VAP3: Variation-Aware Prompt Performance Prediction

Anonymous Author(s)

Abstract

Large Language Models (LLMs) exhibit strong capabilities across various Information Retrieval (IR) and natural language processing tasks. However, they are highly sensitive to prompt variations, where slight rephrasings can significantly alter responses, leading to inconsistent or incorrect outputs. This variability poses challenges for response reliability in real-world applications. Inspired by Query Performance Prediction (QPP) in IR, we focus on Prompt Performance Prediction (PPP), which estimates whether an LLM will generate a correct response for a given prompt before execution. We propose *VAP3 (Variation-Aware Prompt Performance Prediction)*, a novel pre-generation PPP approach that integrates prompt variations with adversarial training to enhance robustness against trivial modifications and better capture prompt sensitivity. Evaluating VAP3 against LLM-based self-evaluation, QPP-inspired baselines, and supervised classification models on the PromptSET-HotpotQA and PromptSET-TriviaQA datasets, we demonstrate that VAP3 consistently outperforms all baselines, achieving stable and reliable performance across datasets.

1 Introduction

Large Language Models (LLMs) have recently demonstrated strong performance in question answering, information retrieval (IR), and broader natural language processing tasks [12, 15, 32, 38, 50, 60]. Their ability to generate fluent and contextually relevant responses has led to their widespread adoption across various applications [8, 26, 29, 56]. However, a significant challenge remains: *prompt sensitivity*, the phenomenon where negligible changes in the wording or structure of a prompt can lead to dramatically different responses [14, 39, 61]. There have been many attempts to study the impact of prompting on the output of different tasks [37, 55]. For example, in the IR community, at the LLME4val workshop at SIGIR 2024 [44] and for the task of LLM-based relevance judgment, different research groups were asked to formulate a prompt for a well-understood objective of providing relevance judgments. Despite all participants being familiar with the task, sharing a background in computer science, and having experience with LLMs, the prompts they generated for the same language model resulted in significantly different performance outcomes and varying levels of agreement with human annotations [45]. This inconsistency raises concerns about the reliability and robustness of LLMs, particularly in domains that demand high precision and accuracy [53].

The IR community has already extensively researched the distinct yet conceptually similar task of *Query Performance Prediction* (QPP), which aims to estimate the effectiveness of a query without having access to ground truth [3, 18, 22, 28]. Inspired by QPP, we focus on *Prompt Performance Prediction* (PPP) as the task of predicting whether an LLM will effectively generate a correct response to a given prompt. Here, we assume that the effectiveness of a prompt is quantifiable and that a correct response exists. Similar to QPP, which is categorized into pre-retrieval and post-retrieval strategies [11], PPP can also be divided into *pre-generation PPP*, and *post-generation PPP*. In QPP, post-retrieval methods benefit from additional contextual information, i.e., retrieved results, allowing for more accurate predictions [17, 34, 41]. We expect a similar trend in PPP, where post-generation approaches would provide more reliable predictions compared to pre-generation ones. However, pre-generation PPP is particularly valuable and it has more applications in real-world scenarios because it enables performance estimation before generation occurs. This can significantly reduce computational cost and save system generation time. Additionally, it allows the system to make informed decisions before presenting an answer to the user. For example, it can guide users or the system to refine prompts [42, 59], ask clarification questions [4, 48], or even select a different LLM to generate the response before execution [52].

To our knowledge, there has been limited research on this topic. In [10, 43], the authors explored prompt performance prediction for image generation. Some studies have explored LLM-based self-evaluation strategies, where the model assesses its own ability to respond to a given prompt [2, 25, 31], but extensive research in this area is still lacking. Additionally, efforts have been made to quantify and understand prompt sensitivity [13, 62]. For instance, Chatterjee et al. [13] introduced a metric to measure prompt sensitivity, analyzing how slight variations in prompts affect LLM responses. Their findings indicate that certain LLMs exhibit inherent sensitivity, leading to inconsistent or incorrect outputs even with minor modifications. Overall, predicting whether a prompt will effectively satisfy a user's information need remains a challenging task.

