setup.* We evaluate query variants using both sparse and dense retrievers to test whether QPP-based selection generalizes across retrieval paradigms.

- **Sparse Retrieval:** BM25 with standard MS MARCO parameters ( $k_1 = 0.9, b = 0.4$ ) via Pyserini. For each query variant, we retrieve the top-100 passages from MS MARCO v2.1.
- **Dense Retrieval:** Cohere embeddings with DiskVectorIndex to retrieve the top-100 passages using a compressed, memory-mapped index<sup>2</sup>.

Both configurations were official TREC RAG 2024 baselines. All retrieval parameters are fixed across experiments to isolate the effect of query reformulation.

*RAG Setup.* For RAG, we follow the official TREC-RAG 2024 baseline using the Ragnarok framework [55]. For each query variant, we retrieve the top-5 documents (sparse or dense) and provide them as context to the generator, which produces an answer conditioned on both the query and the retrieved documents. We avoid additional prompt engineering to ensure differences arise from query reformulation rather than prompt design.

### 3.4 System Performance Evaluation

We first evaluate the performance of each system configuration under retrieval-only and RAG settings using various query variants.

*3.4.1 Retrieval Evaluation.* Since RAG conditions on the top-5 retrieved documents, we report nDCG@5 to align retrieval evaluation with the evidence actually used for generation. We additionally report Recall@100 to capture higher-depth coverage. All metrics are computed using graded relevance judgments on a 0–3 scale provided by human assessors.

*3.4.2 RAG Evaluation.* We adopt the official nugget-based evaluation framework from the TREC-RAG 2024 benchmark [56]. Unlike standard ranking metrics, this framework quantifies how well a

<sup>1</sup><https://github.com/l33-lab/QueryGym>

<sup>2</sup><https://huggingface.co/datasets/Cohere/msmarco-v2-embed-english-3v>

synthesized answer satisfies the underlying information need by identifying atomic factual units, or “nuggets,” within the response. For each query, nuggets are manually created by human assessors based on highly relevant judged documents. Each nugget represents a distinct aspect of the information need and is assigned an importance level: VITAL (essential facts) or OKAY (supplementary details). System performance is measured by the degree of support provided for each nugget, i.e., Full Support, Partial Support, or No Support. These support levels are aggregated into weighted utility measures that provide flexibility in prioritizing critical query aspects. We report performance across two ends of the spectrum:

- **Nugget-All** ( $\mathbb{N}_{\text{All}}$ ): This represents the lenient end of the evaluation spectrum. In this metric, any satisfied nugget—whether VITAL or OKAY—contributes to the final score, rewarding systems for broad semantic coverage.
- **Nugget-Strict** ( $\mathbb{N}_{\text{Strict}}$ ): This represents the most stringent evaluation criteria. It focuses exclusively on the satisfaction of VITAL nuggets, typically requiring a high level of support (e.g., full support) to contribute to the performance score.

For details of metric scoring, we refer readers to [56].

### 3.5 Prediction Performance Evaluation

We evaluate the capability of QPP methods to estimate the relative quality of query variants derived from the same underlying information need. For each information need, we generate 30 distinct query variants. Each variant is executed through the retrieval and RAG pipelines described previously, generating a comprehensive set of ground-truth performance scores. Specifically, we measure retrieval effectiveness using nDCG@5 and Recall@100, and evaluate answer synthesis quality using nugget-based evaluation metrics, including Nugget-All ( $\mathbb{N}_{\text{All}}$ ) and Nugget-Strict ( $\mathbb{N}_{\text{Strict}}$ ).

**3.5.1 Correlation-based Evaluation.** Traditionally, QPP methods are evaluated by computing the correlation between predicted and actual system performance. Given a query  $q$ , a system  $S$  has true performance  $M_s(q)$  under metric  $M$ , while a QPP method predicts a performance score of  $\hat{M}_s(q)$ . Now let  $Q_I = q_1, q_2, \dots, q_n$  denote the  $n$  query variants for information need  $I$ . For each  $q_i \in Q_I$ , we obtain predicted scores  $\hat{M}_s(q_i)$  and true scores  $M_s(q_i)$ . Prediction quality is measured by the correlation between  $\hat{M}_Q$  and  $M_Q$ , denoted as  $\text{Corr}(\hat{M}_Q, M_Q)$ . Both linear and rank-based metrics, such as Pearson’s  $\rho$  and Kendall’s  $\tau$ , are commonly used, and correlations are aggregated across information needs.

While measuring the correlation between predicted and actual performance is standard practice, we argue that this traditional evaluation framework was originally designed for datasets with highly diverse topics (inter-topic difficulty). Therefore, correlation based evaluation may not be the most representative for the query variant selection task. In a production RAG pipeline, a system architect often only requires the predictor to identify the single best-performing variant to execute. If a QPP method successfully selects the optimal variant, the accuracy of the ranking for the remaining sub-optimal variants is of secondary importance. Therefore, we contend that QPP must be evaluated in a more realistic scenario that prioritizes selection accuracy over global correlation. We propose end-to-end evaluation pipelines for assessing the practicality of QPP when choosing among different query variants.

**3.5.2 End-to-End Evaluation.** We evaluate QPP as a practical mechanism for query variant selection rather than a global ranking tool. For each information need  $I_i$ , let  $Q_i = \{q_{i,1}, \dots, q_{i,n}\}$  denote its set of generated variants. We select the query variant predicted to be most effective by a QPP method as:

$$q_i^{\text{QPP}} = \arg \max_{q_{i,j} \in Q_i} \hat{M}_s(q_{i,j}), \quad (1)$$

where  $\hat{M}_s$  denotes the predicted performance score. The selected variant is then executed through the retrieval or RAG pipeline to measure its true system performance  $M_s(q_i^{\text{QPP}})$ .

For comparison, we define the ORACLE variant as:

$$q_i^{\text{ORACLE}} = \arg \max_{q_{i,j} \in Q_i} M_s(q_{i,j}), \quad (2)$$

which represents the maximum achievable performance under perfect selection. We evaluate QPP based on (i) the improvement over the ORIGINAL QUERY and (ii) the remaining gap to the ORACLE. This evaluation is strictly precision-focused, as it assesses the system’s ability to choose the single best variant without being penalized by the correlation of sub-optimal variants within the set.

### 3.6 QPP baselines

We consider a comprehensive set of QPP methods:

**3.6.1 Pre-retrieval QPP Methods.** We include the following pre-retrieval QPP methods as baselines in our experiments. For methods that operate at the query-term level, we aggregate term-level signals using different functions (e.g., average, maximum, sum), which are indicated in parentheses for each method. Methods that operate at the whole-query level do not require such aggregation.

