Improving Zero-shot LLM Re-Ranker with Risk Minimization

Xiaowei Yuan[1,2,3], **Zhao Yang**[1,2], **Yequan Wang**[3,*], **Jun Zhao**[1,2], **Kang Liu**[1,2,*]
[1]The Key Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing Academy of Artificial Intelligence, Beijing, China
yuanxiaowei2022@ia.ac.cn, {zhao.yang, jzhao, kliu}@nlpr.ia.ac.cn
tshwangyequan@gmail.com

Abstract

In the Retrieval-Augmented Generation (RAG) system, advanced Large Language Models (LLMs) have emerged as effective Query Likelihood Models (QLMs) in an unsupervised way, which re-rank documents based on the probability of generating the query given the content of a document. However, directly prompting LLMs to approximate QLMs inherently is biased, where the estimated distribution might diverge from the actual document-specific distribution. In this study, we introduce a novel framework, UR^3, which leverages Bayesian decision theory to both quantify and mitigate this estimation bias. Specifically, UR^3 reformulates the problem as maximizing the probability of document generation, thereby harmonizing the optimization of query and document generation probabilities under a unified risk minimization objective. Our empirical results indicate that UR^3 significantly enhances re-ranking, particularly in improving the Top-1 accuracy. It benefits the QA tasks by achieving higher accuracy with fewer input documents.

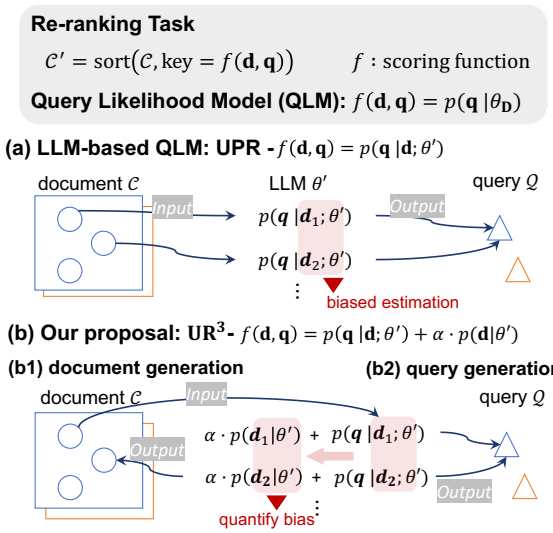


Figure 1: Method comparison in the re-ranking task. (a) The framework of LLM-based QLM method: unsupervised passage re-ranker (UPR). (b) The framework of our proposal: Unsupervised Risk-minimization Re-Ranker (**UR^3**); (b1) calculating document generation probability to quantify the biased model estimation; (b2) calculating the query generation probability to measure relevance.

1 Introduction

Large Language Models (LLMs) exhibit remarkable capabilities but face several challenges including hallucination and outdated knowledge (Zhao et al., 2023; Ji et al., 2023). Retrieval-Augmented Generation (RAG) has emerged as a promising solution by incorporating external knowledge (Ram et al., 2023; Gao et al., 2023). In the RAG system, a re-ranking model can serve as a second-pass document optimizer and refiner for the knowledge retrieval. This is particularly critical in open-domain Question Answering (QA) tasks, where it leads to large gains in performance (Karpukhin et al., 2020; Zhu et al., 2023). The re-ranker assesses the relevance of the documents retrieved by the initial retriever (e.g., BM25 (Robertson and Zaragoza, 2009)) and effectively prioritizes the most relevant items at the top. This not only enhances retrieval efficiency and responsiveness but also resolves the challenge of context window expansion by limiting the total number of documents (Gao et al., 2023).

Most previous approaches trained the re-ranker on manual supervision signals (Karpukhin et al., 2020; Nogueira et al., 2020; Formal et al., 2021), which require significant human efforts and demonstrate weak generalizability (Izacard et al., 2021; Mokrii et al., 2021). As the size of models scales up (e.g., exceeding 10 billion parameters), it becomes increasingly difficult to fine-tune the dedicated re-ranking models. To address this challenge, recent efforts have attempted to leverage the zero-shot language understanding and generation capabilities of

*Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17967–17983
November 12-16, 2024

LLMs to directly enhance document re-ranking in an unsupervised way.