A recent study introduced PromptSET (Prompt Sensitivity Evaluation Task) [47], a specifically designed to examine prompt sensitivity. Constructed from HotpotQA and TriviaQA benchmarks [33, 57], PromptSET comprises a diverse collection of short-form question-answer pairs. Each question is accompanied by slight variations that preserve the same underlying information need. The dataset also includes labels indicating whether an LLM correctly answered each prompt variation. The study aimed to identify prompts that are more sensitive to variations and determine which reformulations of the same information need can be consistently answered by the same LLM. We adopt PromptSET for the PPP task. In addition to providing a large-scale question-answering dataset with LLM-generated responses, PromptSET offers structured prompt variations. These variations are particularly valuable for PPP, enabling us to analyze factors influencing prompt effectiveness and develop robust prediction models.

In addition, Razavi et al. [47] have attempted to apply QPP-inspired baselines on the PromptSET dataset. Their findings indicate that unsupervised QPP methods perform poorly on predicting prompt sensitivity, likely due to the complexity of language generation compared to retrieval. However, supervised approaches demonstrated better performance, suggesting that learned representations are crucial for effective prediction. Building on these insights, we propose a PPP method that integrates *prompt variations* with *adversarial training* to effectively predict prompt performance.

Our PPP method is inspired by findings in [47], which show that when an original prompt is answered correctly, its variations are also likely to receive correct responses. Similarly, LLMs generate more accurate responses for prompts with high similarity to their variations [20, 47], indicating a direct correlation between prompt-variation similarity and response correctness. Incorporating these variations enhances performance estimation and better captures prompt sensitivity. A similar principle in QPP uses query perturbations to estimate retrieval performance [1, 51], where comparing retrieved results from original and perturbed queries provides a strong retrieval effectiveness signal [59]. These insights motivate our *multi-prompt learning framework*, leveraging prompt variations to improve prediction accuracy.

We propose VAP3 (Variation-Aware Prompt Performance Prediction), a novel approach to pre-generation PPP that moves beyond treating prompts in isolation. By incorporating prompt variations, VAP3 more effectively captures sensitivity patterns, leading to more accurate performance predictions. Additionally, our transformer-based model leverages adversarial training to enhance robustness against trivial modifications, ensuring that only slight prompt differences are taken into consideration when predicting prompt performance. We systematically evaluate VAP3 against both LLM-based and QPP-inspired baselines using PromptSET, a dataset derived from HotpotQA and TriviaQA. Experimental results show that VAP3 significantly outperforms existing methods, providing a more reliable approach for predicting prompt effectiveness. To ensure reproducibility, we have released our code at https://anonymous.4open.science/r/vap3-3042.

2 Proposed Approach

Problem Definition. Let p be a prompt and let $\mathcal{G}$ be a language model that generates a response $\mathcal{G}(p)$ for p. We assume that each prompt p has an associated set of ground truth responses R_p. The generated response $\mathcal{G}(p)$ is considered correct if: $\mathcal{G}(p) \in R_p$. We define the prompt performance predictor $\mu(p)$ as a function that quantifies the likelihood of $\mathcal{G}(p) \in R_p$, i.e., $\mu(p, \mathcal{G}(p)) \rightarrow [0, 1]$

Transformer-Based Prediction Model. In the context of QPP in IR [11], several efforts have been made to develop retrieval performance predictors using neural models [27, 49, 58]. Building on these approaches, Khodabakhsh et al. [35] introduced a predictor namely BERTPE, denoted as μ, defined as:

$$\mu(p) = \sigma\big(W_2 \cdot \phi(W_1 \overrightarrow{h}(p)_{[\text{CLS}]} + b_1) + b_2\big) \tag{1}$$

where: $\overrightarrow{h}(p)_{[\text{CLS}]}$ is the final hidden representation of the special $[CLS]$ token of the prompt p from a transformer encoder, e.g., BERT [19], and W_1, W_2 are trainable weight matrices, and b_1, b_2 are bias terms, $\phi(\cdot)$ is a non-linear activation function, such as ReLU, and finally, $\sigma(\cdot)$ is the sigmoid function, ensuring the output is in the range of $[0, 1]$. The prediction model is trained using a binary cross-entropy loss:

$$\mathcal{L}_{ppp}(\theta) = -\sum_{p \in \mathcal{T}} y_p \log \mu(p) + (1 - y_p) \log(1 - \mu(p)) \tag{2}$$