- **IDF:** Aggregates the Inverse Document Frequency of query terms using average, maximum, standard deviation, and sum operations to estimate query specificity [45].
- **ICTF:** Inverse Collection Term Frequency, which penalizes terms that are frequent in the collection but rare in individual documents [45].
- **SCQ:** Collection Query Similarity, which measures the similarity between the query and the collection language model using TF-IDF statistics [86].
- **SCS:** Simplified Clarity Score estimates query ambiguity by approximating the KL-divergence between the query and collection language models [38]. We consider two variants:  $\text{SCS}_{\text{apx}}$ , a lightweight approximation based on query length and average ICTF, and  $\text{SCS}_{\text{full}}$ , the full KL-divergence formulation using query and collection term probabilities.
- **QL:** Query Likelihood uses the retrieval score of the query generated by the QL model as a proxy for performance [54].
- **DM:** Discounted Matryoshka is a neural pre-retrieval method that estimates difficulty based on the distance between the query embedding and a set of reference vectors. We use E5 embeddings to obtain the query vector representation [29].
- **QSD (Pre):** Query Space Distance, which estimates performance by interpolating the effectiveness of semantically similar historical queries in the embedding space [17].

**3.6.2 Post-retrieval QPP Methods.** These methods leverage the scores and content of the top- $k$  retrieved documents.

- **Clarity**: Measures the KL-divergence between the language model of the top- $k$  ( $k=100$ ) retrieved documents and the collection language model [24].
- **WIG**: Weighted Information Gain computes the information gain of the top retrieved documents compared to the corpus average, acting as a proxy for result quality [87].
- **NQC**: Normalized Query Commitment calculates the standard deviation of retrieval scores in the top- $k$  results, assuming that high variance indicates better discrimination between relevant and non-relevant items [72].
- **SMV**: Score Magnitude and Variance combines the mean and variance of retrieval scores to capture both the strength and discriminative power of the retrieval signal [74].
- $\sigma_{\max}$ : The maximum standard deviation of retrieval scores across ranked-list prefixes, capturing peak score dispersion [52].
- $\sigma_{50\%}$ : A QPP method that is based on standard deviation computed over documents scoring at least 50% of the top score, forming a variable-length ranked list [25].
- **RSD**: Ranking Score Distribution, which analyzes the decay curve of retrieval scores to predict difficulty [60].
- **BERT-QPP**: A supervised method that uses a cross-encoder or bi-encoder to aggregate query-document interaction signals into a performance prediction [8].
- **QSD (Post)**: Extends QSD by incorporating retrieved document information to refine the selection of historical nearest neighbors for performance estimation [17].

For score-distribution predictors such as NQC, WIG, and SMV, we include both normalized and non-normalized variants. Normalization refers to scaling retrieval scores by collection-level statistics (e.g., dividing by the average document score or applying a standardization factor) to reduce sensitivity to absolute score magnitude and improve comparability across queries or retrieval models.

## 4 Results and Findings

Table 1 reports end-to-end performance for retrieval-only and RAG when query variants are selected under different QPP settings. For each QPP method, we select, for every information need, the query variant with the highest predicted effectiveness score and execute it under the corresponding evaluation paradigm. We report retrieval effectiveness and nugget-based answer quality. Pre-retrieval QPP methods are applied to both retrieval-only and RAG settings. Post-retrieval QPP methods are likewise applied to both, as they rely on statistics from the retrieved list. We use the following metrics:

- **Nugget-All** ( $\mathbb{N}_{\text{all}}$ ): Lenient end-to-end utility.
- **Nugget-Strict** ( $\mathbb{N}_{\text{strict}}$ ): Strict end-to-end utility.
- **nDCG@5**: Retrieval quality at depth 5.
- **Recall@100**: Retrieval coverage at depth 100.

The first row corresponds to the *Original Query* without reformulation, serving as the baseline for retrieval and RAG performance. The next blocks present results for *pre- and post-retrieval QPP methods*, where for each method, the variant with the highest predicted score is selected. Finally, the last rows provide *oracle upper bounds per metric*, representing the maximum achievable performance if the best variant were selected using true scores. We report oracle selection based on ranking metrics (Oracle nDCG, Oracle Recall)

and answer-level metrics (Oracle Strict, Oracle All). The gap between each QPP method and the oracle quantifies the remaining room for improvement. Table 1 reports the results of QPP-based selection compared to the original query and oracle upper bounds. Improvements over the original query are underlined, and the **best-performing** method within each section is bold. This table enables us to examine whether QPP-based selection improves over the original query, whether retrieval gains translate into RAG gains, and how far current methods are from the theoretical ceiling.

### 4.1 RQ1: Can pre-retrieval QPP methods reliably identify optimal query variants?

We analyze whether pre-retrieval QPP signals are sufficient to select variants that improve retrieval and RAG performance. As shown in Table 1, underlined cells indicate improvements over the original query. We observe that pre-retrieval QPP methods almost consistently improve end-to-end RAG performance compared to the original query, with the exception of certain neural predictors such as DM. Across most nugget-based metrics ( $\mathbb{N}_{\text{strict}}$  and  $\mathbb{N}_{\text{all}}$ ), the majority of pre-retrieval methods yield improvements, indicating reliable selection of stronger query variants. For BM25, the best-performing pre-retrieval method ( $\text{IDF}_{\max}$ ) increases  $\mathbb{N}_{\text{all}}$  from 0.2730 to 0.3980 (+45.8%) and  $\mathbb{N}_{\text{strict}}$  from 0.2270 to 0.3770 (+66.1%) relative to the original query. Even in the dense retrieval setting where ranking is already strong, improvements remain notable. However, while these methods substantially improve nugget-based end-to-end performance, gains in retrieval metrics (e.g., nDCG@5 and Recall@100) are much smaller, especially for dense retrievers. Traditional term-based predictors such as  $\text{IDF}_{\text{avg}}$ ,  $\text{SCQ}_{\text{avg}}$ , and  $\text{ICTF}_{\text{avg}}$  consistently improve answer quality but often fail to improve ranking effectiveness. In the dense case, only SCS variants increase nDCG or Recall, yet nugget gains remain consistent across many methods.

This suggests that improvements in generation quality are not strictly driven by improvements in ranking metrics. Rather, pre-retrieval QPP selects variants that retrieve evidence better aligned with downstream synthesis, even when standard ranking measures show limited change. Moreover, traditional lexical predictors are not only sufficient but frequently outperform neural pre-retrieval methods in selecting variants that enhance answer quality, indicating that signals useful for RAG variant selection differ from those optimized for retrieval difficulty estimation.

**Takeaway RQ1:** Pre-retrieval QPP methods can successfully select query variants that improve end-to-end generation performance over the original query. However, they are less effective at improving intermediate retrieval metrics, particularly for dense retrievers, where gains in answer quality do not necessarily correlate with gains in standard ranking metrics.

