Estimating Query Performance Using Neural Query Space Proximity

AMIN BIGDELI, University of Waterloo, Canada
SAJAD EBRAHIMI, University of Guelph, Canada
NEGAR ARABZADEH, University of Waterloo, Canada
SARA SALAMAT, Toronto Metropolitan University, Canada
SHIRIN SEYEDSALEHI, Toronto Metropolitan University, Canada
MARYAM KHODABAKHSH, Shahrood University of Technology, Iran
FATTANE ZARRINKALAM, University of Guelph, Canada
EBRAHIM BAGHERI, Toronto Metropolitan University, Canada

The varying performance of information retrieval (IR) methods, including state-of-the-art transformer-based neural retrievers, across diverse queries poses a significant challenge for achieving robust and reliable retrieval effectiveness. Query Performance Prediction (QPP) seeks to estimate the effectiveness of a retrieval method for individual queries, enabling adaptive strategies to improve retrieval outcomes, particularly for challenging queries. However, existing QPP approaches face fundamental challenges: pre-retrieval methods often rely on surface-level query features that fail to capture the nuanced relationship between queries and retrieval effectiveness, while post-retrieval methods depend heavily on the quality of retrieved documents, which can be unreliable for difficult queries. To this end, we propose the *Query Space Distance-Based QPP (*`QSD-QPP`*)* framework, which leverages the deterministic and consistent behavior of retrieval methods to estimate query performance by referencing historical queries with known effectiveness. The approach is motivated by the observation that semantically or syntactically similar queries often exhibit consistent retrieval performance, a property that can be exploited to make reliable predictions for unseen queries. `QSD-QPP` operates in two modes: (1) a *lightweight pre-retrieval* instantiation that dynamically constructs a query subspace based on embedding distances to interpolate the performance of proximate historical queries, and (2) an *enriched post-retrieval* instantiation that incorporates contextualized embeddings, document interactions, and historical query associations to enhance prediction accuracy. By utilizing large-scale contextualized embeddings derived from pre-trained language models, `QSD-QPP` efficiently identifies semantically similar queries and leverages their performance for robust predictions. By addressing the inherent limitations of prior approaches, `QSD-QPP` achieves a balanced trade-off between computational efficiency, prediction accuracy, and scalability. We evaluate `QSD-QPP` on four benchmark datasets, including MS MARCO Dev and TREC Deep Learning tracks (2019, 2020, and DL-Hard), demonstrating its superior accuracy and robustness compared to state-of-the-art baselines in both pre-retrieval and post-retrieval QPP tasks. To ensure reproducibility and encourage further research, we publicly release the implementation of our work.

Authors' Contact Information: Amin Bigdeli, University of Waterloo, Waterloo, ON, Canada, abigdeli@uwaterloo.ca; Sajad Ebrahimi, University of Guelph, Guelph, ON, Canada, sebrah05@uoguelph.ca; Negar Arabzadeh, University of Waterloo, Waterloo, ON, Canada, Narabzad@uwaterloo.ca; Sara Salamat, Toronto Metropolitan University, Toronto, ON, Canada, sara.salamat@torontomu.ca; Shirin SeyedSalehi, Toronto Metropolitan University, Toronto, ON, Canada, shirin.seyedsalehi@torontomu.ca; Maryam Khodabakhsh, Shahrood University of Technology, Shahrood, Iran, m_khodabakhsh@shahroodut.ac.ir; Fattane Zarrinkalam, University of Guelph, Guelph, ON, Canada, fzarrink@uoguelph.ca; Ebrahim Bagheri, Toronto Metropolitan University, Toronto, ON, Canada, bagheri@torontomu.ca.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts: • **Information systems** → **Information retrieval**; • **Computing methodologies** → *Machine learning approaches*; *Nearest neighbor algorithms*; • **General and reference** → Performance.

Additional Key Words and Phrases: Query Performance Prediction, Information Retrieval, Pre-retrieval Prediction, Post-retrieval Prediction, Contextualized Representations, Query Space

ACM Reference Format:
Amin Bigdeli, Sajad Ebrahimi, Negar Arabzadeh, Sara Salamat, Shirin SeyedSalehi, Maryam Khodabakhsh, Fattane Zarrinkalam, and Ebrahim Bagheri. 2018. Estimating Query Performance Using Neural Query Space Proximity. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Researchers in the field of Information Retrieval (IR) are focused on designing methods that achieve both high effectiveness and robustness. A key objective is to ensure that these methods consistently deliver strong performance across a wide range of queries, as highlighted by recent studies [4, 13, 48]. *Transformer-based neural retrieval methods* have been a strong step in this direction [26, 65]. In practice, even advanced IR methods, including powerful transformer-based neural retrievers, often exhibit inconsistent effectiveness, excelling with certain queries while struggling with others [4]. As such, identifying difficult queries allows the use of alternative strategies on these queries, such as query routing [58], query reformulation [55], and asking users to clarify their intents, to improve their performance [1, 5]. The task of Query Performance Prediction (QPP) plays a crucial role in addressing this challenge by estimating how well a retrieval method is likely to perform on a given query. Accurate QPP enables retrieval methods to optimize their processing and prioritization of queries, improving both their efficiency and effectiveness. This, in turn, enhances the overall user experience by delivering more reliable and relevant results [15].

QPP methods can essentially be categorized into two classes: pre-retrieval and post-retrieval methods [9]. Pre-retrieval methods estimate query performance before any documents are retrieved, relying on features such as query term statistics, geometric properties of the query in embedding space, and corpus-level statistics [6, 8, 15, 23, 28, 57, 60]. These methods are computationally efficient as they do not need the retrieval method to be executed on the query, making them valuable for early performance predictions and resource conservation. On the other hand, post-retrieval methods predict query performance after the initial retrieval stage by leveraging additional information, such as retrieval scores of the retrieved documents [3, 20, 27, 36, 67]. By leveraging richer data, these methods generally provide more accurate predictions compared to pre-retrieval methods.

Recent advances in both pre-retrieval [6–8, 66] and post-retrieval [3, 20, 27, 67] QPP methods have significantly improved the prediction effectiveness. However, to the best of our knowledge, existing approaches have not fully utilized large-scale labeled query collections to predict the performance of incoming queries. Our work builds on the deterministic nature of retrieval methods, which exhibit consistent behavior when processing similar queries. This observation leads to an *effective intuition*: the performance of an unseen query can be estimated by referencing a set of past queries with known effectiveness that are semantically or syntactically similar. Simply put, when a query that is identical to an already processed query is encountered, the performance of the retrieval method on the same query will remain consistent, and hence predicting its performance again will become trivial for such situations. This idea can be generalized to cases when a new query is observed that has not been observed in the past. In such situations, the performance of the new query could potentially be estimated based on its association with the performance of similar, previously observed and processed queries.

We note that strong empirical support for our intuition can be found in the query set of the MS MARCO Development set (Dev small) [44]. Specifically, for each query, we measured the embedding distance between its representation and the 10 historical queries with the closest proximity. Subsequently, we interpolated the

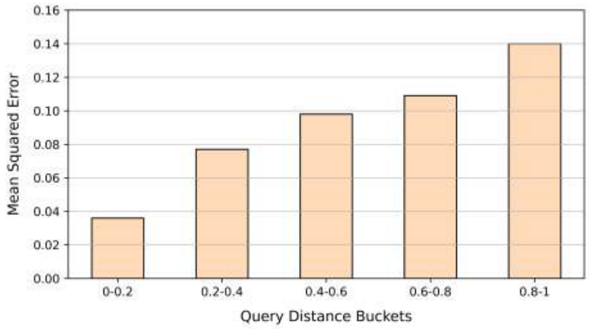

Fig. 1. Relationship between query distances and their retrieval effectiveness, demonstrating that smaller distances are associated with lower MSE in predicting query performance.

performance of these 10 queries to estimate the target query's performance. Analyzing the relationship between these distance scores and the Mean Squared Error (MSE) of the predicted versus actual query performance revealed a clear trend, as depicted in Figure 1. This figure illustrates the relationship between the embedding distances of Dev small queries, divided into five query buckets based on their average distance scores to proximate queries, and the average MSE between their retrieval effectiveness and that of the 10 most proximate historical queries. The results show that as the embedding distance between the target query and queries within its approximate proximity decreases, the MSE also decreases, indicating improved accuracy in query performance prediction.

We believe that our intuition, supported by preliminary empirical evidence, is grounded in the fact that retrieval methods produce deterministic retrieval outcomes. In other words, regardless of how many times a query is submitted to a retrieval method, the retrieved set of documents will remain the same and hence their retrieval effectiveness will not change[1]. This deterministic and consistent retrieval behavior in retrieving and ranking documents for queries can be generalized and employed for estimating the performance of future-yet-unobserved queries based on their proximity to historical queries within a shared query space. On this basis, we propose the Query Space Distance-Based QPP (QSD-QPP) framework, inspired by the deterministic nature of retrieval methods.

Our QSD-QPP framework offers two instantiations tailored for pre-retrieval and post-retrieval scenarios: In the pre-retrieval QPP setting, predictions are made without access to retrieved documents. Through searching the Query Space, the method dynamically constructs a query subspace comprising historical queries with the shortest distances to the input query and interpolates their known performance to estimate the effectiveness of the input query. Our post-retrieval QPP setting incorporates additional information from the retrieved documents. It refines predictions by combining the query's contextualized representation, document characteristics, and associations with historical queries. To efficiently create a query subspace for each input query, our QSD-QPP approach uses contextualized representations derived from large language models (LLMs), which map queries into a high-dimensional representation space. Historical queries within the subspace are identified by measuring the distance between their embeddings in this space, enabling accurate performance estimation for both seen and new queries.

We evaluate our framework on four widely-used query collections, including the MS MARCO passage collection and TREC Deep Learning (DL) query sets from 2019, 2020, and DL-Hard. The experimental results demonstrate

[1] We note that work on personalized search, which may return different results for the same query depending on contextual information, is not the subject of this paper. This paper focuses on *ad hoc retrieval* [54].

that our approach significantly outperforms state-of-the-art methods in both pre-retrieval and post-retrieval QPP tasks, achieving robust and accurate query performance predictions.

The key contributions of this work can be enumerated as follows:

- We introduce a novel Query Space Distance-Based QPP (QSD-QPP) framework that leverages historical query performance data for both pre-retrieval and post-retrieval predictions, enabling robust performance estimation for diverse query scenarios.
- We offer a lightweight pre-retrieval instantiation of QSD-QPP that efficiently predicts query performance by interpolating effectiveness of relevant queries in close proximity to the input query within the Query Space.
- We propose a post-retrieval variation of QSD-QPP that enriches query representations by incorporating contextualized embeddings, document interactions, and historical query associations for more precise predictions.
- We present extensive experiments conducted on diverse datasets, including MS MARCO and the TREC Deep Learning tracks (2019, 2020, and DL-Hard), showcasing the superior performance of our methods compared to state-of-the-art baselines.
- We publicly release the code and models to foster reproducibility and enable further research in query performance prediction tasks[2].