Prompt Variations. To account for prompt sensitivity, we utilize variations of each prompt to assess performance consistency. In the context of the PromptSET dataset, $v(p)$ is defined as a set of variations for a given prompt p, ensuring that each variation $v_{p_i} \in v(p)$ maintains the same intent while differing in phrasing or structure. For each prompt p, an LLM is used to generate a set of $|v(p)|$ prompt variations (in order of 10). To ensure meaningful transformations, each variation v_{p_i} is required to preserve the original information need while maintaining a semantic similarity above a threshold τ, i.e., $\text{Sim}(p, v_{p_i}) > \tau$. This step ensures that generated variations retain the intent of the original prompt without introducing ambiguity. Variations that fail to meet the similarity threshold τ or deviate from the original intent are discarded.

Using variations of prompts introduced in PromptSET, we can consider $\mathcal{T}$ to represent the training dataset, which consists of prompts $p \in P$, their variations $v(p)$, and the corresponding ground truth labels y_p i.e.,:

$$\mathcal{T} = \{[p_i, y_{p_i}] \mid p_i \in \{p \cup v(p) \forall p \in P\}\} \tag{3}$$

Each training sample $t_i \in \mathcal{T}$ is accompanied by a binary label y_{p_i} that indicates whether the generated response is correct. y_{p_i} is only 1 if $\mathcal{G}(p) \in R_p$. This formulation ensures that the model learns to predict whether the response $G(p)$ produced by the language model $\mathcal{G}$ aligns with the ground truth response set R_p.

In this work, we extend the input of the predictor μ beyond a standalone prompt p by incorporating its variations $v(p)$. As discussed in the introduction, leveraging prompt variations and modeling their interactions with the original prompt can enhance performance prediction. To this end, we propose incorporating prompt variations directly into the prediction model. Instead of making independent predictions for each prompt, we estimate the performance by considering both p and its variations, formulating $\mu(p)$ as the expected value over all pairs of p and its variations:

$$\mu(p) = \mathbb{E}_{v_{p_i} \sim v(p)} \left[\sigma(W_2 \cdot \phi(W_1 \overrightarrow{h}(p \oplus v_{p_i})_{[\text{CLS}]} + b_1) + b_2)\right] \tag{4}$$

Here, $\overrightarrow{h}(p \oplus v_{p_i})_{[\text{CLS}]}$ is the final hidden representation of the [CLS] token obtained from encoding the concatenation of the prompt p and its variation v_p, separated by a special token [SEP]. $\mathbb{E}_{v_p \sim \mathcal{V}(p)}$ denotes the expected value taken over all variations $v_{p_i} \in v(p)$. The network applies a non-linear transformation $\phi(\cdot)$ followed by a sigmoid activation $\sigma(\cdot)$ to produce the probability estimate. This approach ensures that instead of making independent predictions for each prompt variation, we model the joint impact of variations on prompt performance.