Recent studies have explored LLMs for permutation generation (Ma et al., 2023; Sun et al., 2023) as re-rankers, which yield significant performance by generating a ranked list of a group of documents. However, these models face high time complexity with long lists and the performance is highly sensitive to the document order in the prompt. (Zhu et al., 2023). In this paper, we consider a unsupervised query generation method based on Query Likelihood Model (QLM) (Ponte and Croft, 1998; Hiemstra, 2001; Zhai and Lafferty, 2001), which judges the relevance of each query-document pair independently, thus offering lower time complexity. The core idea behind QLM is to infer a language model θ_D for each document $\mathbf{d}$, and to rank the documents based on the likelihood of the query according to this model $p(\mathbf{q} \mid \theta_D)$.

The typical LLM-based QLM is called Unsupervised Passage Re-ranker (UPR) (Sachan et al., 2022). It leverages a LLM θ' to score the probability of generating the question $\mathbf{q}$ conditioned on the input document $\mathbf{d}$ as $p(\mathbf{q} \mid \mathbf{d}; \theta')$, highlighting the zero-shot ranking capabilities of the LLM-based QLM. Upon closer examination, **an inherent estimation bias occurs** when employing $p(\mathbf{d}; \theta')$ to approximate $p(\theta_D)$. As illustrated in Figure 1, the estimated distribution $p(\mathbf{d}; \theta')$ might not accurately reflect the actual document-specific distribution, $p(\theta_D)$. This divergence primarily stems from the estimation bias in employing a generalized model, such as θ', which is not specifically tuned to capturing the document characteristics necessary for the query generation task (Bender and Koller, 2020; Wang et al., 2022; Zhong et al., 2023).

To bridge the gap between the estimated distribution by LLM $p(\mathbf{q} \mid \mathbf{d}; \theta')$ and the actual document distribution $p(\theta_D)$, we introduce a novel method called Unsupervised Risk-minimization Re-Ranker ($\mathbf{UR^3}$). It characterizes the document selection as a optimization process based on Bayesian decision theory (Wald, 1950; Zhai and Lafferty, 2006a). In specific, to quantify the estimation bias, UR^3 employs the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) to reformulate the minimization of bias as the maximization of document generation probability. Therefore, this approach allows for the simultaneous maximization of both query and document generation probabilities, treating them as a common objective in term of risk minimization.

To prove the effectiveness of UR^3, we verify it in the re-ranking stage in current RAG models for the open-domain QA tasks. In the re-ranking tasks, the results indicate that our method significantly enhances the Top-1 accuracy on the open-domain NQ (Kwiatkowski et al., 2019), WebQ (Berant et al., 2013), and TriviaQA (Joshi et al., 2017) datasets, with improvements of 6.64%, 6.35%, and 3.18% points compared with UPR. In the QA tasks, the Exact Match (EM) and F1 scores exhibit increases of up to 1.48 and 2.06, respectively, when utilizing the fewest document input (only 1).

The contributions of this paper are as follows:

- From the perspective of risk minimization, this paper presents a theoretical formalization to rank the relevance of query-document pairs. This formalization not only considers query generation but also evaluates the estimation bias through document generation probabilities (See §4.2).
- The enhancement in performance is notable for higher-ranked results, with the most pronounced improvements at the Top-1. This significantly benefits the QA tasks by achieving higher accuracy with fewer input documents (See §4.3).

2 Related Work

Re-rankers serve as the second-pass document filter in IR, based on the relevance between the query and the documents. Recently, LLMs have attracted significant attention in the field of IR, with numerous innovative approaches being proposed for re-ranking tasks (Zhu et al., 2023; Gao et al., 2023). Existing instructions for zero-shot document re-ranking with LLMs can be classified into three types: query generation (Sachan et al., 2021; Zhuang et al., 2023), relevance generation (Liang et al., 2022) and permutation generation (Ma et al., 2023; Sun et al., 2023). However, permutation generation models face high time complexity with long lists, and relevance generation method does not have an advantage in terms of performance compared to others (Zhu et al., 2023). In this paper, we focus on the application of query generation LLMs in an unsupervised way.