The remainder of this paper is organized as follows: In Section 2, we review existing QPP methods and discuss their limitations. Section 3 introduces our Query Space Distance-Based QPP framework, detailing the methodologies for both pre-retrieval and post-retrieval settings. We further describe the experimental setup, datasets, and evaluation metrics, followed by the results and analysis of our approach compared to state-of-the-art methods in Sections 4 and 5. Finally, Section 6 concludes the paper with a summary of our contributions and potential directions for future research.

2 Related Work

QPP methods can be broadly categorized into pre-retrieval and post-retrieval techniques. In the following, we will cover pertinent methods in each of these two classes of QPP methods.

2.1 Pre-retrieval QPP

Pre-retrieval QPP methods can be classified into statistical-based and neural-based methods. Traditional statistical-based techniques such as SCS, IDF, ICTF, VAR, PMI, and SCQ, mainly operate based on the query term distribution within the document collection [28, 29, 40, 68]. These methods can be further classified according to the query aspects they incorporate for prediction such as term importance, specificity, similarity, term-relatedness, and coherency [29–31, 49].

Recent advancements in pre-retrieval QPP methods incorporate external resources, such as neural embeddings and corpus-level statistics, to significantly enhance the accuracy and reliability of query performance predictions [3, 6, 8, 17–19, 27, 57, 61, 67]. For example, the $P_{clarity}$ method uses word embeddings to measure the ambiguity of each query term [57]. It estimates a Gaussian Mixture Model (GMM) from the distribution of the representations of the most similar words to the query terms and uses the posterior probabilities from this GMM to quantify the query's ambiguity. This method works based on the assumption that ambiguous queries encompass several senses and each sense a word carries roughly corresponds to a Gaussian mixture component. In another work, Arabzadeh et al. [6–8] proposed a set of neural-based pre-retrieval QPP metrics that assess query specificity by constructing an ego network based on the query term and its closest neighbors in the embedding space. They use graph centrality and density metrics as indicators of query difficulty. The premise behind these neural-based

[2] https://github.com/sadjadeb/QSD_QPP

QPP methods is that a more specific term is typically surrounded by more closely related concepts, while a more generic term will have more distant neighbors.

Other pre-retrieval QPP studies have primarily focused on leveraging neural query representations to address the QPP task. Zamani et al. [66] proposed an innovative multivariate representation learning framework that utilizes LLMs to generate both mean and variance vector representations for queries and documents. They subsequently employed these vectors to train a retrieval model. Their study revealed a compelling correlation between the learned vector representations of the queries and the corresponding ranking performance of those queries. Furthermore, the BertPE method [38] fine-tuned pre-trained language models to directly predict query performance, utilizing synthetic relevance judgments to overcome the scarcity of labeled training data. Using BERT-based neural embeddings, this method achieves competitive performance across various datasets, showcasing the potential of integrating *Transformer* architectures for robust pre-retrieval predictions.

Our work distinguishes itself from existing approaches by proposing a novel pre-retrieval QPP methodology that creates a Query Space from queries with known performance metrics. By dynamically constructing a subspace of queries with shortest distance to the input query, our method leverages their performance data to accurately predict the potential effectiveness of the new query. Our proposed approach benefits from the semantic neural-based representation of the queries while also maintaining a very low cost, making it applicable to real-world scenarios for pre-retrieval QPP methods.

2.2 Post-retrieval QPP

Post-retrieval methods can be divided into unsupervised [2, 16, 43, 60, 63, 69] and supervised [3, 27, 36, 67] methods. These methods typically exhibit a stronger correlation with retrieval performance compared to pre-retrieval approaches. This is because they utilize additional information from the retrieved documents, combined with insights from the query and corpus, to enhance prediction accuracy.

Traditionally, unsupervised methods predict the effectiveness of queries by analyzing the statistical characteristics of the query, the retrieved documents, and the corpus itself. Their primary advantage is that they do not require labeled data for training. Many traditional post-retrieval QPP methods utilize the distribution of retrieval scores for the retrieved documents as indicators of retrieval performance. For example, Clarity [15] measures the divergence between the language models induced by the retrieved documents and the collection as a whole. Similarly, Weighted Information Gain (WIG) [69] evaluates the disparity between the average retrieval score of the top-ranked documents and the average score of the entire collection. The underlying assumption is that the closer these documents are to the query, in comparison to a typical non-relevant document from the corpus, the more successful the retrieval process will be. The variation in retrieval scores has proven to be an effective indicator of query performance. For instance, the Normalized Query Commitment (NQC) [60] predicts retrieval performance by calculating the standard deviation of scores among the top-retrieved documents. A higher standard deviation suggests that relevant and non-relevant documents are more distinct, indicating better query performance. Similarly, other methods such as SMV [63] and $n(\sigma_x)$ [16] operate on the same principle, taking advantage of the divergence of scores to assess the effectiveness of the query.

Some models focus on the robustness of the ranking of retrieved documents. For instance, Query Feedback (QF) predicts query performance based on the overlap between the documents retrieved for the original query and those retrieved for a perturbed query derived from pseudo-relevance feedback. Singh et al. [61] also explored this idea by comparing the retrieved document lists for the original query with a reranked list using a pairwise neural-based ranker. Recently, Arabzadeh et al. [2] proposed a similar concept tailored for dense retrievers. However, QPP methods specific for dense retrievers are beyond the scope of our paper [24].

Thanks to large-scale relevance judgment collections, such as MS MARCO [45], researchers have explored supervised post-retrieval QPP methods more extensively [3, 17, 27, 67]. For instance, NeuralQPP [67] is one of the pioneering neural frameworks that use unsupervised QPP methods as weak signals to develop a more effective supervised approach. There has also been growing interest in employing contextualized pre-trained language models [21] for the QPP task. These models estimate the performance of a query by fine-tuning a language model [3, 19, 27]. As the pioneer in this group, NQA-QPP [27] integrates three key components to predict query performance: (1) a score-based component to analyze the distribution of retrieval scores, (2) a representation of the query, and (3) interactions between the representations of the query and the top retrieved documents. More recently, Khodabakhsh et al. [37] proposed a method where ranking and QPP tasks are learned together, enhancing each other. As other examples, BERT-QPP [3] and qpp-BERT-pl [19] focus on estimating query performance by fine-tuning language models such as BERT, aiming to learn the effectiveness of the query from the interactions between the query and retrieved document tokens. Building upon this group of studies, Ebrahimi et al. [22] proposed enhancing BERT-QPP by incorporating relevance scores to improve the effectiveness of query performance prediction.

Our approach distinguishes itself from other post-retrieval QPP methods by leveraging information from previously submitted queries that are in close proximity within the Query Space. The proposed method benefits from the enriched representation of the inputs provided by a neural-based model by encoding the performance of these queries, identified based on their distance. We show that the accuracy of query performance estimation improves significantly by incorporating contextual information from nearby queries, as opposed to relying solely on individual query characteristics.

3 Proposed Approach

The primary hypothesis underlying our work is that the deterministic nature of retrieval systems enables the generalization of query performance across similar queries within a structured representation of query space. Specifically, we posit that historical query performance data can effectively serve as a proxy for estimating the performance of new, unobserved queries when contextual and semantic similarities are adequately modeled. To this end, we introduce a *Query Space Distance-Based Query Performance Prediction* (QSD-QPP) framework, which systematically constructs a high-dimensional *Query Space* where queries are embedded. This space is enriched by incorporating historical query performance measures, allowing for the quantification of relationships between new queries and their nearest neighbors in this space.

By leveraging this structured representation, QSD-QPP facilitates robust performance prediction for two scenarios: (1) *Pre-retrieval QPP*, which interpolates performance estimates based on query-level distances and performance trends from similar historical queries, and (2) *Post-retrieval QPP*, which augments the prediction process with contextual embeddings of retrieved documents and query associations. The hypothesis behind our work asserts that our approach not only capitalizes on the deterministic characteristics of retrieval methods but also exploits the inherent continuity in the performance of queries in the query space, enabling accurate and generalizable predictions across diverse query sets.

3.1 Proposed Query Space Distance-Based QPP Framework

A general QPP framework can be formulated as a function μ, which takes four key arguments: the input query q, the retrieval method R, the collection of documents C, and the ranked list of documents D_q retrieved by R in response to q. The function μ predicts the retrieval effectiveness $\widehat{M_q}$ for query q, given the quintuple as follows:

$$\widehat{M_q} = \mu(q, R, C, D_q)$$

In a pre-retrieval QPP setting, the set of retrieved documents D_q is not available since performance prediction is done prior to retrieval. Thus, the function μ can only process a triple $\mu(q, R, C)$, relying solely on the input query

q, the retrieval method R, and the document collection C. In contrast and within the context of post-retrieval QPP, the ranked list of documents D_q retrieved by R is available; hence, $\mu(q, R, C, D_q)$.

The fundamental concept underlying our proposed QSD-QPP method is that both sparse and dense retrieval methods exhibit deterministic behavior, meaning they consistently retrieve and rank documents in a similar manner for queries with shared characteristics. This deterministic property allows the generalization of query performance by leveraging similarities between new and past queries. The main premise is that queries that lie within the same subspaces will exhibit similar retrieval effectiveness. Therefore, the performance of a new query can be inferred based on the performance of queries within that query's subspace.

Our hypothesis spans two extremes in the query space spectrum: *(1) Exact Matches:* At one extreme, a query can be identical to some query that has already been observed in the historical data, and its retrieval performance is precisely known. In such cases, the prediction process simplifies to a deterministic retrieval from the query space, where the precomputed performance metric for the exact match can be directly leveraged. This scenario underscores the deterministic nature of retrieval systems, as identical queries inherently produce identical outcomes under fixed retrieval settings. *(2) Unobserved Queries:* At the other extreme, a query is entirely new, residing in an underpopulated or unpopulated region of the query space. Such queries often arise in dynamic contexts, such as emerging real-world events or unconventional information needs, where no proximal neighbors exist within their query subspace. In such cases, the prediction process is inherently probabilistic and relies on extrapolation. Specifically, retrieval effectiveness can be inferred by taking the performance metrics of the most semantically proximate queries available in the query space into account. These neighbors, though distant, provide the basis for approximating performance trends by leveraging the continuity and structural coherence of the query space. This process must account for diminishing confidence with increasing distance, ensuring that the estimation reflects the uncertainty inherent in extrapolating from distant points in the query space.