Additionally, we want to make our model aware of trivial modifications, such as typos or minor word substitutions, that do not significantly impact LLM responses [9]. A naïve prediction model may overfit to these minor variations, misinterpreting them as meaningful differences and introducing unnecessary sensitivity. To mitigate this issue, we incorporate an *adversarial training module* that systematically refines the training process by reducing the model's sensitivity to irrelevant changes [30, 40]. In addition, this ensures that the model can generalize beyond seen prompt variations. To achieve this, we employ the *Fast Gradient Sign Method (FGSM)* [24] to perturb the token embeddings and generate adversarial prompts. Given an input prompt p, we define its adversarial version $\tilde{p}$ as:

$$\tilde{p} = p + \epsilon \cdot \text{sign}(\nabla_p \mathcal{L}(\theta)) \tag{5}$$

Table 1: Examples of Original prompts and their variations from PromptSET-HotpotQA and PromptSET-TriviaQA dataset. The variations were generated by LLaMA 3.1. On average each prompt is accompanied by 9 different variations.

Original Prompt	Variation 1	Variation 2
What party is Sarah Coburn's father a member of?	Identify the political affiliation of Sarah Coburn's dad.	Determine the parental politics of Sarah Coburn.
1998 was the Chinese year of which creature?	What is the zodiac animal for the Chinese year that started in 1998?	Which animal represents the Chinese New Year that began in 1998 according to the traditional calendar?
Which Monkee was born in Maryland but grew up in Washington DC?	Who is the Monkee with a connection to the nation's capital?	Can you name the Monkee with roots in Washington DC?

where ϵ is the perturbation factor that controls the intensity of adversarial modifications. $\nabla_p \mathcal{L}(\theta)$ is the gradient of the loss function with respect to the input embeddings and $\text{sign}(\cdot)$ extracts the direction of gradient updates to perturb the most sensitive tokens.

We train the model using both original and adversarial variations. Given a batch of training data $\mathcal{T}$, we extend the dataset to include adversarial perturbations:

$$\mathcal{T}_{\text{adv}} = \mathcal{T} \cup \{[\tilde{p}, y_p] \mid p \in \mathcal{T}\} \tag{6}$$

The adversarial loss is computed similar to the PPP loss in Equation 2 but on perturbed versions, i.e.,:

$$\mathcal{L}_{adv}(\theta) = -\sum_{\tilde{p} \in \mathcal{T}_{adv}} y_p \log \mu(\tilde{p}) + (1 - y_p) \log(1 - \mu(\tilde{p})) \tag{7}$$

The final training objective combines both the standard and adversarial losses where λ is a hyperparameter controlling the contribution of adversarial training:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{ppp} + \lambda \mathcal{L}_{\text{adv}} \tag{8}$$

3 Experiments

3.1 Dataset

We evaluate our proposed approach using PromptSET generated by LLaMA [47]. PromptSET is derived from two widely used question-answering benchmarks: HotpotQA[57] and TriviaQA [33]. It consists of original prompts and their systematically generated variations, ensuring that all variations retain the same underlying information need while differing in surface form. Each prompt and its variations are labeled based on whether an LLM successfully generates the expected answer. This dataset provides a robust large-scale evaluation setting for prompt performance prediction. Table 1 shows some prompts and their variations in PromptSET dataset. These variations are generated by LLaMA 3.1 after being prompted to generated alternative variations for the original prompt. PromptSET comprises 8,028 original prompts in the training set and 3,441 original prompts in the test set, each accompanied by nine automatically generated variations by LLaMA. The questions and their variations were then answered by the LLM and the responses were compared against human-annotated answers from HotpotQA and TriviaQA. A variation is labeled correct if the LLM-generated response matches the expected answer. This dataset allows for the evaluation of pre-generation PPP, enabling models to predict performance before response generation. For more details on this dataset, we refer to the original paper.

3.2 Baselines

Since there is no existing method specifically designed for PPP, we explore a set of potential QPP and LLM-inspired baseline. We note that corpus-based qpp metrics are inapplicable in our scenario since no document collection is available. Instead, we focus on query-only metrics, which rely on prompt terms (and their variations).