Language modeling approaches to information retrieval are attractive and promising because they connect the problem of retrieval with that of language model estimation. UPR (Sachan et al., 2022)

introduces instructional query generation methods by LLMs, as the query-document relevance score is determined by the average log-likelihood of generating the actual query tokens based on the document. It has been proven that some LLMs yield significant performance in zero-shot document re-ranking. Recently, research (Zhuang et al., 2023) has also shown that the LLMs that are pre-trained without any supervised instruction fine-tuning (such as LLaMA (Touvron et al., 2023a)) also yield robust zero-shot ranking ability.

Another line is to optimize prompt for better performance. For example, a discrete prompt optimization method Co-Prompt (Cho et al., 2023) is proposed for better prompt generation in re-ranking tasks. Besides, PaRaDe (Drozdov et al., 2023) introduces a difficulty-based method for selecting few-shot demonstrations to include in prompts, demonstrating significant improvements over zero-shot prompts. But the prompt engineering is not within the scope of this paper. Our prompt adheres to the original setup as UPR (e.g., "*Please write a query based on this document*") in a zero-shot manner.

3 UR^3: Unsupervised Risk-minimization Re-Ranker

Existing methods (Sachan et al., 2022; Zhuang et al., 2023) have limited performance in re-ranking due to the oversight of biased estimation when considering a LLM conditioned on the input document $p(\mathbf{d};\theta')$ as the actual document language distribution $p(\theta_D)$.

To tackle the problem, we introduce a novel re-ranking model **UR^3**, which considers not only the query generation probability (§3.3.1) but also the quantification of bias (§3.3.2). For the latter, our method characterizes the distribution discrepancy between an actual document language model $p(\theta_D)$ and the LLM $p(\mathbf{d};\theta')$. Utilizing the distance-based risk-minimization Bayes decision, the estimation bias can be reformulated as the probability of document generation, thereby forming a common optimization objective with the query generation process.

3.1 Problem Formalization

In a retrieval system, a query $\mathbf{q}$ from a user $\mathcal{U}$ is assumed to sampled from a query-based empirical distribution $p(\mathbf{q} \mid \theta_{Q_e})$[1]. A document model θ_D is selected from the document source $\mathcal{S}$ according to the distribution $p(\theta_D \mid \mathcal{S})$, and then this model generates a document according to $p(\mathbf{d} \mid \theta_D)$.

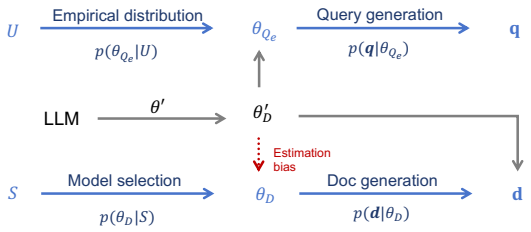


Figure 2: The process for a LLM-based re-ranking method in the view of Bayes decision theory.

Let $\mathcal{C} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k\}$ be a set of candidates from source $\mathcal{S}$, where we assume that the retriever provides the K most relevant documents. For a candidate document $\mathbf{d}$, QLM (Ponte and Croft, 1998) estimates the conditional probability $p(\mathbf{q}|\theta_D)$[2], which captures how well the document "fits" the particular query. Previous QLM-based work (Sachan et al., 2022; Cho et al., 2023; Drozdov et al., 2023) score each document by computing the likelihood of the query conditioned on the input document as $p(\mathbf{q} \mid \mathbf{d};\theta')$. They approximate the document language model θ_D by applying $\mathbf{d}$ as input into a pre-trained LLM θ', formulated as:

$$p(\theta'_D) \stackrel{\text{def}}{=} p(\mathbf{d};\theta') \tag{1}$$

3.2 Bayes Decision Theory

The standard retrieval problem can be regarded as a decision problem where the decision involves choosing the best ranking. Zhai and Lafferty (2006b) formalizes the decision problem within a probabilistic framework of the Bayesian decision theory. A possible action a is to return a single document based on the expected risk R, which is associated with a loss $L(a,\theta)$:

$$R(a \mid \mathcal{U}, \mathbf{q}, \mathcal{S}, \mathcal{C}) = \int_{\Theta} L(a,\theta) p(\theta \mid \mathcal{U}, \mathbf{q}, \mathcal{S}, \mathcal{C}) d\theta \tag{2}$$

The Bayesian decision rule is then to present the document list a^* having the least expected risk:

$$a^* = \arg\min_a R(a|U, \mathbf{q}, S, C) \tag{3}$$

[1]Here we do not define a user-specific query model that encodes detailed knowledge about the user, but rather an empirical distribution θ_{Q_e} for mathematical convenience. The query language model is concentrated on the actual query terms.