### 4.2 RQ2: Do post-retrieval QPP methods improve variant selection accuracy?

We investigate whether incorporating ranked-list signals in post-retrieval QPP methods leads to stronger variant ranking and improved end-to-end selection compared to pre-retrieval predictors. Post-retrieval QPP methods demonstrate stronger improvements in retrieval metrics, particularly nDCG@5 and Recall@100 for dense

Category	Method	Sparse Retriever (BM25)				Dense Retriever (Cohere)			
		RAG		Retrieval		RAG		Retrieval	
		$\bar{N}_{All}$	$\bar{N}_{Strict}$	nDCG@5	Recall@100	$\bar{N}_{All}$	$\bar{N}_{Strict}$	nDCG@5	Recall@100
Original	original	0.273	0.227	0.285	0.178	0.377	0.328	0.557	0.375
Pre-retrieval	IDF <sub>avg</sub>	<u>0.372</u>	<u>0.349</u>	0.274	<u>0.209</u>	<u>0.415</u>	<u>0.386</u>	0.538	0.332
	IDF <sub>max</sub>	<b>0.398</b>	<b>0.377</b>	<b>0.360</b>	<b>0.239</b>	<u>0.398</u>	<u>0.386</u>	0.532	0.353
	IDF <sub>sum</sub>	<u>0.386</u>	<u>0.367</u>	<u>0.329</u>	<u>0.217</u>	<u>0.41</u>	<u>0.384</u>	0.519	0.343
	ICTF <sub>avg</sub>	<u>0.374</u>	<u>0.349</u>	0.275	<u>0.21</u>	<b>0.421</b>	<b>0.392</b>	0.532	0.332
	SCQ <sub>avg</sub>	<u>0.368</u>	<u>0.341</u>	0.275	<u>0.21</u>	<b>0.421</b>	<u>0.391</u>	0.544	0.337
	SCQ <sub>max</sub>	<u>0.362</u>	<u>0.338</u>	<u>0.358</u>	<u>0.213</u>	<u>0.378</u>	<u>0.366</u>	0.521	0.339
	SCQ <sub>sum</sub>	<u>0.384</u>	<u>0.363</u>	<u>0.339</u>	<u>0.219</u>	<u>0.409</u>	<u>0.386</u>	0.513	0.341
	SCS <sub>apx</sub>	<u>0.274</u>	0.227	<u>0.289</u>	0.177	0.376	0.328	<b>0.557</b>	<b>0.375</b>
	SCS <sub>full</sub>	<u>0.362</u>	<u>0.336</u>	<u>0.296</u>	<u>0.21</u>	<u>0.401</u>	<u>0.373</u>	0.517	0.327
	QL	<u>0.394</u>	<u>0.368</u>	<u>0.345</u>	<u>0.228</u>	<u>0.404</u>	<u>0.38</u>	0.527	0.348
	QSD <sub>Pre</sub>	<u>0.354</u>	<u>0.333</u>	<u>0.321</u>	<u>0.199</u>	<u>0.405</u>	<u>0.384</u>	0.531	0.34
	DM	0.273	0.227	0.285	0.178	0.377	0.328	<b>0.557</b>	<b>0.375</b>
Post-retrieval	RSD	<u>0.372</u>	<u>0.345</u>	<u>0.332</u>	<u>0.22</u>	<u>0.387</u>	0.328	<b>0.601</b>	<u>0.386</u>
	clarity	<u>0.333</u>	<u>0.279</u>	0.269	<u>0.197</u>	<b>0.39</b>	0.322	0.525	0.349
	NQC	<b>0.381</b>	<b>0.355</b>	<u>0.331</u>	<u>0.22</u>	<u>0.388</u>	0.321	<u>0.58</u>	<u>0.381</u>
	NQC <sub>norm</sub>	<u>0.35</u>	<u>0.316</u>	<b>0.407</b>	<b>0.254</b>	<u>0.388</u>	0.314	<u>0.583</u>	<u>0.384</u>
	$\sigma_{max}$	<u>0.378</u>	<u>0.353</u>	<u>0.316</u>	<u>0.218</u>	<u>0.384</u>	0.315	<u>0.573</u>	<u>0.382</u>
	$\sigma_{0.5}$	<u>0.377</u>	<u>0.352</u>	<u>0.333</u>	<u>0.215</u>	0.377	0.328	0.557	0.375
	SMV	<u>0.372</u>	<u>0.345</u>	<u>0.332</u>	<u>0.22</u>	<u>0.387</u>	0.328	<b>0.601</b>	<u>0.386</u>
	SMV <sub>norm</sub>	<u>0.348</u>	<u>0.313</u>	0.4	<u>0.252</u>	0.377	0.317	<u>0.593</u>	<b>0.387</b>
	WIG	<u>0.361</u>	<u>0.333</u>	<u>0.296</u>	<u>0.203</u>	0.377	0.328	0.557	0.375
	WIG <sub>norm</sub>	0.273	0.227	0.285	0.178	0.377	0.328	0.557	0.375
	QSD <sub>post</sub>	<u>0.357</u>	<u>0.322</u>	<u>0.391</u>	<u>0.207</u>	<u>0.385</u>	<b>0.343</b>	<u>0.568</u>	0.352
	BERTQPP <sub>bi-encoder</sub>	<u>0.31</u>	<u>0.297</u>	<u>0.304</u>	<u>0.218</u>	0.366	0.324	<u>0.564</u>	0.346
	BERTQPP <sub>cross-encoder</sub>	<u>0.336</u>	<u>0.32</u>	<u>0.353</u>	<u>0.215</u>	0.366	0.328	0.536	0.361
Oracle	Oracle-ndcg@5	<u>0.395</u>	<u>0.344</u>	<b>0.644</b>	<u>0.257</u>	<u>0.392</u>	<u>0.354</u>	<b>0.723</b>	0.374
	Oracle-recall@100	<u>0.383</u>	<u>0.363</u>	<u>0.493</u>	<b>0.333</b>	<u>0.378</u>	<u>0.332</u>	<u>0.608</u>	<b>0.444</b>
	Oracle- $\bar{N}_{All}$	<b>0.511</b>	<u>0.485</u>	<u>0.414</u>	<u>0.235</u>	<b>0.533</b>	<u>0.485</u>	<u>0.568</u>	0.356
	Oracle- $\bar{N}_{Strict}$	<u>0.463</u>	<b>0.536</b>	<u>0.374</u>	<u>0.231</u>	<u>0.486</u>	<b>0.569</b>	<u>0.571</u>	0.344

**Table 1: End-to-end performance of QPP-based query variant selection under retrieval-only and RAG settings. For each method, the highest-scoring variant is executed. Results are reported for BM25 and dense retrievers across ranking (nDCG@5, Recall@100) and nugget-based metrics ( $\bar{N}_{strict}$ ,  $\bar{N}_{all}$ ). Underlined values indicate improvements over the original query; bold denotes the best method per section. Oracle rows represent upper bounds under perfect selection.**

retrievers, as reflected by the greater number of underlined scores in the right-hand columns of the table compared to pre-retrieval methods. For example, under dense retrieval, nDCG@5 increases from 0.5570 to 0.6010 (+7.9%) with RSD. However, the corresponding improvement in  $\bar{N}_{All}$  is only +2.7%, which is smaller than the gains achieved by the best pre-retrieval methods. This indicates that stronger ranking prediction does not necessarily translate into proportionally stronger end-to-end answer quality.