In **pre-retrieval QPP**, the challenge will be to estimate the performance of the retrieval method based solely on the input query q, the retrieval method R, and the collection C without having access to any retrieved documents. Within our proposed framework, the goal will be to predict query performance by interpolating the effectiveness of past queries that reside within a similar subspace, comprising queries in the proximity of the input query. We denote the variation of our proposed framework for pre-retrieval QPP as QSD-QPP$_{\mathtt{Pre}}$.

For **post-retrieval QPP**, the objective is to obtain a more accurate prediction of query performance using additional information from the set of retrieved documents D_q. The quality and characteristics of these documents provide valuable signals about potential retrieval performance. Hence, post-retrieval QPP exploits the relationship between the query and the retrieved documents, as well as the historical performance of queries within similar query subspaces to estimate query performance. We refer to the variation of our proposed framework for post-retrieval QPP as QSD-QPP$_{\mathtt{Post}}$.

In both scenarios, the effectiveness of the input query can be generalized by utilizing past performance data. In the pre-retrieval setting, this involves leveraging the query's representation alone, while in the post-retrieval setting, additional contextual information from the retrieved documents is also incorporated.

To effectively predict query performance, we introduce the concept of a *Query Space*, a high-dimensional embedding space derived from a set of historical queries and their associated retrieval effectiveness measures. The Query Space facilitates search by identifying a dynamically constructed *subspace* of queries within a given distance threshold from the input query. Let us define $Q = \{q_1, q_2, ..., q_n\}$ to be a list of n previously retrieved queries and $M_Q = \{M_{q_1}, M_{q_2}, ..., M_{q_n}\}$ to be a list of computed retrieval effectiveness measures where M_{q_i} represents the effectiveness score of query q_i with respect to a chosen evaluation metric M. In order to build the Query Space, a translation function $\mathcal{T}$ is used to map each query $q_i \in Q$ into a fixed-length vector representation (d) within a high-dimensional embedding space, denoted as the embedding vector $\mathbf{e}(.)$ as defined below:

$$\mathcal{T}(q_i) = \mathbf{e}(q_i), \quad \mathcal{T} : Q \to \mathbb{R}^d. \tag{1}$$

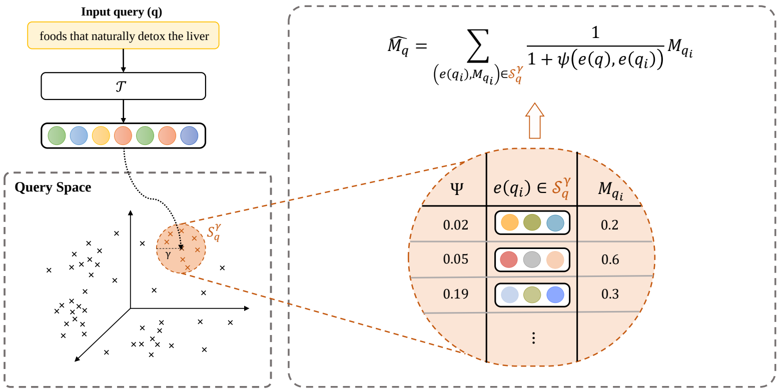
Input query (q)
foods that naturally detox the liver
$\mathcal{T}$
Query Space
$\mathcal{S}_q^\gamma$
γ
$\widehat{M_q} = \sum_{(e(q_i),M_{q_i})\in\mathcal{S}_q^\gamma} \frac{1}{1+\psi(e(q),e(q_i))} M_{q_i}$
Ψ
$e(q_i) \in \mathcal{S}_q^\gamma$
M_{q_i}
0.02
0.2
0.05
0.6
0.19
0.3
⋮

Fig. 2. The overview of the pre-retrieval instantiation of our QSD-QPP method, i.e., QSD-QPP$_{\texttt{Pre}}$.

The Query Space $\mathcal{QS}$ is constructed by applying the translation function $\mathcal{T}$ to map all past queries in Q into a d-dimensional vector space, as defined below:

$$\mathcal{QS} = \left\{ \left(\mathcal{T}(q_i), M_{q_i}\right) \mid q_i \in Q,\ M_{q_i} \in M_Q \right\}, \quad \mathcal{QS} \subseteq \mathbb{R}^d \times \mathbb{R}. \tag{2}$$

For a new query q, represented as $\mathbf{e}(q)$, we dynamically construct a *subspace* $\mathcal{S}_q^\gamma$ by identifying all queries in $\mathcal{QS}$ whose distance from $\mathbf{e}(q)$ is below a specified distance threshold γ. Formally:

$$\mathcal{S}_q^\gamma = \left\{ \left(\mathbf{e}(q_i), M_{q_i}\right) \in \mathcal{QS} \mid \psi\left(\mathbf{e}(q), \mathbf{e}(q_i)\right) \leq \gamma \right\}, \tag{3}$$

where $\psi\left(\mathbf{e}(q), \mathbf{e}(q_i)\right)$ is a distance function defined as:

$$\psi\left(\mathbf{e}(q), \mathbf{e}(q_i)\right) = \left(\sum_{j=1}^{d} \left(\mathbf{e}(q)[j] - \mathbf{e}(q_i)[j]\right)^2 \right)^{\frac{1}{2}}. \tag{4}$$

The parameter γ, referred to as the *distance threshold*, determines the maximum permissible proximity between the input query q and past queries in the Query Space for inclusion in the subspace $\mathcal{S}_q^\gamma$. This ensures that only queries that are within the vicinity of q in the query space to be considered relevant for query performance prediction.

Let $k = |\mathcal{S}_q^\gamma|$ represent the number of queries in the subspace $\mathcal{S}_q^\gamma$. We show how these k previously retrieved queries in $\mathcal{S}_q^\gamma$ for query q can be incorporated into both QSD-QPP$_{\texttt{Pre}}$ and QSD-QPP$_{\texttt{Post}}$ scenarios to effectively estimate $\widehat{M_q}$ of query q. The details of how our main hypothesis is applied in pre-retrieval and post-retrieval QPP scenarios are explained in Sections 3.2 and 3.3, respectively.

3.2 Pre-retrieval Instantiation

Within the context of a pre-retrieval QPP scenario, the pre-retrieval instantiation of our proposed framework, QSD-QPP$_{\texttt{Pre}}$, follows two steps for each input query (see Figure 2):

(1) *Dynamic Subspace Construction*: The input query q is embedded into the high-dimensional Query Space $\mathcal{QS}$ as $\mathbf{e}(q)$, which captures both semantic and syntactic features of the query. To localize the inference

process, a query subspace $\mathcal{S}_q^\gamma$ is dynamically constructed, comprising historical queries within a predefined distance threshold γ from $\mathbf{e}(q)$ as described in Equation 3. By filtering queries based on their geometric distances in $\mathcal{QS}$, $\mathcal{S}_q^\gamma$ provides a focused subspace conducive to robust prediction.

(2) *Performance Prediction*: Once the subspace $\mathcal{S}_q^\gamma$ is constructed, the retrieval effectiveness of q, denoted as $\widehat{M}_q$, is estimated by aggregating the effectiveness scores M_{q_i} of the queries in $\mathcal{S}_q^\gamma$, weighted by their respective proximities to $\mathbf{e}(q)$. This process is guided by an interpolation function $\mathcal{I}(\mathbf{e}(q), \mathcal{S}_q^\gamma)$, defined as:

$$\widehat{M}_q = \mathcal{I}(\mathbf{e}(q), \mathcal{S}_q^\gamma) = \sum_{(\mathbf{e}(q_i), M_{q_i}) \in \mathcal{S}_q^\gamma} \omega\left(\mathbf{e}(q), \mathbf{e}(q_i)\right) M_{q_i}. \tag{5}$$

Here, $\omega(\mathbf{e}(q), \mathbf{e}(q_i))$ represents a weighting function that assigns importance to each historical query q_i based on its distance to $\mathbf{e}(q)$. A common choice for $\omega(\cdot)$ is an inverse distance-based weighting scheme [11], expressed as:

$$\omega(\mathbf{e}(q), \mathbf{e}(q_i)) = \frac{1}{1 + \psi\left(\mathbf{e}(q), \mathbf{e}(q_i)\right)}. \tag{6}$$

In the absence of weighting (uniform interpolation), the predicted retrieval effectiveness can be simplified to the arithmetic mean of the effectiveness scores within $\mathcal{S}_q^\gamma$:

$$\widehat{M}_q = \frac{1}{|\mathcal{S}_q^\gamma|} \sum_{(\mathbf{e}(q_i), M_{q_i}) \in \mathcal{S}_q^\gamma} M_{q_i}, \tag{7}$$

where $\widehat{M}_q$ is the predicted score for the performance of q based on the average performance of the past queries inside its query subspace $\mathcal{S}_q^\gamma$. The principles underpinning $\texttt{QSD-QPP}_{\texttt{Pre}}$ resonate with other tasks within the context of machine learning, where task-specific predictions are often derived by leveraging geometric proximity in a learned representation space. For instance, in semi-supervised regression [39, 70], methods like embedding space mapping and pseudo-label smearing [25, 41] have been proposed to ensure that unlabeled data contribute positively to the prediction process. Similarly, studies on unsupervised feature selection [62] exemplify the concept of learning based on similar subspaces. Our proposed $\texttt{QSD-QPP}_{\texttt{Pre}}$ builds upon these ideas by dynamically constructing a focused query subspace for each input query, enabling localized inference based on historical query performance.

3.3 Post-retrieval Instantiation

In the post-retrieval QPP scenario, our proposed method, $\texttt{QSD-QPP}_{\texttt{Post}}$, extends the pre-retrieval approach by integrating additional layers of contextual and historical information to refine the estimation of query effectiveness (see Figure 3). This approach incorporates three core components into an enriched query representation:

(1) *Contextualized Query Representation*: The individual characteristics of the input query q are captured through its contextualized embedding representation $\mathbf{e}(q)$. This ensures that queries with similar semantic subspaces are mapped to proximate regions in the high-dimensional embedding space.

(2) *Document Characteristics*: Information about the documents retrieved for q, denoted as $D_q = \{d_1, d_2, \ldots, d_m\}$, is incorporated through their contextualized embeddings. These embeddings represent the properties of the retrieved documents, which serve as critical signals for inferring the effectiveness of the query.

(3) *Historical Query Association*: The relationship between q and previously observed queries is captured by measuring the geometric distance between $\mathbf{e}(q)$ and the embeddings of historical queries $\{\mathbf{e}(q_i) \mid q_i \in Q\}$ in the Query Space $\mathcal{QS}$. Queries with smaller distances to $\mathbf{e}(q)$ are considered more relevant, and their known effectiveness scores M_{q_i} contribute significantly to the prediction of q's retrieval effectiveness.

