Robust Query Performance Prediction with Context-Aware Models

Abbas Saleminezhad

Supervisors: Dr. Ebrahim Bagheri
Dr. Soosan Beheshti

PhD Candidate at Toronto Metropolitan University, April 2025

Saleminezhad, A., Arabzadeh, N., Rad, R.H. *et al.* Robust query performance prediction for dense retrievers via adaptive disturbance generation. *Mach Learn* 114, 65 (2025).

Research Motivation

What is QPP?

- predict the retrieval quality of a search system for a query without human relevance judgments

Why Query Performance Prediction (QPP) Matters?

- Information retrieval systems (search engines, QA systems, RAG models) often struggle with poorly performing queries.
- Queries vary in effectiveness; some retrieve highly relevant documents, others fail.

Impact of QPP:

- Helps improve retrieval effectiveness by identifying difficult queries.
- Enables query reformulation, retrieval model adaptation, and better ranking strategies.

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Problem definition

- Estimating how well the retrieved documents meet the informational needs expressed by the query.
- Predictor μ has to estimate the performance of q
  - A Collection C
  - A Query q
  - A list of retrieved documents Dq

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

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Research Problem & Objective

Research Problem

- How can we predict the performance of queries in both sparse and dense retrieval settings?
- Need for robust QPP techniques that generalize across retrieval models.

Research Objectives

- Overcoming Perturbation-Based QPP limitations:
  Relies on lexical query modifications, making it ineffective for dense retrievers.
  Sensitive to dataset-specific perturbations, leading to inconsistent performance.
- Utilizing contextualized embeddings for a consistent performance

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

ADG-QPP

Foundations of ADG-QPP

Robust vs. Non-Robust Queries:

- **Robust Queries:** Retrieval remains **consistent** despite perturbations.
- **Non-Robust Queries:** Small perturbations **significantly change** retrieval results.

Challenge in Dense Retrieval:

- Sparse retrievers handle **lexical changes** (e.g., word swaps).
- **Dense retrievers require embedding-level perturbations.**

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

The Proposed Main Approach

Query Embedding & Dense Retrieval

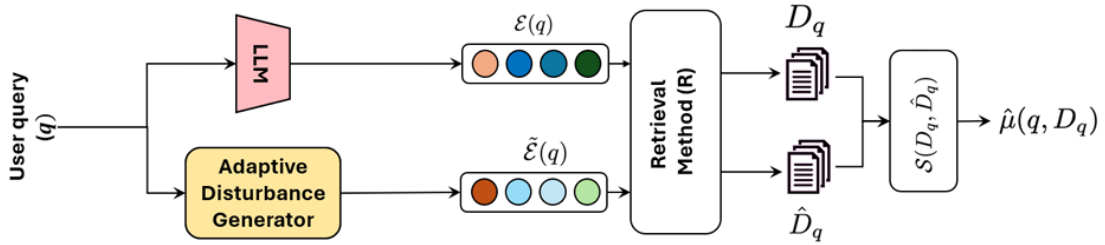
- Query q is mapped to a dense vector using LLM
- Retrieves top-k documents from corpus C with dense retriever R

Perturbation of Query Embeddings

- Perturbations are added to the query embedding to create a disturbed representation

Similarity

- Quantifies the differences between the retrieved lists from original and perturbed query



TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Proposed Approach

Query Representation in Embedding Space:

- A function **maps queries into dense vectors** for retrieval.
- Dense retriever R retrieves **top-k documents** based on these vectors.

Performance Estimation:

- Compare retrieved results **before and after perturbation**.
- A stable retrieval set = **robust query**, unstable set = **difficult query**.

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Adaptive Disturbance Generation (ADG)

Baseline Perturbation: Gaussian Noise (AWGN)

- Uniformly applies **noise** to query embeddings:
- **Issue:** Assumes **all queries are equally sensitive** to noise.

Adaptive Disturbance Generation (ADG)

- Adjusts noise **based on query context** in embedding space.
- Ensures **more meaningful** perturbations per query.

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Focal Networks for Adaptive Noise

Focal Network Constructs Query Context:

- **Query-based Focal Network (QFN)** – Captures similarity with **other queries**.
- **Document-based Focal Network (DFN)** – Captures similarity with **retrieved documents**.