1. *LLM-Based Self-Evaluation:* In this baseline, the LLM is explicitly asked to predict its ability to generate a correct response to the given prompt. This self-evaluation strategy has been employed previously [2, 25, 31] and is known for exhibiting overconfidence, as LLMs tend to overestimate their own reliability. This is implemented using LLaMA 3.1, where the model outputs a confidence score reflecting its perceived capability to produce an accurate response.
2. *Text-Based Classification:* This baseline introduced in PromptSET, frames PPP as a binary text classification task, where a classifier is trained to predict whether a given prompt will be successfully. The baseline fine-tunes a BERT-based classifier from simple transformers [46] that learns from different prompts and their corresponding answerability labels.
3. *Specificity-Based QPP Methods:* Inspired by specificity metrics [5–7], we explore prompt specificity as a proxy for performance. The intuition is that specific prompts, being less ambiguous, are easier to address, whereas broader prompts introduce interpretative uncertainty. We implement four neural-based specificity QPP methods—Closeness Centrality (CC), Degree Centrality (DC), Inverse Edge Frequency (IEF), and PageRank—quantifying the centrality of prompt terms in the embedding space, with the hypothesis that higher specificity correlates with greater answerability.
4. *Clarity Score:* In QPP, the query clarity hypothesis posits that queries with lower Kullback-Leibler (KL) divergence from a document corpus retrieve more relevant results [16]. Inspired by this, we introduce the *Prompt Clarity Score*, which quantifies the KL divergence between a prompt and its variations. A low divergence indicates semantic consistency across rephrasings, suggesting a well-defined concept that is easier for an LLM to address. We evaluate KL divergence as a predictor of prompt performance. 5. *Entropy:* Entropy has been used in prior work as a signal for detecting hallucination in LLM-generated responses [23, 36, 54], where higher entropy indicates greater uncertainty in the model's predictions. The underlying intuition is that a prompt with higher entropy suggests lower predictability and answerability, meaning that the LLM is more uncertain about producing a relevant and precise response. Following this intuition, we hypothesize that by measuring entropy in token representations, we capture query variability and uncertainty. Significant entropy shifts across prompt variations indicate instability, suggesting lower reliability and response quality.
6. *Geometric-Based Methods:* Faggioli et al. [21] have introduced a geometric framework for QPP, measuring query-document embedding volumes to assess retrievability. We adapt it for PPP by constructing embedding volumes from prompts and their variations. We explore two methods: (i) Reciprocal Volume, which measures the inverse hypercube volume around a query and its variants,

Table 2: Results of baselines and VAP3 on TriviaQA and HotPotQA.

Category	Method	TriviaQA				HotPotQA			
		Accuracy	F1	Recall	Precision	Accuracy	F1	Recall	Precision
Random	Random	0.5017	0.507	0.580	0.510	0.498	0.308	0.485	0.226
LLM-based	Self-evaluation	0.4656	0.6239	0.9798	0.4577	0.1696	0.2050	0.9419	0.1150
Specificity-based	CC	0.506	0.453	0.452	0.454	0.549	0.209	0.524	0.130
	DC	0.484	0.448	0.463	0.434	0.565	0.199	0.475	0.126
	IEF	0.505	0.462	0.469	0.455	0.535	0.204	0.526	0.127
	PageRank	0.481	0.444	0.458	0.431	0.533	0.153	0.370	0.096
Geometric-based	Reciprocal Volume	0.5629	0.532	0.594	0.512	0.533	0.320	0.477	0.241
	Discounted Matryoshka	0.5526	0.529	0.592	0.511	0.532	0.326	0.491	0.244
Query-variations	Query Clarity	0.5503	0.538	0.606	0.532	0.520	0.350	0.561	0.254
Interactions	Entropy	0.4554	0.502	0.606	0.559	0.424	0.300	0.536	0.209
Text Classification	BERT	0.660	0.659	0.620	0.654	0.526	0.360	0.017	0.813
Supervised	BERTPE	0.648	0.627	0.644	0.611	0.710	0.318	0.594	0.217
Proposed	VAP3	0.693	0.781	0.819	0.746	0.730	0.501	0.589	0.436

with smaller volumes indicating higher semantic coherence; and (ii) Discounted Matryoshka Representation, which progressively aggregates volumes to capture retrieval consistency. These methods hypothesize that compact prompt embeddings indicate higher answerability, while greater dispersion suggests increased uncertainty and low answerability.