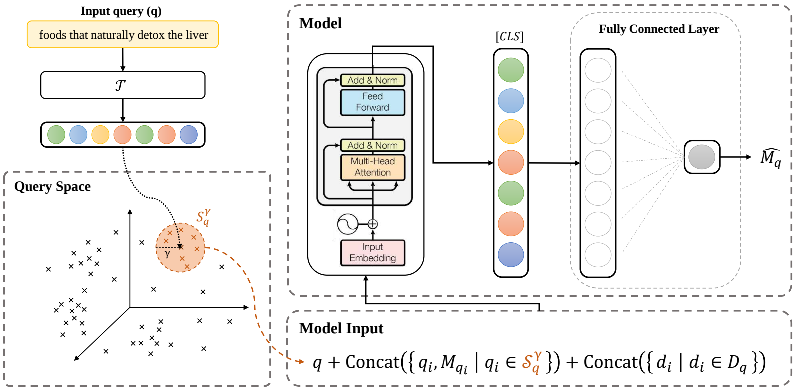

Fig. 3. The overview of the post-retrieval instantiation of our QSD-QPP method, i.e., QSD-QPP$_{\text{Post}}$.

In particular, for an input query q, we let the cross-encoder architecture estimate the performance of q, through regression, based on the contextualized representation of q, D_q, and a set of top-k historical queries retrieved from the Query Space $\mathcal{QS}$ with known effectiveness. The historical queries are selected based on their proximity to $\mathbf{e}(q)$ within the query subspace $\mathcal{S}_q^\gamma$, as described in Equation 3. We encode q, D_q, and queries within $\mathcal{S}_q^\gamma$ and their known retrieval effectivness by concatenating them using a special separator token. This concatenated vector is then passed through a linear transformation to estimate the retrieval effectiveness. Formally, the prediction of the retrieval effectiveness $\widehat{M_q}$ is defined as follows:

$$\widehat{M_q} = \text{Linear}\left(\mathcal{T}\left(\text{Concat}\left(q, \text{Concat}\left(\left\{q_i, M_{q_i} \mid q_i \in \mathcal{S}_q^\gamma\right\}\right), \text{Concat}\left(\left\{d_i \mid d_i \in D_q\right\}\right)\right)\right)\right). \tag{8}$$

For training QSD-QPP$_{\text{Post}}$, we leverage a sigmoid activation layer and optimize the model with a one-class binary cross-entropy loss function. Let M_q denote the target ranking metric, such as average precision, which serves as the ground-truth effectiveness score for query q. To minimize the prediction error, we define the following loss function for training the model to predict query performance:

$$\ell(\widehat{M_q}, M_q) = -w\left[M_q \cdot \log\left(\sigma(\widehat{M_q})\right) + (1 - M_q) \cdot \log\left(1 - \sigma(\widehat{M_q})\right)\right], \tag{9}$$

where $\sigma(\cdot)$ represents the sigmoid function. This loss function effectively guides the model to learn a fine-grained regression of query performance based on historical query information.

In the context of QSD-QPP$_{\text{Post}}$, the reliance on historical queries aligns with the broader concept of subspace-based learning [10, 32], where the predictive power stems from modeling relationships between a target and its most relevant neighbors. It closely resembles methods that maximize the likelihood of correct predictions by leveraging neighborhood samples, aligning the latent space structure with task-specific similarities [64] or techniques such as Elastic Net Subspace Clustering (ENSC) [47, 52], which represent each data point as a sparse combination of its neighbors, enabling more robust modeling of relationships within local subspaces. As such, QSD-QPP$_{\text{Post}}$ goes beyond existing post-retrieval QPP methods by enabling the post-retrieval model to learn not only the association between the query and its relevant retrieved document but also its relationship with other historical queries within similar query subspaces.

4 Experiments

4.1 Dataset

For evaluation purposes, we use the widely adopted MS MARCO passage collection dataset [45] that consists of 8.8 million documents and over 500k queries with at least one relevance judgments. We used these 500k queries to build our Query Space $\mathcal{QS}$ introduced in Section 3.1 using different language models to implement $\mathcal{T}(.)$. The MS MARCO collection also includes a Dev query set containing 6,980 queries on which we test the performance of our proposed $\texttt{QSD-QPP}_{\texttt{Pre}}$ and $\texttt{QSD-QPP}_{\texttt{Post}}$ methods. Besides, we use three other TREC Deep Learning track query sets, namely (1) TREC DL 2019 [14] that consists of 43 queries, (2) TREC DL 2020 [12] consisting of 54 queries, and (3) DL-Hard, which includes 50 queries [42]. The main differences between these three TREC datasets and the MS MARCO Dev set are (1) the number of queries included in each dataset; (2) the number of available relevance judgments per query and (3) the binary vs. graded levels of judgments. While most of the MS MARCO Dev set queries have sparse relevance judgments (almost one relevant document per query), queries of the other three Deep learning Tracks have numerous relevance judged documents. Therefore, testing our proposed methods on this range of datasets can show the robustness of our approach in terms of both the query size and the number of relevant documents per query.

4.2 Evaluation Methodology and Metrics

We evaluate the performance of our proposed methods using a common QPP evaluation strategy, that is, computing the linear and ranked-based correlation between the predicted performance given by our methods and the actual performance of queries over the list of documents. For comparison, we measure linear correlation by Pearson's ρ, rank correlation by Kendall's τ, and rank correlation coefficient by Spearman ρ. As a common practice [9, 28], we predict the performance of the widely used BM25 retriever for each query using the metrics reported in [9] for the pre-retrieval setting and [37] for the post-retrieval setting. A higher correlation value shows more accurate query performance prediction.

4.3 Baselines

To evaluate the performance of our proposed pre-retrieval and post-retrieval QPP methods, we compare them against a variety of established baselines in each category.

4.3.1 Pre-retrieval Baselines. We adopt widely used pre-retrieval query performance prediction baselines that have shown promising performance on various well-known corpora and query sets [6, 9]. These baselines can be categorized into two categories as follows:

(1) **Traditional Baselines**: The majority of pre-retrieval QPP methods that belong to this group rely on term frequencies, index statistics and analyze how query terms are distributed within a collection. Two commonly used statistics in this context are the inverse document frequency (IDF) and the inverse collection term frequency (ICTF) of the query terms [40]. These statistics are widely recognized as measures of query terms' relative importance. The simplified clarity score (SCS) [29] is a metric that measures query specificity based on the Kullback-Leibler divergence between the query language model and the collection language model. Zhao et al. [68] proposed SCQ in which they calculate the vector-space-based query similarity by treating the collection as a single document composed of concatenated individual documents. PMI (Pointwise Mutual Information) is a term-relatedness based predictor, which analyzes the co-occurrence statistics of terms [28]. These predictors assume that frequent co-occurrence of query terms in the collection indicates good performance, assuming that all query terms are related to the same topic. We also consider VAR [68], which is a coherency-based method that works based on the variance of the term weight distribution over documents containing that term.

(2) **Neural-embedding based Baselines**: Closeness Centrality (CC), Degree Centrality (DC) and Inverse Edge Frequency (IEF) are based on the idea that a query term surrounded by many other similar terms in the embedding space is more specific, while a term with fewer closely surrounded terms is more generic and a more specific query is easier to address than a more generic one since a generic query might have several aspects to satisfy [6–8]. In these baselines, an ego network representation helps capture the context of a term using neural embedding representations and by considering other similar terms as alter nodes. These metrics helps determine the specificity of the represented term, which can indicate the difficulty of a query as well. Another neural-based metric is $P_{clarity}$ which works based on the idea that the number of clusters around the neighborhood of a query term is a potential indicator of its specificity [57]. MRL is another pre-retrieval QPP method that utilizes the learned multivariate vector representations of a query for predicting its performance [66]. We also included the most recently proposed baseline by Khodabakhsh et al. [38], the BertPE method, which utilizes a cross-encoder architecture to fine-tune a pre-trained language model for directly predicting query performance.

We note that for all baselines, we used the best-performing hyperparameters on reported in the original papers usually tuned on other TREC collections such as Robust04 and Clueweb09 and 12 and Gov2 as reported in [6, 9]. Additionally, some baselines works on per query term level. In such cases and where required, the maximum aggregation function over all query terms has been applied to achieve the predicted performance.

4.3.2 Post-retrieval Baselines. We evaluate our post-retrieval QPP model against the leading baselines. Similar to pre-retrieval baselines, these methods can be categorized into traditional baselines as well as neural-embedding based baselines as follows:

(1) **Traditional Baseline**: This group of baselines mostly rely on the statistical features of the retrieved document and the query. For example, Clarity, has been proposed based on the KL-divergence between language models derived from retrieved documents and the overall corpus. We also consider a variety of score-based methods such as WIG [69], NQC [60], $n(\sigma\%)$ [16], RSD [56], and SMV [63], which predict query performance by analyzing various statistics of the retrieval scores of top-ranked documents. In addition, we consider the Utility Estimation Framework (UEF) [59] which is designed to complement robust QPP baselines such as NQC [60].

(2) **Neural-embedding-based Baselines**: NeuralQPP [67] is a supervised QPP method that leverages existing unsupervised QPP methods as signals for weakly-supervised learning. Additionally, NQA-QPP [27] employs a BERT model to learn representations of queries and documents, while BERT-QPP [3] fine-tunes BERT to directly predict query retrieval scores. QppBERT-PL[20], a recent BERT-based method, uses point-wise training on individual queries and list-wise training over top-ranked pseudo-relevant documents. Furthermore, we include QPP-PRP [61], which assesses the performance of neural rankers by evaluating the agreement level between a pairwise neural reranker, such as DuoT5 [50], and the ranked list generated by the neural ranker for a given query.

4.4 Experimental Setup and Hyperparameters

In this section, we provide a detailed description of our configurations and hyperparameters used in our experiments. In order to index the queries within the Query Space, we adopt an inverted model with asymmetric distance computation [33]. This enables the dynamic construction of a subspace comprising queries that are within a defined proximity to an input query q with $\mathbf{e}(q) \in R^d$. We leveraged the implementation provided by FAISS [34] which offers efficient mechanisms for searching over high-dimensional vectors. For each target query, we adopt the L^2 distance function for $\psi(.,.)$ as recommended in [35] to identify and locate queries within the Query Space with the closest proximity, forming a dynamically constructed subspace.