This pattern is expected, as post-retrieval predictors directly exploit ranked-list signals (e.g., score distributions, clarity) and are therefore optimized to estimate retrieval effectiveness. While they are more effective at identifying variants that improve ranking quality, this advantage does not consistently result in superior downstream answer synthesis ( $\bar{N}_{Strict}$  and  $\bar{N}_{All}$ ) compared to the best pre-retrieval methods.

From a practical perspective, this distinction is crucial. If the ultimate goal is end-to-end answer quality, pre-retrieval QPP methods offer a compelling alternative. First, they avoid the latency

cost of executing an initial retrieval step before prediction. Second, they achieve competitive, and even in some cases stronger, end-to-end improvements. In fact, comparing pre-retrieval vs. post-retrieval performance suggests that selecting a semantically richer query variant may have a larger impact on generation quality than marginally improving the retrieval ranking itself. Additionally, comparing neural post-retrieval methods (e.g., BERT-QPP) with non-neural ones (e.g., NQC, WIG) reveals that traditional statistical predictors often perform surprisingly well, suggesting they may offer better generalizability across different retrieval models than supervised neural predictors trained on specific distributions.

**Takeaway RQ2:** Post-retrieval QPP methods outperform pre-retrieval methods in improving retrieval ranking metrics, particularly for dense retrievers. However, their advantage in end-to-end RAG performance is modest, suggesting that when the goal is answer quality, the additional latency of post-retrieval prediction may not be justified.

Category	Method	Sparse Retriever (BM25)				Dense Retriever (Cohere)			
		RAG		Retrieval		RAG		Retrieval	
		$N_{All}$	$N_{Strict}$	nDCG@5	Recall@100	$N_{All}$	$N_{Strict}$	nDCG@5	Recall@100
Pre-retrieval	IDF <sub>avg</sub>	0.205	0.131	-0.137	-0.138	<b>0.160</b>	0.132	-0.131	-0.210
	IDF <sub>max</sub>	0.243	<b>0.189</b>	<b>0.100</b>	<b>0.118</b>	0.059	0.058	0.008	-0.058
	IDF <sub>sum</sub>	0.244	0.167	0.006	0.076	0.102	0.094	-0.066	-0.119
	ICTF <sub>avg</sub>	0.227	0.153	-0.113	-0.099	0.154	<b>0.137</b>	-0.127	-0.192
	SCQ <sub>avg</sub>	0.201	0.123	-0.148	-0.138	0.159	0.128	-0.127	-0.212
	SCQ <sub>max</sub>	0.244	0.155	0.045	0.076	0.059	0.048	-0.047	-0.109
	SCQ <sub>sum</sub>	0.244	0.168	0.010	0.081	0.101	0.094	-0.063	-0.115
	SCS <sub>apx</sub>	-0.232	-0.181	-0.086	-0.183	-0.036	-0.053	0.011	<b>0.058</b>
	SCS <sub>full</sub>	0.168	0.089	-0.093	-0.065	0.115	0.079	-0.151	-0.212
	QL	<b>0.244</b>	0.172	0.029	0.108	0.092	0.092	-0.046	-0.091
	QSD <sub>pre</sub>	0.070	0.071	0.011	0.027	-0.016	-0.024	-0.106	-0.066
	DM	-0.041	-0.011	0.099	0.060	-0.073	-0.058	0.027	0.056
Post-retrieval	RSD	0.230	0.158	-0.010	0.073	-0.024	-0.002	0.320	0.417
	clarity	0.061	-0.002	-0.102	-0.093	<b>0.046</b>	<b>0.009</b>	-0.022	-0.002
	NQC	0.233	0.160	-0.007	0.068	-0.038	-0.015	<b>0.329</b>	<b>0.421</b>
	NQC <sub>norm</sub>	0.086	0.094	<b>0.218</b>	<b>0.325</b>	-0.023	-0.003	0.291	0.370
	$\sigma_{max}$	<b>0.234</b>	0.161	-0.005	0.068	-0.063	-0.036	0.290	0.372
	$\sigma_{0.5}$	0.234	<b>0.168</b>	-0.024	0.056	-0.069	-0.067	0.077	0.169
	SMV	0.230	0.158	-0.010	0.073	-0.024	-0.002	0.320	0.417
	SMV <sub>norm</sub>	0.085	0.093	0.199	0.324	-0.012	0.005	0.290	0.380
	WIG	0.220	0.133	-0.092	-0.050	-0.068	-0.063	0.033	0.119
	WIG <sub>norm</sub>	-0.253	-0.191	-0.052	-0.113	-0.057	-0.054	0.019	0.106
	QSD <sub>post</sub>	0.070	0.034	-0.082	-0.028	0.012	-0.024	0.061	0.079
	BERTQPP <sub>bi-encoder</sub>	0.082	0.106	0.046	0.071	-0.070	-0.062	0.077	0.074
BERTQPP <sub>cross-encoder</sub>	0.191	0.139	0.147	0.139	-0.004	-0.015	-0.006	-0.020	

Table 2: Pearson correlation of QPP methods with retrieval and RAG metrics. Highest correlation in each section is bolded.

### 4.3 RQ3: Does strong retrieval prediction translate into strong RAG performance prediction?

We examine whether choosing, for each information need, the query variant that maximizes retrieval effectiveness also yields the strongest end-to-end RAG performance. In other words, *if we optimize the query variant for retrieval metrics, do we also optimize for answer quality?*

The results in Table 1 provide critical insight through the Oracle rows. By comparing oracle selection based on retrieval metrics (e.g., Oracle-recall@100 and Oracle-ndcg@5) with oracle selection based on nugget metrics (e.g., Oracle- $N_{All}$  and Oracle- $N_{Strict}$ ), we observe a substantial divergence between retrieval-optimal and answer-optimal variant selection.