[2]For convenience, the subscript i is omitted in subsequent notations.

We extend the framework to allow for a consideration of the the approximate document model θ'_D with LLM θ', as illustrated in Figure 2. The **expected risk of action** a can be formulated as:

$$R(\mathbf{d};\mathbf{q}) \overset{\text{def}}{=} R(a=\mathbf{d} \mid U,\mathbf{q},\mathcal{S},C,\theta') \propto L(\hat{\theta}_{\mathbf{q}},\theta'_D,\hat{\theta}_{\mathbf{d}}) \quad (4)$$

where the distribution of θ'_D is determined by the distribution $p(\mathbf{d};\theta')$, and

$$\hat{\theta}_{\mathbf{q}} = \arg\max_{\theta_{Q_e}} p(\theta_{Q_e} \mid \mathbf{q},\mathcal{U})$$
$$\hat{\theta}_{\mathbf{d}} = \arg\max_{\theta_D} p(\theta_D \mid \mathbf{d},\mathcal{S})$$

The detailed derivations are presented in Appendix A.

To summarize, the document set $\mathcal{C}$ is represented through a series of k sequential decisions. This process yields a list of documents ranked in ascending order according to the $R(\mathbf{d};\mathbf{q})$. A smaller loss L means a better ranking for the document.

3.3 Distance-based Loss Functions

In this section, we conceptualize the loss function, L, as a distance-based function, Δ, quantified using KL divergence, initially introduced by Lafferty and Zhai (2001).

Based on the dependency relationships illustrated in Figure 2, the distance among models can be split into the sum of the following two terms, where the details refer to Appendix B.

$$L(\hat{\theta}_{\mathbf{q}},\theta'_D,\hat{\theta}_{\mathbf{d}}) \approx c_1\Delta(\hat{\theta}_{\mathbf{q}},\theta'_D)+c_2\Delta(\theta'_D,\hat{\theta}_{\mathbf{d}}) \quad (5)$$

where $c_1 > 0$ and $c_2 > 0$ are constants. Therefore, the following formula can be derived:

$$R(\mathbf{d};\mathbf{q}) \propto \Delta(\hat{\theta}_{\mathbf{q}},\theta'_D) + \alpha\Delta(\hat{\theta}_{\mathbf{d}},\theta'_D) \quad (6)$$

where the α is proportional to $c_2/\ c_1$. Then we will characterize that the minimum risk ranking criterion as the sum of probability of query generation (§3.3.1) and document generation (§3.3.2), respectively.

3.3.1 Probability of Query Generation

Given $\hat{\theta}_{\mathbf{q}}$ is a distribution that represents an empirical distribution of query $\mathbf{q}$, where $\mathbf{q} = q_1 q_2 \dots q_m$, we have[3]:

$$\begin{aligned}\Delta(\hat{\theta}_{\mathbf{q}},\hat{\theta}'_{\mathbf{d}}) &\overset{\text{def}}{=} \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{q}}) \parallel p(\theta'_D)\,] \\ &\propto -\log p(\mathbf{q} \mid \theta'_D) + c_{\mathbf{q}} \\ &\propto -\frac{1}{m}\sum_{i=1}^{m}\log p(q_i \mid \mathbf{q}_{<i},\mathbf{d};\theta')\end{aligned} \quad (7)$$

where the constant $c_{\mathbf{q}}$ presents the entropy of the query model. This is precisely the log-likelihood criterion that has been in used in the language modeling approaches of query generation (Sachan et al., 2022; Zhuang et al., 2023).