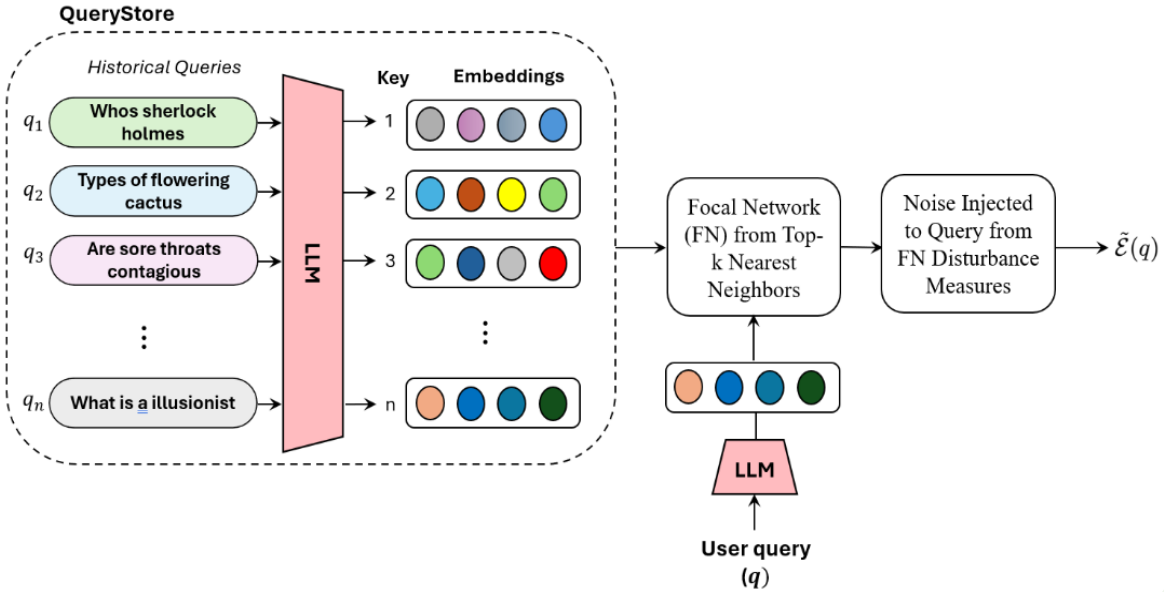
Graph-Based Noise Personalization:

- Use **network metrics** to **adjust disturbance level**

TMU

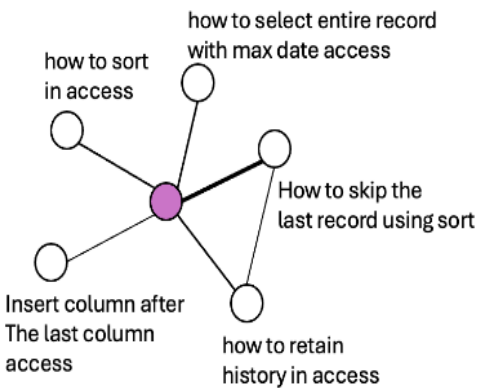
47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Query-based Focal Network (QFN)

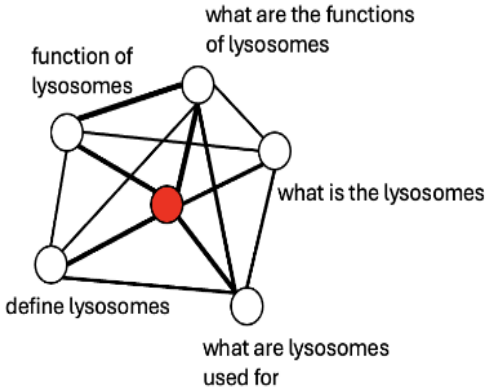


Focal Networks - QFN

- Main query and k most similar queries in Query Store
- QFN insights for query robustness
  - Sparse QFNs → poor query performance, high sensitivity to disturbances
  - Dense QFNs → robust queries, less affected by noisy perturbations



Query: 'access, how to go to most recent record'



Query: 'what important job do the lysosomes have'