Even in the ideal scenario where, for each information need, we select the query variant that achieves the highest retrieval performance (Oracle-ndcg@5), end-to-end answer quality does not reach its maximum potential. For instance, with the BM25 retriever, nDCG increases from 0.2850 (original query) to 0.6440 under Oracle-ndcg@5. However, the corresponding  $N_{Strict}$  score for these “retrieval-optimized” queries reaches only 0.3440. In contrast, practical QPP methods such as IDF<sub>max</sub> (pre-retrieval) and NQC (post-retrieval) achieve higher  $N_{Strict}$  scores of 0.3770 and 0.3550, respectively. Importantly, this answer quality is not only far from the theoretical maximum, but it is also outperformed by practical QPP methods that do not rely on ground-truth labels. As another example, for dense retrievers, the post-retrieval method Clarity

achieves a  $N_{Strict}$  score of 0.3900, which is higher than the score obtained by optimizing directly for Oracle-recall@100 (0.3780). This demonstrates that selecting the variant that produces the best-ranked list does not necessarily produce the best final answer.

Conversely, examining oracle selection based on answer quality (Oracle- $N_{Strict}$ ) reveals a dramatically different outcome. Optimizing directly for Strict Nugget yields a score of 0.5360 for BM25 and 0.5690 for the Dense Retriever—substantially higher than the retrieval-optimized selection (0.3440 for BM25 and 0.3540 for Dense when optimized using Oracle-ndcg@5). However, these “answer-optimized” variants often exhibit lower retrieval effectiveness (e.g., BM25 nDCG decreases from 0.5698 to 0.3480). This contrast confirms that the query variant maximizing ranking metrics often differs from the one maximizing end-to-end answer utility. In other words, retrieval-optimal and RAG-optimal selection are fundamentally distinct objectives. Factors that improve ranking quality—such as early precision—do not fully capture downstream answer needs, which may depend on coverage, complementarity, and alignment with the generative model. Therefore, even perfect retrieval prediction would not guarantee optimal RAG performance if the optimized objective is misaligned with answer quality.

**Takeaway RQ3:** Selecting the query variant that maximizes retrieval effectiveness does not guarantee optimal end-to-end RAG performance even in ideal scenario. Retrieval-optimized variants yield substantially lower end-to-end quality than answer-optimized or QPP-selected variants, revealing a fundamental misalignment between retrieval-optimal and RAG-optimal selection.

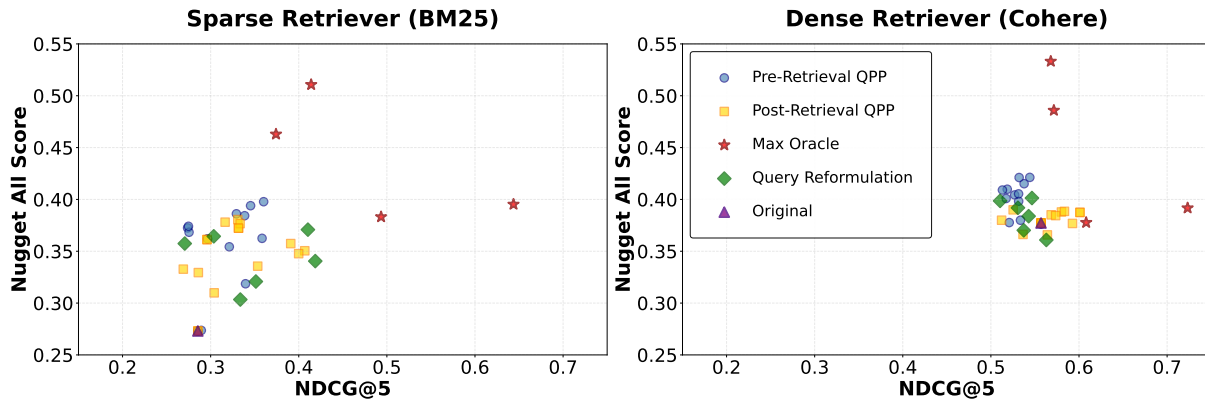


Figure 1: Relationship between retrieval effectiveness (nDCG@5) and end-to-end RAG utility (Nugget-All) under sparse and dense retrieval. Each point corresponds to a query variant selected by different strategies (pre-retrieval QPP, post-retrieval QPP, single reformulation, original query, oracle).

#### 4.4 RQ4: Does correlation-based evaluation capture QPP’s utility in RAG variant selection?

There has been a longstanding debate in the QPP community regarding the suitability of correlation coefficients as the primary evaluation metric [30, 36, 37]. Similar to [76, 77], We extend this discussion to the RAG setting, asking whether strong correlation with retrieval metrics reliably reflects a method’s usefulness for selecting the best-performing query variant in terms of end-to-end answer quality. Table 2 reports the Pearson correlation between QPP predictions and true performance across query variants, measured against both retrieval metrics (nDCG@5, Recall@100) and nugget-based RAG metrics. The results reveal a clear disconnect between these objectives.

For instance, the post-retrieval method *nqc* applied to the dense retriever exhibits a strong positive correlation with retrieval effectiveness ( $r = 0.3286$  for nDCG@5,  $r = 0.4206$  for Recall@100), yet its correlation with end-to-end Nugget scores is negligible ( $-0.00384$  for  $N_{All}$  and  $-0.0145$  for  $N_{Strict}$ ). This illustrates a critical failure mode that a QPP method may accurately predict ranking quality while providing no useful signal for selecting variants that improve answer quality. Conversely, *SCQ<sub>max</sub>* in BM25 shows the opposite pattern. It achieves moderate correlation with Nugget scores ( $r = 0.2436$ ), despite having near-zero correlation with nDCG@5 ( $r = 0.0447$ ). This suggests that the signals governing downstream answer utility differ from those that explain ranking difficulty.

These findings show that accurate retrieval prediction does not guarantee accurate RAG performance prediction. The relationship between ranking quality and answer utility is non-linear, influenced by factors such as complementarity, coverage, redundancy, and the model’s sensitivity to context composition [58]. More importantly, correlation does not align with the goal of query variant selection. While it captures average association, selection is a decision problem—requiring the identification of the single best variant per query. A method may show reasonable correlation yet still fail to rank the optimal variant correctly.

**Takeaway RQ4:** Correlation with retrieval metrics is not a sufficient proxy for evaluating QPP in RAG-based variant selection. A method that predicts ranking quality well may still fail to identify the variant that maximizes end-to-end answer utility. Evaluation should therefore align with the final decision objective i.e., selecting variants that improve answer fidelity, rather than relying solely on intermediate ranking correlations.

## 5 Discussion

A central practical question emerging from our results is how a RAG system should use query reformulation in deployment. If multiple reformulations can be generated for the same information need, should the system simply commit to the single reformulation strategy that performs best on average, or should it perform query-dependent variant selection? Table 3 reports the strongest configuration from each category for both sparse (BM25) and dense retrieval. Specifically, for each retriever, we present: (i) the original query, (ii) the best-performing single reformulator, (iii) the strongest pre-retrieval QPP method, (iv) the strongest post-retrieval QPP method, and (v) the Oracle upper bounds. Rather than averaging across methods, Table 3 emphasizes the best achievable performance within each paradigm, enabling a direct comparison between committing to a single reformulation strategy and performing query-dependent variant selection. Figure 1 complements this analysis by plotting retrieval effectiveness (nDCG@5) against end-to-end RAG utility ( $N_{All}$ ) under both sparse and dense retrieval.