3.3.2 Probability of Document Generation

Following previous studies (Ponte and Croft, 1998; Hiemstra, 2001; Zhai and Lafferty, 2001), $p(\mathbf{d})$ is assumed to be uniformly distributed, if we view θ' as a stochastic variable, then

$$P(\mathbf{d},\theta') = P(\mathbf{d})P(\theta' \mid \mathbf{d}) \propto P(\theta' \mid \mathbf{d}) \quad (8)$$

Therefore, the distance from the approximate distribution θ'_D to the actual posterior distribution $\hat{\theta}_{\mathbf{d}}$ is formulated as:

$$\begin{aligned}\Delta(\hat{\theta}_{\mathbf{d}},\theta'_D) &\overset{\text{def}}{=} \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\mathbf{d},\theta')\,] \\ &\propto \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\theta' \mid \mathbf{d})\,]\end{aligned} \quad (9)$$

The calculation of Formula 9 can be equivalently reformulated as the computation of the Evidence Lower Bound (ELBO) via variational inference (Hoffman et al., 2013)[3]:

$$\mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\theta' \mid \mathbf{d})] = -\mathrm{ELBO}(\theta) + \log p(\mathbf{d}) \quad (10)$$

where

$$\mathrm{ELBO}(\theta) = \mathbb{E}[\log p(\mathbf{d} \mid \theta')] - \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\theta')]$$

Since the latter KL divergence term in ELBO(θ) is same for all $\mathbf{d}$ for a specific LLM, the following formula can be derived:

$$\Delta(\hat{\theta}_{\mathbf{d}},\hat{\theta}'_{\mathbf{d}}) \propto -\mathbb{E}[\log p(\mathbf{d} \mid \theta')] \quad (11)$$

Let $\mathbf{d} = d_1, d_2, ..., d_n$, the final risk minimization object can be formulated as the proportional sum of query and document generation probabilities based on Formula 7 and 11:

$$\begin{aligned}R(\mathbf{d};\mathbf{q}) \propto &-\frac{1}{m}\sum_{i=1}^{m}\log p(q_i \mid \mathbf{q}_{<i},\mathbf{d};\theta') \\ &-\alpha\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\log p(d_i \mid \mathbf{d}_{<i};\theta')\right)\end{aligned} \quad (12)$$

[3]The theoretical derivations are detailed in Appendix C.

where α is a hyperparameter. The expectation of the term in Formula 11 is calculated as the document generation probability on LLM θ', which synchronizes the computation of the query and the document within one-time inference. The detailed instructions are included in Appendix D.

4 Experiments

4.1 Experimental Setup

Datasets. For the document retrieval in the QA task, we use the three popular datasets of open-domain QA: NaturalQuestions (NQ; (Kwiatkowski et al., 2019)), WebQuestions (WebQ; (Berant et al., 2013)) and TriviaQA (Joshi et al., 2017). For re-ranking, we utilize the pre-processed English Wikipedia dump from December 2018, as released by Karpukhin et al. (2020), as the source of evidence documents. Then we apply the ranking results to generate answers for questions to evaluate the QA performance.

We additionally employ the BEIR Benchmark (Thakur et al., 2021) for a comprehensive retrieval evaluation in Appendix E.

Retrievers. In our re-ranking experiments, we retrieve documents from both unsupervised and supervised retrievers, including three unsupervised retrievers—Contriever (Izacard et al., 2021), BM25 (Robertson and Zaragoza, 2009), and MSS (Sachan et al., 2021)—and one supervised retriever, DPR (Karpukhin et al., 2020).

Baselines We adopt three unsupervised re-ranking methods as the baselines: RankGPT (RG) (Sun et al., 2023), UPR (Sachan et al., 2022) and Interpolation (Int.) (Zhuang et al., 2023).

- RankGPT aims to directly rank a list of documents employing a sliding window strategy to re-rank subsets of candidate documents based on a LLM[4].
- UPR leverages a LLM to obtained the query-document relevance score, which is determined by the log-likelihood of generating the actual query tokens based on the document.
- Interpolation method linearly combines the UPR score with the scores from the first-stage retriever using a weighted sum of scores. We apply the method to both UPR and UR^3 methods with the same weight configuration.

[4]For a fair comparison, the implementation of the RankGPT method is based on the LLaMA2-7B-Chat model.

For UR^3, the values of α is set to 0.25. Detailed analyses about the hyperparameter are provided in Appendix H.

Metrics Following previous work (Thakur et al., 2021; Sachan et al., 2022), we compute the conventional Top-K retrieval accuracy, nDCG@K and MAP@K metrics to evaluate the re-ranking performance. And we use the EM and F1 scores for evaluating the QA performance of LLMs.