We used a 24GB NVIDIA GeForce RTX 3090 GPU to train the QSD-QPP$_{\text{Post}}$ and to create the embeddings that were needed for building $\mathcal{QS}$. It took approximately 90 minutes for our QSD-QPP$_{\text{Post}}$ model to be fine-tuned on the training set. For the purpose of creating vector representations and searching the Query Space to construct the subpace, we use four language models including (1) all-mpnet-base-v2, (2) all-MiniLM-v2,

(3) paraphrase-MiniLM-v2, and (4) deberta-v3-base. The first three language models are trained based on a Sentence Transformer [53], while deberta-v3-base introduces disentangled attention and an enhanced mask decoder to improve over BERT and RoBERTa. To predict the performance of a query based on k queries within its constructed subspace, we investigate the impact of hyperparameter k with different values in range of 1 to 10.

Additionally, in the Ablation Study section, we investigate the impact of Query Space size by building Query Space with varying numbers of queries included in them. This analysis allows us to assess how the size of Query Space affects the performance of $\texttt{QSD-QPP}_{\texttt{Pre}}$ and $\texttt{QSD-QPP}_{\texttt{Post}}$ methods.

4.4.1 Expanding Relevance Judgements. One of the difficulties is ensuring a balance in relevance judgements between Query Space and various test sets. Given that TREC DL 2019, TREC DL 2020, and DL-Hard query sets have several relevant documents for each query, whereas other query sets like MS MARCO training queries and the dev set in MS MARCO have an average of only 1.06 relevant documents per query, it is necessary to close this gap between Query Space and the test sets. Therefore, we increase the number of relevant documents for the queries in Query Space. For this purpose, as suggested in the literature [51], top-1000 BM25 documents are first re-ranked using a neural-based pointwise reranker, i.e., MonoBERT [46] and then DuoT5 [51] is used for re-ranking the top-50 ranked documents returned by MonoBERT. Finally, the top-50 documents are selected and added to the existing list of relevant documents.

4.4.2 Subspace Construction. To construct the subspace for each query, we utilized the weights of the aforementioned models and defined the distance threshold γ to ensure that each query retrieves 10 queries with the smallest distances from the Query Space. This approach has been applied to all sets of the queries we had. By providing our Query Space as a FAISS index, we were able to search over it and create a subspace with 10 queries for each input query. We cached all 10 queries within the subspace of each query to make the pipeline cost-efficient and prevent running the retriever every time we changed a parameter (i.e., number of the queries within the subspace of each query that has been fed to our model).

4.5 Comparison with Baselines

In the following subsections, we report the performance of our $\texttt{QSD-QPP}_{\texttt{Pre}}$ and $\texttt{QSD-QPP}_{\texttt{Post}}$ methods in comparison with the baselines.

4.5.1 Pre-retrieval Model. We compare the performance of $\texttt{QSD-QPP}_{\texttt{Pre}}$ with the baselines over four different datasets and three correlation measures. Table 1 presents the results of our experiments. To implement $\texttt{QSD-QPP}_{\texttt{Pre}}$ in Table 1, we used the paraphrase-MiniLM-v2 language model to obtain contextualized representations of the query while considering $k = 10$ queries within the subspace for each input query. The performance of other variations of $\texttt{QSD-QPP}_{\texttt{Pre}}$ based on different k values and other language models are reported in Section 5 of this paper.

Based on the results in Table 1, we find that our method consistently outperforms the baselines across all correlation measures and datasets. A notable observation is that, in general, the query performance predictors exhibit inferior performance on the MS MARCO Dev set compared to the other three TREC datasets. For instance, IDF, one of the top-performing baselines, and our proposed method register a Pearson correlation of 0.117 and 0.219, respectively, on the MS MARCO Dev set. In contrast, these methods achieve a Pearson correlation of 0.440 and 0.483, respectively, on TREC DL 2019. A similar trend is noticeable across the other three DL datasets. We speculate that this may be due to incomplete relevance judgements on the MS MARCO Dev set. Given that queries in this dataset have, on average, only one relevant judged document, while judgements for the other three query sets are extremely thorough, these incomplete judgements might introduce noise into the performance, which may consequently affect the QPP performances.

Table 1. Performance comparison between our QSD-QPP$_{\mathtt{Pre}}$ method and various pre-retrieval QPP baselines over MS MARCO Dev small and TREC DL 2019 (Upper Table) and Trec DL 2020 and TREC DL HARD (Lower Table) in terms of Pearson ρ, Kendall τ, and Spearman ρ. The highest value in each column is marked in bold and superscript † denotes that the correlation values obtained by QSD-QPP$_{\mathtt{Pre}}$ are statistically significant with p-value of 0.05 using a paired t-test.

QPP Method	**MS MARCO Dev small**			**TREC DL 2019**		
	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
SCS	0.021	0.058	0.085	0.471	0.262	0.354
$P_{clarity}$	0.052	0.007	0.009	0.109	0.119	0.139
VAR	0.067	0.081	0.119	0.290	0.141	0.187
PMI	0.030	0.033	0.048	0.155	0.065	0.079
IDF	0.117	0.138	0.200	0.440	0.276	0.389
SCQ	0.029	0.022	0.032	0.395	0.114	0.157
ICTF	0.105	0.136	0.198	0.435	0.259	0.365
DC	0.071	0.044	0.065	0.132	0.083	0.092
CC	0.085	0.066	0.076	0.079	0.068	0.023
IEF	0.110	0.090	0.118	0.140	0.090	0.134
MRL	0.022	0.046	0.067	0.176	0.079	0.140
BertPE	0.010	0.003	0.004	0.255	0.190	0.281
QSD-QPP$_{\mathtt{Pre}}$	**0.219**†	**0.214**†	**0.309**†	**0.483**†	**0.349**†	**0.508**†

QPP Method	**TREC DL 2020**			**DL Hard**		
	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
SCS	0.447	0.310	0.448	0.247	0.159	0.240
$P_{clarity}$	0.069	0.052	0.063	0.095	0.209	0.272
VAR	0.047	0.051	0.063	0.023	0.014	0.001
PMI	0.021	0.012	0.003	0.093	0.027	0.042
IDF	0.413	0.236	0.345	0.200	0.197	0.275
SCQ	0.193	0.005	0.004	0.335	0.106	0.152
ICTF	0.409	0.236	0.348	0.192	0.195	0.272
DC	0.100	0.118	0.150	0.155	0.091	0.115
CC	0.172	0.065	0.089	0.155	0.093	0.111
IEF	0.110	0.025	0.037	0.018	0.071	0.139
MRL	0.093	0.078	0.117	0.046	0.052	0.038
BertPE	0.013	0.031	0.038	0.054	0.025	0.050
QSD-QPP$_{\mathtt{Pre}}$	**0.452**†	**0.319**†	**0.457**†	**0.364**†	**0.234**†	**0.340**†

Furthermore, SCS and IDF show superior performance compared to the rest of the baselines, but still trail our proposed QSD-QPP$_{\mathtt{Pre}}$ method by a significant margin. We posit that this is because both metrics operate based on term importance signals, such as inverse document frequency (IDF) and inverse collection term frequency (ICTF) in SCS. As a result, we can conclude that these two signals perform best for these query subsets, as they have previously demonstrated promising performance on other TREC collections [9].

Moreover, two neural-based baselines, $P_{clarity}$ and DC, did not yield impressive results on the four query sets under study. These methods work by identifying terms in the embedding space that are most similar to individual query terms. We hypothesize this is because these methods operate at the level of individual query terms. The queries in our study, which are generated by real users, lack structured form and are heavily dependent on context. Consequently, the signals for term relatedness are weaker compared to those from the more structured queries previously experimented within TREC collections where these two methods exhibited robust performance. In contrast, our approach embeds the entire query into one representation, taking into account the context and all query terms as a unified whole. Two other neural-based baselines, MRL and BertPE, exhibited subpar performance on three of the query sets, showing adequate effectiveness only on TREC DL 2019. The performance of MRL has been extensively evaluated on TREC DL 2019 in its original paper [66], but the method struggles to generalize effectively across other query sets. Similarly, while BertPE demonstrates reasonable effectiveness when evaluated on a combination of datasets as shown in [38], its performance on individual query sets reveals significant weaknesses. However, QSD-QPP$_{\text{Pre}}$ demonstrates consistently superior and stable performance across all four query sets, excelling in making accurate predictions regardless of the query difficulty level.

We also observe that, unlike some of the baselines that achieve satisfactory Kendall correlation but poor Pearson correlation, our model consistently delivers strong performance across all three correlations on all four datasets. This demonstrates the robustness of our method. For instance, when considering the correlation metrics of $P_{clarity}$ on DL-Hard or PMI on TREC DL 2019, it is evident that these methods perform well on either Pearson or Kendall correlations, but not on both. However, in contrast, our proposed method exhibits strong performance across all three correlation metrics and all four datasets. This underscores the robustness and consistency of our proposed QSD-QPP$_{\text{Pre}}$ method, further distinguishing it from the baselines.

4.5.2 Post-retrieval Model. Before presenting our findings, we highlight that the influence of hyperparameters—such as k, which specifies the number of queries considered within the subspace of q—and the choice of language models is thoroughly analyzed in Section 5. Here, and in Table 2, we focus on the results of QSD-QPP$_{\text{Post}}$ using the deberta-v3-base language model with $k = 1$, comparing its performance against state-of-the-art post-retrieval QPP baselines.

Neural-based methods have demonstrated significant improvements over traditional models, showcasing notable advancements in predictive performance. For instance, starting from NeuralQPP, more recent neural models, such as BERT-QPP and qpp-BERT-PL, achieve a substantial boost in Pearson correlation, improving by approximately 0.37 on the Dev small dataset. Despite these advancements, many baseline methods exhibit inconsistencies across datasets. For example, SMV, while excelling in rank correlation on DL 2020, fails to maintain competitive performance on MS MARCO Dev. Similarly, qpp-BERT-PL, which outperforms others on MS MARCO Dev, struggles on DL-Hard, and qpp-PRP, despite strong performance elsewhere, performs poorly on DL 2019. In contrast, our proposed QSD-QPP$_{\text{Post}}$ delivers consistently robust results across all datasets and metrics. This reliability highlights its adaptability to varying data distributions, ensuring steady performance even in the face of diverse and challenging conditions.