*Query Variants Selection vs. Single Reformulation.* Here, we study whether variant selection truly improves over simply committing to the strongest reformulator. In Table 3, MuGI provides the best overall reformulation baseline for BM25 (averaged across five trials), while in dense retrieval the strongest reformulator similarly establishes a competitive baseline. QPP-based selection consistently improves over these single-reformulator strategies. For BM25, both the best pre-retrieval and best post-retrieval predictors outperform MuGI, demonstrating that selective routing yields additional gains beyond committing to a single reformulation. In dense retrieval, the



Category	Method	Sparse Retriever (BM25)				Dense Retriever (Cohere)			
		RAG		Retrieval		RAG		Retrieval	
		$N_{All}$	$N_{Strict}$	nDCG@5	Recall@100	$N_{All}$	$N_{Strict}$	nDCG@5	Recall@100
Original	original	0.273	0.227	0.285	0.178	0.377	0.328	0.557	0.375
Best Pre-retrieval	IDF <sub>max</sub>	<b>0.398</b>	<b>0.377</b>	<b>0.360</b>	<b>0.239</b>	0.398	0.386	0.532	0.353
	IDF <sub>avg</sub>	0.373	0.349	0.264	0.209	<b>0.415</b>	0.386	0.510	0.332
Best Post-retrieval	NQC	<b>0.381</b>	<b>0.355</b>	0.331	0.22	0.388	0.321	0.58	0.381
	Clarity	0.333	0.279	0.269	0.197	<b>0.39</b>	0.322	0.525	0.349
Best Vanilla Query Reformulation	MuGI	<b>0.371</b>	<b>0.349</b>	0.387	<b>0.254</b>	0.384	0.364	0.535	0.364
	GenQR	0.357	0.322	0.255	0.181	<b>0.401</b>	<b>0.375</b>	0.520	0.335
Oracle	Oracle-ndcg@5	0.395	0.344	<b>0.644</b>	0.257	0.392	0.354	<b>0.723</b>	0.374
	Oracle-recall@100	0.383	0.363	0.493	<b>0.333</b>	0.378	0.332	0.608	<b>0.444</b>
	Oracle- $N_{All}$	<b>0.511</b>	0.485	0.414	0.235	<b>0.533</b>	0.485	0.568	0.356
	Oracle- $N_{Strict}$	0.463	<b>0.536</b>	0.374	0.231	0.486	<b>0.569</b>	0.571	0.344

**Table 3: Best-performing configuration from each paradigm including Original, pre-retrieval QPP, post-retrieval QPP, single reformulation, and Oracle upper bounds, reported separately for BM25 and dense retrieval. Bold values are carried over from Table 1 and denote the best-performing method within their corresponding block in Table 1.**

improvement over the strongest reformulator is smaller (e.g.,  $N_{All}$  improving from approximately 0.4013 to 0.4152), yet still consistent. This suggests that when a reformulator is already strong, the marginal gain from QPP may be moderate—but still meaningful.

Figure 1 further clarifies this behavior. Single reformulators appear scattered across the plane, indicating that no fixed strategy dominates across information needs. QPP-based selection shifts configurations upward in  $N_{All}$ , even when nDCG gains are limited. Notably, this improvement does not require consistent increases in retrieval metrics, reinforcing that *RAG Performance Prediction* is distinct from traditional retrieval prediction.

Because both plots share the same scale, we can also directly compare sparse and dense retrievers. Dense retrieval achieves higher overall nDCG, but the points are tightly clustered, making configurations harder to distinguish using ranking metrics alone. In contrast, sparse retrieval exhibits greater dispersion, providing a more discriminative landscape for variant selection.

*Oracle Gap and Future Promise.* Despite the improvements that are achieved by QPP-based selection, a substantial gap remains between current methods and the Oracle upper bounds. The oracle results, computed by selecting the variant with the highest true score under each target metric, reveal that retrieval-optimal and answer-optimal variants frequently differ. Importantly, the oracle analysis in Tables 2 and 3 confirms that the best-performing variant is already present among the generated reformulations. The ceiling is therefore not hypothetical—it is attainable within the existing variant pool. In particular, the gains under  $N_{All}$  and  $N_{Strict}$  demonstrate that substantially higher end-to-end performance is achievable if the correct variant can be reliably identified.

While current QPP methods provide consistent but incremental improvements over single reformulation strategies, the remaining oracle gap highlights significant future potential. This positions QPP as a high-impact prediction problem: even modest advances in generation-aware or utility-aligned predictors could translate into large end-to-end gains without increasing the number of executed variants.

*Efficiency and Optimization Objectives.* Finally, these findings have direct system implications. Executing retrieval and generation for all variants and then selecting the best answer is computationally expensive. QPP shifts this decision upstream, selecting the best *input query* before incurring generation costs. Pre-retrieval predictors are particularly attractive due to their low latency.

More broadly, as illustrated in Figure 1, optimizing for retrieval (nDCG) is not equivalent to optimizing for RAG utility. Future work should, therefore, develop generation-aware predictors that treat retrieval and generation as a coupled system, directly estimating answer-grounding potential rather than ranking quality alone.

## 6 Conclusion

This work provides a comprehensive and fully reproducible study of QPP for query variant selection in RAG systems. We integrate pre- and post-retrieval QPP methods, multiple reformulation strategies, sparse and dense retrievers, and end-to-end generation evaluation within a unified framework. By standardizing datasets, metrics, retriever configurations, reformulation protocols, and random seeds, we enable controlled comparison across paradigms and isolate the effect of variant selection.

Our findings highlight two key insights. First, variant selection consistently improves robustness over committing to a single reformulation strategy, confirming the value of selective execution in RAG pipelines, though gains remain bounded relative to the Oracle ceiling. Second, retrieval metrics and generation utility are not structurally aligned: variants that optimize ranking effectiveness do not necessarily maximize answer quality, underscoring the need for end-to-end evaluation.

More broadly, our results position QPP as a decision-making component in modern RAG systems rather than merely a predictor of retrieval difficulty. By enabling upstream selection of query variants, QPP can reduce computational cost while maintaining or improving answer quality. However, the gap to oracle performance highlights substantial room for improvement, particularly in designing generation-aware predictors that account for context composition and downstream utility.