LLMs For the re-ranking task, our experiments are conducted on LLaMA2 (7B) (Touvron et al., 2023b), Mistral (7B) (Jiang et al., 2023) and GPT-Neo (2.7B) (Gao et al., 2020) models.

For the QA task, a reader processes the documents retrieved by the retriever to generate the answer to the query. We respectively employ the LLaMA2 (7B and 13B), Mistral (7B) and Gemma (7B) (Mesnard et al., 2024) models as the reader.

4.2 Document Re-ranking

We evaluate the performance of our UR^3 method across all evaluated datasets and retrievers.

4.2.1 Overall Performance

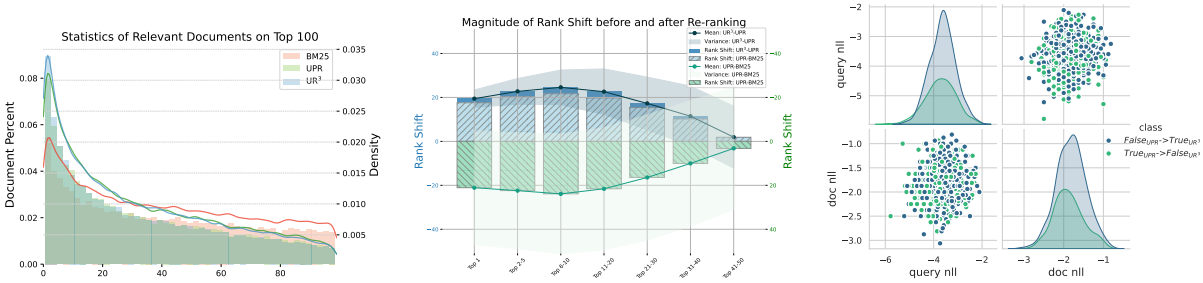
Comprehensive better than UPR. As shown in the Table 1, the results demonstrate that UR^3 enhances the overall rankings of the Top-100 documents, as reflected by an average increase of 1-2% in the MAP@100 metric. Furthermore, improvements are observed across all nDCG@K metrics, indicating that UR^3 prioritizes relevant documents more effectively compared to the UPR method. Closer examination of the Top-K metrics reveals that UR^3 shows greater accuracy enhancements for rankings closer to the top, with the most substantial increase (up to 6.64) observed at Top-1 accuracy. This significantly enhances the suitability of our method for open-domain question answering tasks. Additionally, it potentially alleviates the issues associated with the limited input window length of large models, as our method achieves higher relevance scores with fewer input documents.

Why does RankGPT perform poorly? Interestingly, the RankGPT method yields lower ranking results than the initial retrieval. This can be attributed to the observation that competitive performance is predominantly realized by model based on GPT-4 (Zhu et al., 2023). When utilizing smaller parameterized language models, such as LLaMA2-7B, the RankGPT method underperforms compared to other methods.