7. *Supervised QPP Baseline (BERTPE)*: Finally, we include BERTPE, a state-of-the-art supervised QPP model that has shown strong performance in prior studies [35]. BERTPE learns to predict retriever performance by fine-tuning contextualized representations of queries. In our setting, BERTPE is trained on individual prompts (both original and variations) to learn their intrinsic difficulty. This serves as a strong supervised baseline for PPP.

3.3 Experimental Setup

For training, we tuned key hyperparameters, including the number of epochs $\in [1, 2, 3, 4, 5]$, $\epsilon \in [1e{-}5, 5e{-}5, 1e{-}4, 5e{-}4, 1e{-}3, 5e{-}3]$ (from Equation 5), and $\lambda \in [0.1, 0.3, 0.5, 0.7, 0.9]$ (from Equation 8), the learning rate $[1e-6, 1e-5, 1e-4]$, batch size $[8, 16, 32]$, and dropout rate $[0.1, 0.2, 0.3]$ for fine-tuning bert-base-uncased with adversarial training. We conducted a hyperparameter sweeping using 10% of the training set as a validation set, optimizing for the lowest validation loss. The best-performing configuration, selected based on validation performance, was then applied to the full test set of each dataset.

4 Results

In Table 2, we present the performance of all baseline models introduced in the previous section on the PromptSET-TriviaQA and PromptSET-HotpotQA [47]. The evaluation metrics include accuracy, F1-score, recall, and precision. To provide a reference point, we include a random baseline in the first row, which assigns labels randomly. From this table, we make the following observations:
(1) Among the baselines, the unsupervised methods exhibit notably poor performance, with accuracy values hovering around 50%, which is close to random chance. Geometric-based methods, perform slightly better than specificity-based methods but still fail to provide meaningful predictive power. Furthermore, query-variant interaction-based approaches, including Clarity Score and Entropy, also fail to yield effective predictions for the PPP task. While these features have demonstrated strong correlations with QPP in the retrieval setting, their predictive effectiveness does not translate well to prompt performance prediction, highlighting fundamental differences between retrieval-based and generation-based performance prediction paradigms.
(2) The LLM self-evaluation baseline exhibits extremely high recall but very low precision and accuracy. This confirms that the model overestimates its ability to answer prompts correctly, predicting most prompts as answerable even when they are not.
(3) The supervised baselines, including BERT-based text classification and BERTPE [35], perform significantly better than the unsupervised baselines. The BERT-based text classification model achieves 66.0% accuracy on TriviaQA, outperforming all other baselines. However, it struggles on HotpotQA, with accuracy dropping to 52.6%, indicating a lack of robustness across datasets.
(4) Our proposed method, VAP3, achieves the highest accuracy, F1-score, and a balanced recall-precision tradeoff across both datasets. VAP3 reaches 69.3% accuracy on TriviaQA and 73.0% on HotpotQA, outperforming BERTPE, which achieves 64.8% on TriviaQA and 71.0% on HotpotQA. F1-scores follow a similar trend, reflecting a balanced ability to predict both answerable and unanswerable prompts. This indicates that VAP3 is more robust to dataset variations, likely due to its use of prompt variations and adversarial training, which improve generalization. Overall, our findings demonstrate that VAP3 significantly outperforms all existing baselines, effectively addressing the prompt sensitivity issue in LLMs.

5 Concluding Remarks

In this paper, we propose the task of Prompt Performance Prediction (PPP), which aims to predict whether an LLM will generate a correct response to a given prompt or not. By enabling proactive performance estimation, PPP allows for better prompt formulation, adaptive model selection, and efficient resource utilization. We proposed VAP3 (Variation-Aware Prompt Performance Prediction), a novel pre-generation PPP approach that integrates prompt variations with adversarial training to effectively estimate prompt performance. Our experimental results on PromptSET (HotpotQA & TriviaQA) demonstrate that by leveraging prompt variations, VAP3 consistently outperforms existing baselines, achieving higher accuracy and better generalization.