## References

- [1] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. 2024. Generative information retrieval evaluation. In *Information access in the era of generative ai*. Springer, 135–159.
- [2] Negar Arabzadeh and Ebrahim Bagheri. 2025. VAP3: Variation-Aware Prompt Performance Prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2794–2799.
- [3] Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. 2024. Adapting standard retrieval benchmarks to evaluate generated answers. In *European Conference on Information Retrieval*. Springer, 399–414.
- [4] Negar Arabzadeh, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2021. Query Performance Prediction Through Retrieval Coherency. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*. Springer, 193–200.
- [5] Negar Arabzadeh and Charles LA Clarke. 2024. A comparison of methods for evaluating generative ir. *arXiv preprint arXiv:2404.04044* (2024).
- [6] Negar Arabzadeh, Radin Hamidi Rad, Maryam Khodabakhsh, and Ebrahim Bagheri. 2023. Noisy perturbations for estimating query difficulty in dense retrievers. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 3722–3727.
- [7] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 2857–2861. doi:10.1145/3459637.3482063
- [8] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized Pre-trained transformers for Query Performance Prediction. In *CIKM*.
- [9] Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. 2024. Query performance prediction: Techniques and applications in modern information retrieval. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 291–294.
- [10] Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. 2025. Query performance prediction: Theory, techniques and applications. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 991–994.
- [11] Negar Arabzadeh, Mahsa Seifkar, and Charles LA Clarke. 2022. Unsupervised question clarity prediction through retrieved item coherency. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.
- [12] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras Al-Obeidat, and Ebrahim Bagheri. 2020. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management* 57, 4 (2020), 102248. doi:10.1016/j.ipm.2020.102248
- [13] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. (2024).
- [14] Amin Bigdeli, Negar Arabzadeh, and Ebrahim Bagheri. 2024. Learning to jointly transform and rank difficult queries. In *European Conference on Information Retrieval*. Springer, 40–48.
- [15] Amin Bigdeli, Negar Arabzadeh, Ebrahim Bagheri, and Charles LA Clarke. 2024. Evaluating relative retrieval effectiveness with normalized residual gain. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 64–71.
- [16] Amin Bigdeli, Sajad Ebrahimi, Negar Arabzadeh, Sara Salamat, Shirin SeyedSalehi, Maryam Khodabakhsh, Fattane Zarrinkalam, and Ebrahim Bagheri. 2025. Query Performance Prediction Using Neural Query Space Proximity. *ACM Trans. Intell. Syst. Technol.* (Sept. 2025). doi:10.1145/3762197
- [17] Amin Bigdeli, Sajad Ebrahimi, Negar Arabzadeh, Sara Salamat, Shirin SeyedSalehi, Maryam Khodabakhsh, Fattane Zarrinkalam, and Ebrahim Bagheri. 2025. Query Performance Prediction Using Neural Query Space Proximity. *ACM Transactions on Intelligent Systems and Technology* 17, 1 (2025), 1–25.
- [18] Amin Bigdeli, Mert Incesu, Negar Arabzadeh, Charles LA Clarke, and Ebrahim Bagheri. 2026. ReFormer: Learning and Applying Explicit Query Reformulation Patterns. In *European Conference on Information Retrieval*. Springer, 400–408.
- [19] Amin Bigdeli, Radin Hamidi Rad, Mert Incesu, Negar Arabzadeh, Charles LA Clarke, and Ebrahim Bagheri. 2025. QueryGym: A Toolkit for Reproducible LLM-Based Query Reformulation. *arXiv preprint arXiv:2511.15996* (2025).
- [20] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services, Vol. 2. Morgan & Claypool Publishers. 1–89 pages.
- [21] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What makes a query difficult?. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 390–397.
- [22] Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. *arXiv preprint arXiv:2404.00610* (2024).
- [23] Adrian-Gabriel Chifu, Sébastien Déjean, Moncef Garouani, Josiane Mothe, Diégo Ortiz, and Md Zia Ullah. 2025. Uncovering the Limitations of Query Performance Prediction: Failures, Insights, and Implications for Selective Query Processing. *ACM Transactions on Information Systems* (2025).
- [24] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 299–306.
- [25] Ronan Cummins, Joemon Jose, and Colm O’Riordan. 2011. Improved query performance prediction using standard deviation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 1089–1090.
- [26] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A’Pointwise-Query, Listwise-Documents’ based Query Performance Prediction Approach. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2148–2153.
- [27] Kaustubh D Dhole and Eugene Agichtein. 2024. Genqensemble: Zero-shot llm ensemble prompting for generative query reformulation. In *European Conference on Information Retrieval*. Springer, 326–335.
- [28] Sajad Ebrahimi, Maryam Khodabakhsh, Negar Arabzadeh, and Ebrahim Bagheri. 2024. Estimating query performance through rich contextualized query representations. In *European Conference on Information Retrieval*. Springer, 49–58.
- [29] Guglielmo Faggioli, Nicola Ferro, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonello. 2023. 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*. 1355–1365.
- [30] Guglielmo Faggioli, Oleg Zendej, J Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: a new paradigm to evaluate and understand query performance prediction methods. *Information Retrieval Journal* 25, 2 (2022), 94–122.
- [31] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* 30, 11 (1987), 964–971.
- [32] Debasis Ganguly, Suchana Datta, Mandar Mitra, and Derek Greene. 2022. An analysis of variations in the effectiveness of query performance prediction. In *European Conference on Information Retrieval*. Springer, 215–229.
- [33] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. 2, 1 (2023).
- [34] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2019. Performance Prediction for Non-Factoid Question Answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 55–58.
- [35] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. In *SIGIR Forum*, Vol. 44. 88.
- [36] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *European conference on information retrieval*. Springer, 301–312.
- [37] Claudia Hauff, Leif Azzopardi, Djoerd Hiemstra, and Franciska de Jong. 2010. Query performance prediction: Evaluation contrasted with effectiveness. In *European Conference on Information Retrieval*. Springer, 204–216.
- [38] Ben He and Iadh Ounis. 2004. Inferring Query Performance Using Pre-retrieval Predictors. In *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings*. 43–54. doi:10.1007/978-3-540-30213-1\_5
- [39] Seyed Mohammad Hosseini, Negar Arabzadeh, Morteza Zihayat, and Ebrahim Bagheri. 2024. Enhanced retrieval effectiveness through selective query generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 3792–3796.
- [40] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653* (2023).
- [41] Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. 2018. CLEF 2018 technologically assisted reviews in empirical medicine overview. In *CEUR workshop proceedings*, Vol. 2125. CEUR-WS.
- [42] Maryam Khodabakhsh, Fattane Zarrinkalam, and Negar Arabzadeh. 2024. BertPE: a BERT-based pre-retrieval estimator for query performance prediction. In *European Conference on Information Retrieval*. Springer, 354–363.
- [43] Julian Killingback and Hamed Zamani. 2025. Benchmarking Information Retrieval Models on Complex Retrieval Tasks. *arXiv preprint arXiv:2509.07253* (2025).
- [44] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohanany, Steven Schwarz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 4745–4759.
- [45] Kuilam L Kwok. 1996. A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*. 187–195.