Despite the demonstrated strengths of neural-based methods, including their ability to significantly improve predictive performance, our experiments with the TREC DL 2020 dataset reveal that achieving universal dominance across all metrics remains elusive. While models like BERT-QPP and qpp-BERT-PL show remarkable progress, no single approach consistently excels across Pearson, Kendall, and Spearman correlations. For example, NQA-QPP performs well in Pearson correlation, while SMV leads in Kendall and Spearman correlations, underscoring the fragmented nature of current state-of-the-art baselines. In this context, our proposed QSD-QPP$_{\text{Post}}$ offers a competitive alternative, delivering results on par with supervised baselines that fine-tune contextual embedding models, such as BERT-QPP, qpp-BERT-PL, and qpp-PRP. Interestingly, traditional non-contextualized models

Table 2. Performance comparison between our QSD-QPP$_{\text{Post}}$ method and various post-retrieval QPP baselines over MS MARCO Dev small and TREC DL 2019 (Upper Table) and Trec DL 2020 and TREC DL HARD (Lower Table) in terms of Pearson ρ, Kendall τ, and Spearman ρ. The highest value in each column is marked in bold and superscript † denotes that the correlation values obtained by QSD-QPP$_{\text{Post}}$ are statistically significant with p-value of 0.05 using a paired t-test.

	MS MARCO Dev small			**TREC DL 2019**		
QPP Method	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
Clarity	0.149	0.258	0.345	0.149	0.099	0.126
WIG	0.154	0.170	0.227	0.331	0.260	0.348
QF	0.170	0.210	0.264	0.210	0.164	0.217
NeuralQPP	0.193	0.171	0.227	0.173	0.111	0.134
n($\sigma_{\%}$)	0.221	0.217	0.284	0.195	0.120	0.147
RSD	0.310	0.337	0.447	0.362	0.322	0.469
SMV	0.311	0.271	0.357	0.375	0.269	0.408
NQC	0.315	0.272	0.358	0.384	0.288	0.417
UEF$_{NQC}$	0.316	0.303	0.398	0.359	0.319	0.463
NQA-QPP	0.451	0.364	0.475	0.386	0.297	0.418
BERT-QPP	0.517	0.400	0.520	0.404	0.345	0.472
qpp-BERT-PL	0.520	0.413	0.522	0.330	0.266	0.390
qpp-PRP	0.302	0.311	0.412	0.090	0.061	0.063
QSD-QPP$_{\text{Post}}$	**0.573**†	**0.434**†	**0.561**†	**0.434**†	**0.412**†	**0.508**†

	TREC DL 2020			**TREC DL Hard**		
QPP Method	Pearson	Kendall	Spearman	Pearson	Kendall	Spearman
Clarity	0.360	0.215	0.296	0.271	0.229	0.332
WIG	0.204	0.117	0.166	0.310	0.158	0.226
QF	0.358	0.266	0.366	0.295	0.240	0.340
NeuralQPP	0.248	0.129	0.179	0.289	0.159	0.224
n($\sigma_{\%}$)	0.480	0.329	0.478	0.371	0.256	0.377
RSD	0.426	0.364	0.508	0.460	0.262	0.394
SMV	0.450	**0.391**	**0.539**	0.495	0.289	0.440
NQC	0.464	0.294	0.423	0.466	0.267	0.399
UEF$_{NQC}$	**0.511**	0.347	0.476	0.507	0.293	0.432
NQA-QPP	0.507	0.347	0.496	0.348	0.164	0.255
BERT-QPP	0.467	0.364	0.448	0.491	0.289	0.412
qpp-BERT-PL	0.427	0.280	0.392	0.432	0.258	0.361
qpp-PRP	0.189	0.157	0.229	0.321	0.181	0.229
QSD-QPP$_{\text{Post}}$	0.462	0.318	0.448	**0.519**†	**0.318**†	**0.459**†

like SMV and RSD still outperform neural-based methods in rank correlation metrics for this dataset, further highlighting the challenges in achieving consistent and universal superiority.

Table 3. Comparing our QSD-QPP$_{\text{Pre}}$ and QSD-QPP$_{\text{Post}}$ methods with various pre-retrieval and post-retrieval QPP baselines based on their run time per query in seconds.

Pre-retrieval Method	SCS	$P_{clarity}$	VAR	PMI	IDF	SCQ	ICTF	DC	CC	IEF	MRL	BertPE	QSD-QPP$_{\text{Pre}}$
Run Time (sec)	0.006	1.966	0.031	0.009	0.001	0.002	0.001	0.556	0.344	0.531	0.008	0.003	0.009

Post-retrieval Method	Clarity	WIG	NeuralQPP	NQC	UEF_NQC	NQA-QPP	BERT-QPP	qpp-PRP	qpp-BERT-PL	QSD-QPP$_{\text{Post}}$
Run Time (sec)	1.130	0.017	0.213	0.003	0.088	0.025	0.003	0.827	0.435	0.003

4.6 Query Latency

In this section, we evaluate the efficiency of our proposed methods by comparing their query latency with that of baseline methods. In order to run all methods and calculate inference time, we used an RTX 3090 GPU and an AMD EPYC 7662 CPU. For the sake of the experiments, we first measured the time taken to index all 500k MS MARCO queries using the paraphrase-MiniLM-v2 language model on the GPU, which amounted to a total of 97 seconds. Subsequently, we proceeded to compute the time needed for the subspace construction and query performance prediction using the QSD-QPP methods. Therefore, we measured both the time necessary for the subspace construction process and query performance prediction using our pre-retrieval and post-retrieval QPP methods as follows:

4.6.1 Pre-retrieval Model. Our findings are presented in Table 3 where we also provide comparisons with the baseline methods. As shown in this table, it is evident that the QSD-QPP$_{\text{Pre}}$ method not only outperforms the baselines in terms of performance but also demonstrates remarkable efficiency. These characteristics make it well-suited for real-world applications.

4.6.2 Post-retrieval. The running time per query for our post retrieval QPP method and the baselines are also presented in Table 3. By comparing the runtime of our method with baseline supervised methods, we can observe that our method, QSD-QPP$_{\text{Post}}$, takes slightly more time to process each query compared to BERT-QPP. This is because, as explained in Section 3.3, we augmented the input of our method by concatenating the queries within the subspace of each query and their performance, resulting in a longer input length. As a result, our method should process more input in comparison with BERT-QPP. However, it is still much faster than other models such as NeuralQPP and NQA-QPP, which are neural-based models as well.

5 Ablation Studies

The performance of our proposed QSD-QPP$_{\text{Pre}}$ and QSD-QPP$_{\text{Post}}$ methods may be impacted by the choice of (1) the base language model used for creating the Query Space, (2) the subspace size, defined as the number of queries or samples inside the subspace during inference, (3) the size of the Query Space, which determines the pool of previously retrieved queries available for performance prediction, and (4) the impact of maximum token length on QSD-QPP$_{\text{Post}}$. To better understand these effects, we investigate their impact on the overall performance of our proposed methods in the following sections.

5.1 Impact of Choice of Language Model and Subspace Size

In this section, we first investigate the impact of the base language model and the size of subspace represented as k. To explore the impact of the base language model, we adopt four different large language models, namely (1) all-mpnet-base-v2, (2) all-MiniLM-v2, (3) paraphrase-MiniLM-v2 and (4) deberta-v3-base and build the Query Space independently for each of them and measure the performance of QSD-QPP$_{\text{Pre}}$ and QSD-QPP$_{\text{Post}}$. These

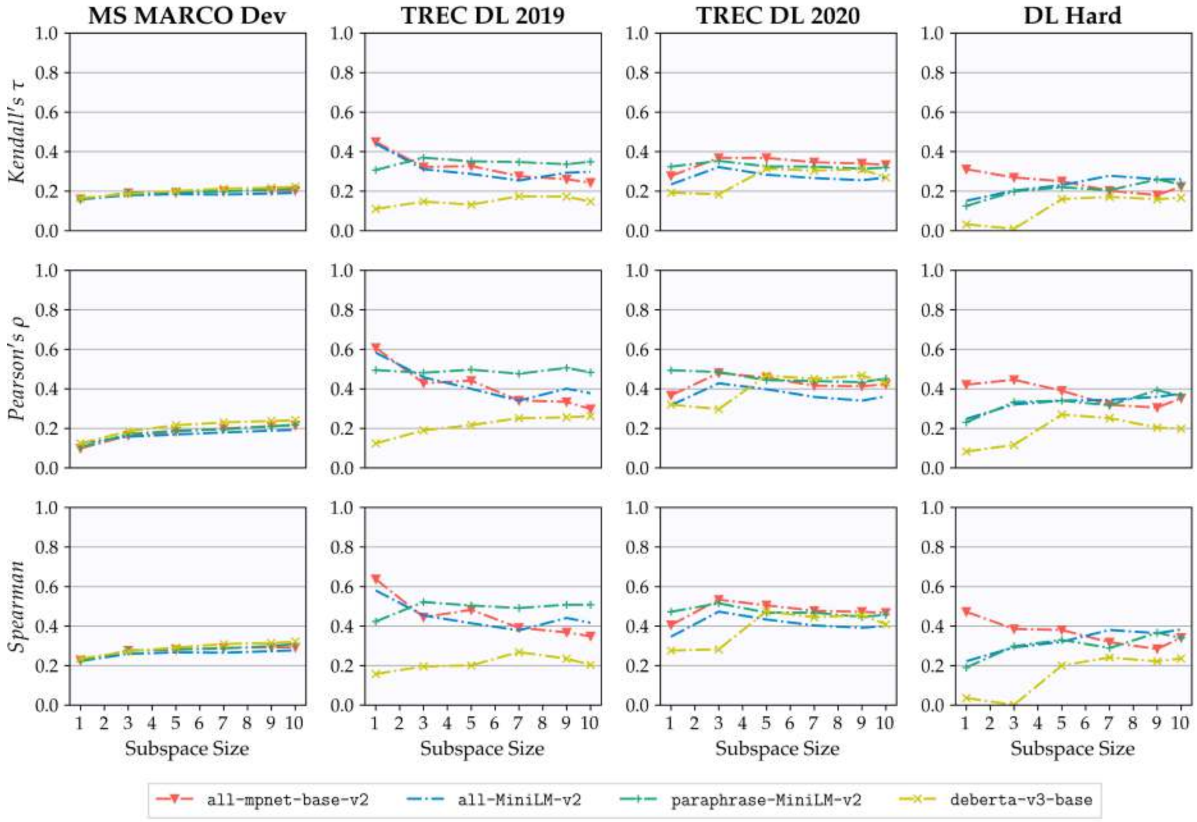

Fig. 4. Impact of Subspace Size and Choice of Language Model across all four datasets on $\texttt{QSD-QPP}_{\texttt{Pre}}$ method.

language models were selected because they have demonstrated effective performance in various downstream IR and NLP tasks, ensuring the generalizability of our findings.