Network Metrics

Node-based Disturbances

- Insight to structural importance and connectivity of the query node

Edge-based Disturbances

- Assess the structure and strength of connections

Cluster-based Disturbances

- Understanding the overall interconnectedness of network

	Metric Name	Formula	Description
Node-based	Semantic Network Size (SNS)	$\lvert V_\zeta \rvert$	A large network size surrounding a particular query node suggests that the query is embedded in a dense semantic space, potentially enhancing its robustness due to multiple relational pathways.
	Degree Centrality (DC)	$\lvert \{e_{i,j} \in E_\zeta\} \rvert$	Extent to which a query connects to other nodes, identifying whether the node is popular in the focal network. Popular queries are likely to be robust to disturbance.
	Closeness Centrality (CC)	$\left[\sum_{j \in V_\zeta, j \neq i} d(q,j)\right]^{-1}$	Highlights how quickly it is possible to move from the query node to others in the focal network. Queries with high connectivity maintain short paths even when perturbed and remain robust.
	PageRank (PR)	$\frac{1-d}{\lvert V_\zeta \rvert} + d \sum_{j \in V_\zeta} \frac{PR(j)}{\deg(j)}$	A high PR measures how well a query is not only connected to many other nodes but also connected to other highly connected ones. A high PageRank may be a sign of being robustness to disturbance.
Edge-based	Connectivity Score (CS)	$\lvert E_\zeta(q) \rvert$	A high edge count indicates that the query is well-connected, suggesting that it will be less sensitive to noisy perturbations.
	Query Connectivity Strength (QCS)	$\sum_{e_{q,j} \in E_\zeta} \gamma(e_{q,j})$	A high connectivity strength value indicates numerous relevant connections within the network, which can be a sign of the robustness of query.
	Average Query Connectivity (AQC)	$\frac{1}{\lvert E_\zeta(q) \rvert} \sum_{e_{q,j} \in E_\zeta} \gamma(e_{q,j})$	This metric computes the average strength of connections for the query node where high average strengths point to more resilient queries against disturbance.
	Rare Path Index (RPI)	$\frac{1}{\lvert E_\zeta(q) \rvert} \sum_{e_{q,j} \in E_\zeta} \frac{1}{\gamma(e_{q,j})}$	A High RPI suggests that specific connections remain stable and relevant when query is altered.
Cluster-based	Inter-Cluster Connectivity (ICC)	$\frac{1}{\binom{k}{2}} \sum_{I \neq j} \max_{i,j} \gamma(e_{i,j})$	The strength of connections between different clusters reflects the overall interconnectedness of the network pointing to resilience against disturbance.
	Centroid Cluster Weight (CCW)	$\frac{\sum_{(u,v) \in E_{C_k}} w(u,v)}{\lvert E_{C_k} \rvert}$	measures strength within the most cohesive cluster, a sign of robustness against disturbances in at least one aspect of the query.

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Query Performance Estimation via Ranked Bias Overlap (RBO)

Retrieval Stability as a Performance Indicator:

- Compare **retrieved document lists** before & after perturbation.
- Use **Ranked Bias Overlap (RBO)** to measure similarity.

Interpreting Results:

- **High RBO** → Query is **robust** (retrieved results are stable).
- **Low RBO** → Query is **difficult** (results change significantly).

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Evaluation

Dataset

MS MARCO Passage Collection V1 containing 8.8 Million passages

500k train query set with known performances

Query test sets

TREC Deep Learning Track 2019

TREC Deep Learning Track 2020

TREC DL Hard

Performance Metric

MRR@10 and ndcg@10

Correlation Metrics

Kendall's τ

Pearson's ρ

Spearman's ρ

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

ADG-QPP Findings

Table 3: Performance comparison between our best-performed proposed approach and SOTA baselines when predicting the performance of S-BERT dense retriever. All correlations are statistically significant at $\alpha = 0.5$ except the *italic* ones. The highest value in each column is in bold.

	DL-Hard			DL-2019			DL-2020		
	$P-\rho$	$K-\tau$	$S-\rho$	$P-\rho$	$K-\tau$	$S-\rho$	$P-\rho$	$K-\tau$	$S-\rho$
Clarity	0.232	0.110	0.162	0.217	0.111	0.151	0.196	0.137	0.188
QF	0.044	0.051	0.060	0.071	0.022	0.043	0.148	0.029	0.052
NQC	0.418	0.276	0.381	0.560	0.419	0.598	0.285	0.194	0.289
WIG	0.093	0.072	0.105	0.139	0.071	0.116	0.153	0.032	0.051
$n(\sigma_x)$	0.400	0.259	0.369	0.501	0.361	0.532	0.242	0.158	0.232
SMV	0.396	0.314	0.438	0.577	0.428	0.600	0.360	0.246	0.357
UEF	0.441	0.298	0.412	0.607	0.428	**0.601**	0.336	0.228	0.329
NeuralQPP	0.232	0.080	0.103	0.209	0.057	0.057	0.152	0.015	0.003
Pclarity_NQC	0.088	0.053	0.083	0.428	0.314	0.451	0.084	0.202	0.292
NQAQPP	0.113	0.240	0.359	0.269	0.129	0.160	0.221	0.159	0.234
BERTQPP	0.435	0.181	0.256	0.334	0.143	0.194	0.378	0.273	0.411
qppBERT-PL	0.405	0.171	0.225	0.299	0.131	0.183	0.344	0.224	0.335
Deep-QPP	0.096	*0.049*	*0.065*	0.139	0.103	0.106	0.262	0.197	0.291
QPP-PRP	0.181	0.099	0.144	0.203	0.204	0.281	0.181	0.143	0.219
AWGN (Dense-QPP)	0.371	0.254	0.384	0.572	0.414	0.574	0.331	0.199	0.318
Our Approach	**0.469**	**0.319**	**0.449**	**0.684**	**0.439**	0.598	**0.401**	**0.298**	**0.424**

TMU

47TH EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL ECIR 2025

Thank you!