- [46] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
- [47] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2023. Pseudo relevance feedback with deep language models and dense retrievers: Successes and pitfalls. *ACM Transactions on Information Systems* 41, 3 (2023).
- [48] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 5303–5315.
- [49] Chuan Meng, Negar Arabzadeh, Mohammad Aliannejadi, and Maarten de Rijke. 2023. Query Performance Prediction: From Ad-hoc to Conversational Search. *arXiv preprint arXiv:2305.10923* (2023).
- [50] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2025. Query performance prediction using relevance judgments generated by large language models. *ACM Transactions on Information Systems* 43, 4 (2025), 1–35.
- [51] Bhaskar Miuira and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends<sup>W</sup> in Accounting* 13, 1 (2018), 1–126.
- [52] Joaquín Pérez-Iglesias and Lourdes Araujo. 2010. Standard deviation as a query hardness estimator. In *International Symposium on String Processing and Information Retrieval*. Springer, 207–212.
- [53] Fabio Petroni, Federico Siciliano, Fabrizio Silvestri, and Giovanni Trappolini. 2024. IR-RAG@ SIGIR24: Information retrieval’s role in RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3036–3039.
- [54] Jay M Ponte and W Bruce Croft. 2017. A language modeling approach to information retrieval. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 202–208.
- [55] Ronak Pradeep, Nandan Thakur, Sahel Sharifmoghammad, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2025. Ragnarök: A reusable RAG framework and baselines for TREC 2024 retrieval-augmented generation track. In *European Conference on Information Retrieval*. Springer.
- [56] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework. *arXiv preprint arXiv:2411.09607* (2024).
- [57] Siddhant Ray, Rui Pan, Zhuohan Gu, Kuntai Du, Shaoting Feng, Ganesh Anantharayanan, Ravi Netravali, and Junchen Jiang. 2025. Metis: fast quality-aware rag systems with configuration adaptation. In *Proceedings of the ACM SIGOPS 31st symposium on operating systems principles*. 606–622.
- [58] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking prompt sensitivity in large language models. In *European Conference on Information Retrieval*. Springer, 303–313.
- [59] Haggai Roitman, Shai Erera, and Guy Feigenblat. 2019. A study of query performance prediction for answer quality determination. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 43–46.
- [60] Haggai Roitman, Shai Erera, and Bar Weiner. 2017. Robust standard deviation estimation for query performance prediction. In *Proceedings of the acm sigir international conference on theory of information retrieval*. 245–248.
- [61] Sara Salamat, Negar Arabzadeh, Shirin Seyedsalehi, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2023. Neural Disentanglement of Query Difficulty and Semantics. In *CIKM*. 4264–4268.
- [62] Sara Salamat, Negar Arabzadeh, Shirin Seyedsalehi, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2025. A contrastive neural disentanglement approach for query performance prediction. *Machine Learning* 114, 4 (2025), 109.
- [63] Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2400.
- [64] Abbas Saleminezhad, Negar Arabzadeh, Soosan Beheshti, and Ebrahim Bagheri. 2024. Context-aware query term difficulty estimation for performance prediction. In *European Conference on Information Retrieval*. Springer, 30–39.
- [65] Abbas Saleminezhad, Negar Arabzadeh, Soosan Beheshti, and Ebrahim Bagheri. 2026. Learning Context-aware Term Importance for Query Performance Prediction. *ACM Transactions on Intelligent Systems and Technology* 17, 2 (2026), 1–30.
- [66] Abbas Saleminezhad, Negar Arabzadeh, Seyed Mohammad Hosseini, Soosan Beheshti, and Ebrahim Bagheri. 2026. Structure-Aware Pre-retrieval Performance Prediction on Query Affinity Graphs. In *European Conference on Information Retrieval*. Springer, 547–556.
- [67] Payel Santra, Partha Basuchowdhuri, and Debasis Ganguly. 2026. Beyond Correlations: A Downstream Evaluation Framework for Query Performance Prediction. *arXiv preprint arXiv:2601.17339* (2026).
- [68] Payel Santra, Partha Basuchowdhuri, and Debasis Ganguly. 2026. Breaking Flat: A Generalised Query Performance Prediction Evaluation Framework. *arXiv:2601.17359* [cs.IR] <https://arxiv.org/abs/2601.17359>
- [69] Harrison Scells, Leif Azzopardi, Guido Zuccon, and Bevan Koopman. 2018. Query variation performance prediction for systematic reviews. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1089–1092.
- [70] Wonduk Seo, Hyunjin An, and Seunghyun Lee. 2025. A New Query Expansion Approach via Agent-Mediated Dialogic Inquiry. *arXiv:2502.08557* [cs.IR] <https://arxiv.org/abs/2502.08557>
- [71] Anna Shtok, Oren Kurland, and David Carmel. 2010. Using statistical decision theory and relevance models for query-performance prediction. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 259–266.
- [72] Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems (TOIS)* 30, 2 (2012), 1–35.
- [73] Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaie Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136* (2025).
- [74] Yongquan Tao and Shengli Wu. 2014. 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*. 1891–1894.
- [75] Nandan Thakur, Ronak Pradeep, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. Assessing Support for the TREC 2024 RAG Track: A Large-Scale Comparative Study of LLM and Human Evaluations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2759–2763.
- [76] Fangzheng Tian, Jinyuan Fang, Debasis Ganguly, Zaiqiao Meng, and Craig Macdonald. 2025. Am I on the Right Track? What Can Predicted Query Performance Tell Us about the Search Behaviour of Agentic RAG. (2025).
- [77] Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. 2025. Is Relevance Propagated from Retriever to Generator in RAG?. In *European Conference on Information Retrieval*. Springer, 32–48.
- [78] Ellen M Voorhees. 1993. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. 171–180.
- [79] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678* (2023).
- [80] Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023. Generative query reformulation for effective adhoc search. *arXiv preprint arXiv:2308.00415* (2023).
- [81] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).
- [82] Fanghua Ye, Meng Fang, Shenghui Li, and Emine Yilmaz. 2023. Enhancing conversational search: Large language model-aided informative query rewriting. *arXiv preprint arXiv:2310.09716* (2023).
- [83] Oleg Zende, Anna Shtok, Fiana Raiber, Oren Kurland, and J. Shane Culpeper. 2019. Information Needs, Queries, and Query Performance Prediction. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21–25, 2019*. 395–404. doi:10.1145/3331184.3331253
- [84] Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. *arXiv preprint arXiv:2401.06311* (2024).
- [85] Qingfei Zhao, Ruobing Wang, Yukuo Cen, Daren Zha, Shicheng Tan, Yuxiao Dong, and Jie Tang. 2024. Longrag: A dual-perspective retrieval-augmented generation paradigm for long-context question answering. (2024).
- [86] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30–April 3, 2008. Proceedings*. 52–64.
- [87] Yun Zhou and W Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 543–550.