Further, to investigate the impact of subspace size on our proposed methods, we evaluate varying numbers of queries within the subspace, selecting $k = \{1, 3, 5, 7, 9, 10\}$ based on their distance scores relative to the input query, across all four datasets. We limit our analysis to $k \leq 10$ to maintain cost-effectiveness, as our experiments showed that including more than 10 queries in the subspace does not result in significant performance improvements. The results of the variations based on differing language models and subspace size are reported in Figure 4 for $\texttt{QSD-QPP}_{\texttt{Pre}}$ and Figure 5 for $\texttt{QSD-QPP}_{\texttt{Post}}$ respectively. The figures include performance based on Kendall τ and Pearson ρ, and Spearman correlations. The predicted performance and run files of each model have been included on our Github repository at https://github.com/sadjadeb/QSD_QPP.

5.1.1 Pre-retrieval Model. Based on Figure 4, our $\texttt{QSD-QPP}_{\texttt{Pre}}$ method demonstrates stability and robustness across all four datasets, irrespective of the subspace size k or the choice of language model. Among the evaluated models, `Paraphrase-MiniLM-v2` stands out as the most consistent and stable, exhibiting reliable performance

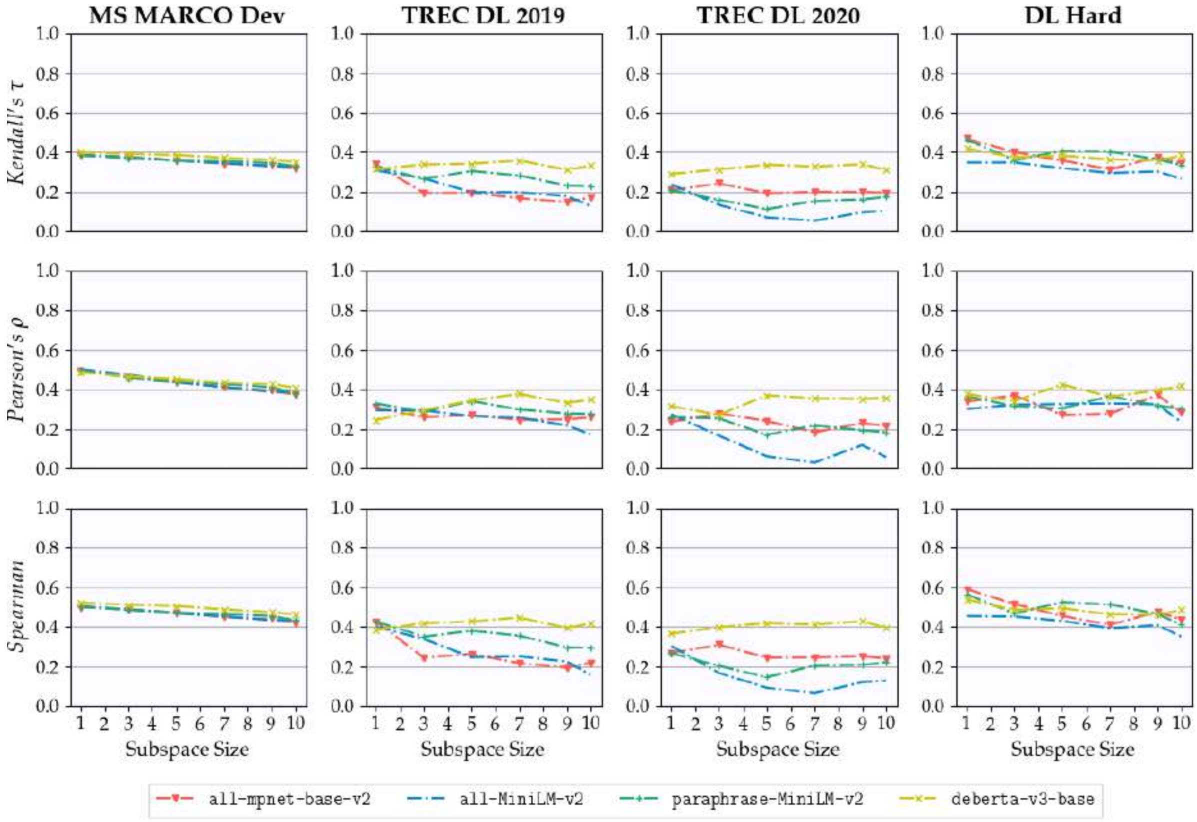

Fig. 5. Impact of Subscpace Size and Choice of Language Model across all four datasets by QSD-QPP$_{\text{Post}}$.

across various configurations. While it does not always achieve the highest scores, its robustness makes it a particularly suitable choice for our purposes. As such, we selected this model for comparison with the baselines in Table 1. We hypothesize that its enhanced reliability stems from being fine-tuned for a relevant task, as `Paraphrase-MiniLM-v2` is specifically optimized for paraphrase mining. An exception to this overall robustness is observed with the `deberta-v3-base` model, which underperforms relative to others when k is below 5 on the TREC DL 2019 and DL Hard datasets. However, as k increases and approaches 10, its performance aligns with that of the other models, demonstrating its ability to adapt to larger subspace sizes.

5.1.2 Post-retrieval Model. Similarly, Figure 5 highlights the robustness of QSD-QPP$_{\text{Post}}$ across various configurations. Interestingly, we observed no significant correlation between increasing the subspace size and the performance of our method across all language models. Notably, augmenting the model's input with more similar queries proved beneficial only when `deberta-v3-base` was employed.

Table 4. Performance of the QSD-QPP$_{\mathtt{Pre}}$ method when various percentage of queries were used for Query Space Construction on MS MARCO Dev queries.

Percentage of Queries	Pearson	Kendall	Spearman
50%	0.200	0.191	0.278
60%	0.200	0.197	0.286
70%	0.196	0.199	0.290
80%	0.216	0.209	0.302
90%	0.215	0.207	0.299
100%	0.219	0.214	0.309

Table 5. Performance of the QSD-QPP$_{\mathtt{Post}}$ method when various percentages of queries were used for Query Space Construction on MS MARCO Dev queries.

Percentage of Queries	Pearson	Kendall	Spearman
50%	0.549	0.421	0.546
60%	0.546	0.426	0.552
70%	0.559	0.426	0.552
80%	0.563	0.429	0.555
90%	0.565	0.429	0.555
100%	0.573	0.434	0.561

5.2 Impact of Query Space size

The performance of our proposed QSD-QPP$_{\mathtt{Pre}}$ and QSD-QPP$_{\mathtt{Post}}$ methods may also be impacted by the size of Query Space used for constructing the subspace. Thus, to explore the impact of Query Space size on the performance of our proposed approaches, we employ a random sampling approach to select various percentages of queries from the pool of 500k MS MARCO queries. For each subset of queries, we construct distinct versions of the Query Space using the paraphrase-MiniLM-v2 language model. Subsequently, we evaluate the performance of QSD-QPP$_{\mathtt{Pre}}$ and QSD-QPP$_{\mathtt{Post}}$ methods on the MS MARCO Dev query dataset.

5.2.1 Pre-retrieval Model. Table 4 illustrates the impact of Query Space size on the performance of our QSD-QPP$_{\mathtt{Pre}}$ method. The results indicate that enlarging the Query Space leads to a slight improvement in performance. Notably, even when utilizing only 50% of $\mathcal{QS}$, our method outperforms all baselines. This demonstrates that, regardless of the size of $\mathcal{QS}$, our approach consistently achieves higher performance than the current baselines.

5.2.2 Post-retrieval Model. Similarly, Table 5 reveals the same pattern for QSD-QPP$_{\mathtt{Post}}$. This model shows similar performance to the pre-retrieval approach and manages to outperform all baselines even when we cut the Query Space in half. This demonstrates that our method is robust and effective across different configurations of the Query Space size, maintaining superior performance compared to existing baselines. This consistency highlights the strength of our approach in achieving high correlation metrics, regardless of the Query Space size.

Overall, we observe that both QSD-QPP$_{\mathtt{Pre}}$ and QSD-QPP$_{\mathtt{Post}}$ show stable and robust performance regardless of the adopted language model, the subspace size k, and the size of constructed Query Space and hence can be reliably used for performing QPP on a range of different query datasets including those with sparse labels (MS MARCO), extensive labels (DL 2019 and 2020), and difficult queries (DL-Hard).

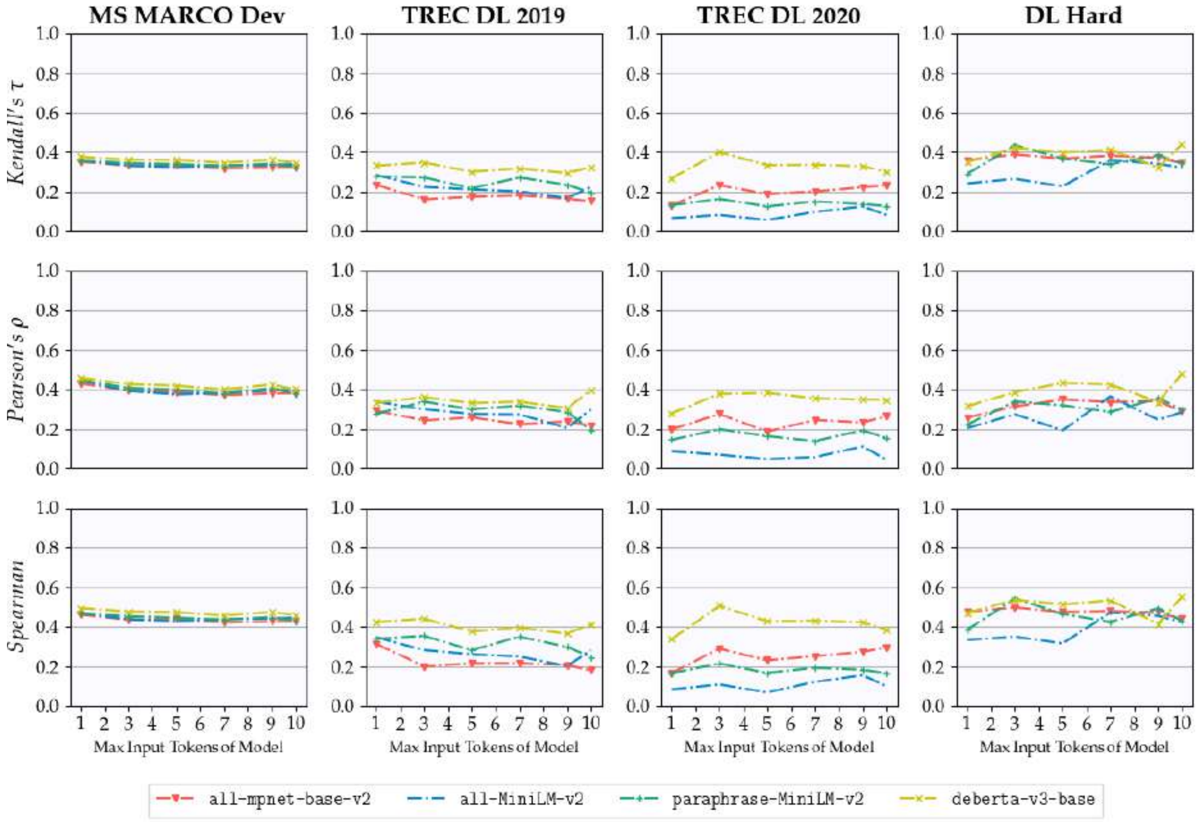

Fig. 6. Impact of Max Token Length across all four datasets by QSD-QPP$_{\text{Post}}$.