References

[1] Negar Arabzadeh, Radin Hamidi Rad, Maryam Khodabakhsh, and Ebrahim Bagheri. 2023. Noisy Perturbations for Estimating Query Difficulty in Dense Retrievers. In *CIKM*. 3722–3727.

[2] Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Hassan Awadallah, Charles L. A. Clarke, and Julia Kiseleva. 2024. Assessing and Verifying Task Utility in LLM-Powered Applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 21868–21888. doi:10.18653/v1/2024.emnlp-main.1219

[3] Negar Arabzadeh, Chuan Meng, Mohammad Aliannejadi, and Ebrahim Bagheri. 2024. Query Performance Prediction: From Fundamentals to Advanced Techniques. In *European Conference on Information Retrieval*. Springer, 381–388.

[4] Negar Arabzadeh, Mahsa Seifikar, and Charles LA Clarke. 2022. Unsupervised Question Clarity Prediction Through Retrieved Item Coherency. In *CIKM*. 3811–3816.

[5] 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.

[6] 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 (CIKM '19)*. ACM, 2109–2112. doi:10.1145/3357384.3358152

[7] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. 2020. *Neural Embedding-Based Metrics for Pre-retrieval Query Performance Prediction*. Springer International Publishing, 78–85. doi:10.1007/978-3-030-45442-5_10

[8] Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. 2024. Aligning Large Language Model with Direct Multi-Preference Optimization for Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 76–86. doi:10.1145/3627673.3679611

[9] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting Is Programming: A Query Language for Large Language Models. *Proceedings of the ACM on Programming Languages* 7, PLDI (June 2023), 1946–1969. doi:10.1145/3591300

[10] Nicolas Bizzozzero, Ihab Bendidi, and Olivier Risser-Maroix. 2024. Prompt Performance Prediction for Image Generation. arXiv:2306.08915 [cs.IR] https://arxiv.org/abs/2306.08915

[11] 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.

[12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.

[13] Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. *arXiv preprint arXiv:2410.02185* (2024).

[14] Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. 2024. POSIX: A Prompt Sensitivity Index For Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 14550–14565. doi:10.18653/v1/2024.findings-emnlp.852

[15] Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937* (2023).

[16] Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting Query Performance. In *SIGIR*. 299–306.

[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 *WSDM*. 201–209.

[18] Suchana Datta, Sean MacAvaney, Debasis Ganguly, and Derek Greene. 2022. A 'Pointwise-Query, Listwise-Document'based Query Performance Prediction Approach. In *SIGIR*. 2148–2153.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. 4171–4186.

[20] Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. *arXiv preprint arXiv:2406.12334* (2024).

[21] Guglielmo Faggioli, Nicola Ferro, Cristina Ioana Muntean, Raffaele Perego, and Nicola Tonellotto. 2023. A Geometric Framework for Query Performance Prediction in Conversational Search. In *SIGIR*. 1355–1365.

[22] Guglielmo Faggioli, Oleg Zendel, J Shane Culpepper, Nicola Ferro, and Falk Scholer. 2022. sMARE: A New Paradigm to Evaluate and Understand Query Performance Prediction Methods. *Information Retrieval Journal* 25, 2 (2022), 94–122.

[23] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

[24] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [stat.ML] https://arxiv.org/abs/1412.6572

[25] Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-Assessment Tests are Unreliable Measures of LLM Personality. arXiv:2309.08163 [cs.CL] https://arxiv.org/abs/2309.08163

[26] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (2023).

[27] Helia Hashemi, Hamed Zamani, and W Bruce Croft. 2019. Performance Prediction for Non-factoid Question Answering. In *ICTIR*. 55–58.

[28] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *CIKM*.