5.3 Impact of Maximum Token Length on QSD-QPP$_{\text{Post}}$

In this section, we examine the impact of varying the maximum token lengths by concatenating the top retrieved documents and training and testing our QSD-QPP$_{\text{Post}}$ model with different token configurations. Specifically, we expand the size of D_q from one to ten while ensuring that the input does not exceed 512 tokens, the maximum length processable by our base models. As illustrated in Figure 6, our method's performance on the MS MARCO Dev set decreases as the number of processed tokens increases. For example, the Pearson correlation drops from 0.458 to 0.399 when the maximum token length is increased from 50 to 500. These results suggest that the optimal configuration for $|D_q|$ is 1, as increasing it results in weaker performance compared to the default setting. Nevertheless, this impact is not the same on the TREC DLs. We see improvement over all of these sets the maximum token lengths became larger. This trend has been repeated over all 4 base language models that we used. The performance of our QSD-QPP$_{\text{Post}}$ increases from 0.332 to 0.396, from 0.279 to 0.347, and from 0.315 to 0.478 on the TREC DL-2019, TREC DL-2020, and TREC DL-Hard, respectively.

We note that given our QSD-QPP$_{\text{Pre}}$ method is an unsupervised method and only relies on out-of-the-box language models for building $\mathcal{QS}$, we were only able to evaluate the effect of changing the maximum token numbers that our model used for the QSD-QPP$_{\text{Post}}$ model.

6 Concluding Remarks

This paper introduced the Query Space Distance-Based QPP (QSD-QPP) framework, a novel approach to query performance prediction that leverages the deterministic behavior of retrieval methods and historical query performance data. By referencing semantically similar queries, QSD-QPP provides robust predictions through both lightweight pre-retrieval and enriched post-retrieval modes, significantly outperforming state-of-the-art methods on multiple benchmark datasets. The results demonstrate the practical utility of embedding historical query knowledge into QPP tasks, achieving a balance between computational efficiency and prediction accuracy.

Building on this work and as future work, we will explore dynamic query subspace construction techniques that adaptively select and weigh historical queries based on their contextual relevance and domain-specific characteristics, enabling more accurate predictions in heterogeneous datasets. Additionally, we aim to develop cross-domain QPP transfer learning methods, leveraging shared embedding spaces and pre-trained models to adapt QSD-QPP for domains with limited labeled query data, ensuring broader applicability and robustness across diverse information retrieval tasks especially for low-resource domains.

References

[1] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *EMNLP*.

[2] 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.

[3] 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*. 2857–2861.

[4] Negar Arabzadeh, Bhaskar Mitra, and Ebrahim Bagheri. 2021. MS MARCO Chameleons: Challenging the MS MARCO Leaderboard with Extremely Obstinate Queries. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4426–4435.

[5] Negar Arabzadeh, Mahsa Seifikar, 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*. 3811–3816.

[6] 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.

[7] Negar Arabzadeh, Fattaneh Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2019. Geometric estimation of specificity within embedding spaces. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2109–2112.

[8] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2020. Neural embedding-based metrics for pre-retrieval query performance prediction. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 78–85.

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

[10] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. 2009. Similarity-based Classification: Concepts and Algorithms. *J. Mach. Learn. Res.* 10 (2009), 747–776. https://api.semanticscholar.org/CorpusID:8559164

[11] Scott Cost and Steven L. Salzberg. 2004. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning* 10 (2004), 57–78. https://api.semanticscholar.org/CorpusID:35426433

[12] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR* abs/2102.07662 (2021). arXiv:2102.07662 https://arxiv.org/abs/2102.07662

[13] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2021. Ms marco: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1566–1576.

[14] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[15] 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.

[16] 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.

[17] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. 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*. 201–209.

[18] Suchana Datta, Debasis Ganguly, Mandar Mitra, and Derek Greene. 2022. A Relative Information Gain-based Query Performance Prediction Framework with Generated Query Variants. *ACM Transactions on Information Systems* 41, 2 (2022), 1–31.

[19] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. 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*. 2148–2153.

[20] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document' based QPP Approach. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3477495.3531821

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[22] 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. https://doi.org/10.1007/978-3-031-56066-8_6

[23] Guglielmo Faggioli, Nicola Ferro, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto. 2023. A Geometric Framework for Query Performance Prediction in Conversational Search. (2023).

[24] Guglielmo Faggioli, Thibault Formal, Stefano Marchesin, Stéphane Clinchant, Nicola Ferro, and Benjamin Piwowarski. 2023. Query Performance Prediction for Neural IR: Are We There Yet?. In *European Conference on Information Retrieval*. Springer, 232–248.

[25] Sally A. Goldman and Yan Zhou. 2000. Enhancing Supervised Learning with Unlabeled Data. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:1215747

[26] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A Deep Look into neural ranking models for information retrieval. *Inf. Process. Manag.* 57, 6 (2020), 102067. https://doi.org/10.1016/J.IPM.2019.102067

[27] 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.

[28] Claudia Hauff. 2010. Predicting the effectiveness of queries and retrieval systems. In *SIGIR Forum*, Vol. 44. 88.

[29] 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. https://doi.org/10.1007/978-3-540-30213-1_5

[30] Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems* 31, 7 (2006), 585–594.

[31] Jiyin He, Martha Larson, and Maarten de Rijke. 2008. Using Coherence-Based Measures to Predict Query Difficulty. In *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*. 689–694. https://doi.org/10.1007/978-3-540-78646-7_80

[32] Qiang He, Zongxia Xie, Qinghua Hu, and Congxin Wu. 2011. Neighborhood based sample and feature selection for SVM classification learning. *Neurocomputing* 74 (2011), 1585–1594. https://api.semanticscholar.org/CorpusID:1517659

[33] Herve Jegou, Matthijs Douze, and Cordelia Schmid. 2010. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence* 33, 1 (2010), 117–128.

[34] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[35] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172* (2019).

[36] Maryam Khodabakhsh and Ebrahim Bagheri. 2021. Semantics-enabled query performance prediction for ad hoc table retrieval. *Inf. Process. Manag.* 58, 1 (2021), 102399. https://doi.org/10.1016/J.IPM.2020.102399

[37] Maryam Khodabakhsh and Ebrahim Bagheri. 2023. Learning to rank and predict: Multi-task learning for ad hoc retrieval and query performance prediction. *Information Sciences* 639 (2023), 119015.

[38] Maryam Khodabakhsh, Fattane Zarrinkalam, and Negar Arabzadeh. 2024. BertPE: A BERT-Based Pre-retrieval Estimator for Query Performance Prediction. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 14610)*. Springer, 354–363. https://doi.org/10.1007/978-3-031-56063-7_27

[39] Georgios Kostopoulos, Stamatis Karlos, Sotiris Kotsiantis, and Omiros Ragos. 2018. Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems* 35, 2 (2018), 1483–1500.

[40] K. 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, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*. 187–195. https://doi.org/10.1145/243199.243266

[41] Liyan Liu, Jin Zhang, Kun Qian, and Fan Min. 2024. Semi-supervised regression via embedding space mapping and pseudo-label smearing. *Applied Intelligence* 54, 20 (2024), 9622–9640.

[42] Iain Mackie, Jeffrey Dalton, and Andrew Yates. 2021. How deep is your learning: The DL-HARD annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2335–2341.

[43] 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).

[44] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[45] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

[46] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[47] Yannis Panagakis and Constantine Kotropoulos. 2014. Elastic Net subspace clustering applied to pop/rock music structure analysis. *Pattern Recognition Letters* 38 (2014), 46–53. https://doi.org/10.1016/j.patrec.2013.10.021

[48] Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *European conference on information retrieval*. Springer, 397–412.

[49] Vassilis Plachouras, Ben He, and Iadh Ounis. 2004. University of Glasgow at TREC 2004: Experiments in Web, Robust, and Terabyte Tracks with Terrier.. In *TREC*.

[50] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. arXiv:2101.05667 [cs.IR]

[51] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667* (2021).

[52] Bradley S Price and Ben Sherwood. 2018. A cluster elastic net for multivariate regression. *Journal of Machine Learning Research* 18, 232 (2018), 1–39.

[53] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[54] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[55] 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.

[56] 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, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017*, Jaap Kamps, Evangelos Kanoulas, Maarten de Rijke, Hui Fang, and Emine Yilmaz (Eds.). ACM, 245–248. https://doi.org/10.1145/3121050.3121087

[57] Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth JF Jones. 2019. Estimating gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information processing & management* 56, 3 (2019), 1026–1045.

[58] Surendra Sarnikar, Zhu Zhang, and J Leon Zhao. 2014. Query-performance prediction for effective query routing in domain-specific repositories. *Journal of the Association for Information Science and Technology* 65, 8 (2014), 1597–1614.

[59] 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.

[60] 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.

[61] Ashutosh Singh, Debasis Ganguly, Suchana Datta, and Craig Macdonald. 2023. Unsupervised Query Performance Prediction for Neural Models utilising Pairwise Rank Preferences. *def* 1 (2023), 2.

[62] Saúl Solorio-Fernández, J Ariel Carrasco-Ochoa, and José Fco Martínez-Trinidad. 2020. A review of unsupervised feature selection methods. *Artificial Intelligence Review* 53, 2 (2020), 907–948.

[63] 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.

[64] Eric P. Xing, A. Ng, Michael I. Jordan, and Stuart J. Russell. 2002. Distance Metric Learning with Application to Clustering with Side-Information. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:2643381

[65] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=zeFrfgyZln

[66] Hamed Zamani and Michael Bendersky. 2023. Multivariate representation learning for information retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 163–173.

[67] Hamed Zamani, W Bruce Croft, and J Shane Culpepper. 2018. Neural query performance prediction using weak supervision from multiple signals. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 105–114.

[68] 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. https://doi.org/10.1007/978-3-540-78646-7_8

[69] 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.

[70] Xiaojin Zhu and Andrew B. Goldberg. 2009. Introduction to Semi-Supervised Learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*. https://api.semanticscholar.org/CorpusID:40097546