[29] Ling Huang, Wanqiu Deng, Yiling Jiang, and Qinghua Zhong. 2025. Development trends of large language models and their applications in green digital intelligence of supply chains. In *Proceedings of the 2024 5th International Conference on Computer Science and Management Technology (ICCSMT '24)*. Association for Computing Machinery, New York, NY, USA, 770–774. doi:10.1145/3708036.3708165

[30] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. 43–58.

[31] Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. 2023. Retrieving Supporting Evidence for Generative Question Answering. *arXiv preprint arXiv:2309.11392* (2023).

[32] Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024. Enhancing question answering for enterprise knowledge bases using large language models. In *International Conference on Database Systems for Advanced Applications*. Springer, 273–290.

[33] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551 [cs.CL] https://arxiv.org/abs/1705.03551

[34] 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.

[35] Maryam Khodabakhsh, Fattane Zarrinkalam, and Negar Arabzadeh. 2024. *BertPE: A BERT-Based Pre-retrieval Estimator for Query Performance Prediction*. Springer Nature Switzerland, 354–363. doi:10.1007/978-3-031-56063-7_27

[36] Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927* (2024).

[37] Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? arXiv:2311.01967 [cs.CL] https://arxiv.org/abs/2311.01967

[38] Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for zero-shot open-domain QA. *arXiv preprint arXiv:2212.08635* (2022).

[39] Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. How are Prompts Different in Terms of Sensitivity?. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 5833–5856. doi:10.18653/v1/2024.naacl-long.325

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083 [stat.ML] https://arxiv.org/abs/1706.06083

[41] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Query Performance Prediction using Relevance Judgments Generated by Large Language Models. *arXiv preprint arXiv:2404.01012* (2024).

[42] Dipasree Pal and Debasis Ganguly. 2021. Effective Query Formulation in Conversation Contextualization: A Query Specificity-based Approach. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 177–183.

[43] Eduard Poesina, Adriana Valentina Costache, Adrian-Gabriel Chifu, Josiane Mothe, and Radu Tudor Ionescu. 2024. PQPP: A Joint Benchmark for Text-to-Image Prompt and Query Performance Prediction. arXiv:2406.04746 [cs.CV] https://arxiv.org/abs/2406.04746

[44] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. 2024. Llm4eval: Large language model for evaluation in ir. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3040–3043.

[45] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles L. A. Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. 2024. LLMJudge: LLMs for Relevance Judgments. arXiv:2408.08896 [cs.IR] https://arxiv.org/abs/2408.08896

[46] Thilina C Rajapakse, Andrew Yates, and Maarten de Rijke. 2024. Simple Transformers: Open-source for All. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 209–215.

[47] Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. 2025. Benchmarking Prompt Sensitivity in Large Language Models. arXiv:2502.06065 [cs.CL] https://arxiv.org/abs/2502.06065

[48] Haggai Roitman, Shai Erera, and Guy Feigenblat. 2019. A Study of Query Performance Prediction for Answer Quality Determination. In *ICTIR*. 43–46.

[49] 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. *IPM* 56, 3 (2019), 1026–1045.

[50] Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2395–2400.

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

[52] Surendra Sarnikar, Zhu Zhang, and J Leon Zhao. 2014. Query-performance Prediction for Effective Query Routing in Domain-specific Repositories. *JASIST* 65, 8 (2014), 1597–1614.

[53] Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics* 12 (2024), e55318.

[54] Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 23–31.

[55] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv:2302.11382 [cs.SE]

[56] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[57] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2369–2380. doi:10.18653/v1/D18-1259

[58] Hamed Zamani, W Bruce Croft, and J Shane Culpepper. 2018. Neural Query Performance Prediction Using Weak Supervision from Multiple Signals. In *SIGIR*. 105–114.

[59] Yun Zhou and W Bruce Croft. 2007. Query Performance Prediction in Web Search Environments. In *SIGIR*. 543–550.

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

[61] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. arXiv:2410.12405 [cs.CL] https://arxiv.org/abs/2410.12405

[62] Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. arXiv:2410.12405 [cs.CL] https://arxiv.org/abs/2410.12405