Datasets	Metric	Contriever Orig.*	Contriever RG	Contriever UPR (+Int.)	Contriever UR^3 (+Int.)	BM25 Orig.	BM25 RG	BM25 UPR (+Int.)	BM25 UR^3 (+Int.)	MSS Orig.	MSS RG	MSS UPR (+Int.)	MSS UR^3 (+Int.)	DPR Orig.	DPR RG	DPR UPR (+Int.)	DPR UR^3 (+Int.)
NQ	Top-1	22.16	13.07	32.38 (32.49)	**37.67** (36.37)	22.11	17.51	32.69 (33.10)	**38.01** (37.42)	19.28	15.35	32.83 (33.49)	**37.48** (36.43)	46.34	37.06	37.65 (48.45)	44.29 (**52.24**)
	Top-5	47.29	46.87	61.41 (61.00)	63.96 (**64.10**)	43.77	38.25	59.83 (59.50)	**61.97** (61.19)	41.25	35.76	59.28 (59.22)	**61.08** (60.61)	68.28	67.67	69.20 (73.85)	71.99 (**74.99**)
	Top-20	67.87	67.51	76.12 (76.26)	**76.57** (76.59)	62.94	62.94	**73.16** (72.63)	72.96 (72.88)	59.97	60.22	71.30 (70.97)	**71.47** (71.25)	80.06	79.70	82.66 (83.10)	82.99 (**83.32**)
	nDCG@1	22.16	13.07	32.38 (32.49)	**37.67** (36.37)	22.11	17.51	32.69 (33.10)	**38.01** (37.42)	19.28	15.35	32.83 (33.39)	**37.48** (36.43)	46.34	17.06	37.65 (48.45)	44.29 (**52.24**)
	nDCG@5	21.70	19.10	33.35 (33.08)	**36.89** (36.36)	21.63	17.43	33.89 (33.89)	**37.12** (36.51)	18.97	15.38	34.39 (34.45)	**37.12** (36.53)	40.62	32.79	38.94 (45.51)	43.05 (**47.97**)
	nDCG@20	26.15	24.20	39.08 (38.79)	**41.60** (41.26)	25.75	23.45	39.27 (39.17)	**41.27** (40.87)	22.88	21.10	39.36 (39.18)	**41.15** (40.36)	42.42	36.43	44.78 (49.34)	47.66 (**50.95**)
	MAP@100	20.71	18.68	31.56 (31.18)	**33.94** (33.46)	20.78	18.37	32.13 (32.05)	**34.05** (33.66)	18.11	16.27	32.32 (31.98)	**34.10** (33.23)	34.89	28.35	36.64 (41.39)	39.38 (**42.91**)
WebQ	Top-1	19.98	18.65	26.62 (28.05)	**32.53** (30.81)	18.90	17.32	27.56 (28.54)	**33.91** (33.56)	11.66	11.96	26.38 (25.44)	**29.38** (27.66)	44.83	37.16	39.32 (46.26)	42.18 (**48.03**)
	Top-5	43.45	41.39	54.92 (55.07)	**58.71** (58.12)	41.83	40.16	54.13 (54.33)	55.17 (**55.76**)	29.04	28.54	48.67 (49.02)	49.85 (**50.44**)	65.01	59.30	66.83 (68.21)	66.88 (**68.95**)
	Top-20	65.70	65.50	72.69 (72.44)	**73.43** (72.79)	62.40	62.35	68.50 (68.55)	**69.54** (69.14)	49.21	49.51	63.19 (63.24)	**62.40** (**62.40**)	74.61	74.46	76.67 (76.53)	76.96 (**77.36**)
	nDCG@1	19.98	18.65	26.62 (28.05)	**32.53** (30.81)	18.90	17.32	27.56 (28.54)	**33.91** (33.56)	11.66	11.96	26.38 (25.44)	**29.38** (27.66)	44.83	37.16	39.32 (46.26)	42.18 (**48.03**)
	nDCG@5	18.64	17.44	26.78 (26.90)	**30.82** (29.89)	19.36	17.95	27.39 (28.27)	30.72 (**30.86**)	11.57	10.81	26.67 (26.08)	**28.21** (27.53)	39.76	34.35	38.66 (42.59)	40.34 (**43.67**)
	nDCG@20	22.22	21.53	31.18 (31.06)	**33.79** (33.21)	22.12	21.41	31.44 (31.82)	**33.62** (33.43)	14.84	14.45	32.46 (31.83)	**33.20** (32.45)	38.95	36.21	41.81 (44.32)	42.65 (**44.74**)
	MAP@100	18.79	18.22	25.92 (25.62)	**27.82** (27.24)	19.15	18.39	26.63 (26.81)	**28.09** (28.02)	12.03	11.53	26.20 (25.32)	**26.84** (25.95)	33.32	30.44	36.46 (38.58)	36.82 (**38.66**)
TriviaQA	Top-1	34.16	25.17	51.77 (51.17)	**54.95** (53.99)	46.30	35.10	55.85 (57.76)	58.70 (**59.80**)	30.76	21.19	52.84 (52.74)	**54.35** (53.83)	57.47	37.16	62.55 (66.77)	63.47 (**67.23**)
	Top-5	59.49	50.99	73.81 (73.69)	**74.31** (74.02)	66.28	57.64	75.60 (75.98)	**76.04** (75.86)	52.65	43.16	70.94 (70.78)	**71.12** (70.78)	72.40	58.84	78.74 (79.06)	78.84 (**79.19**)
	Top-20	73.91	74.10	80.08 (80.01)	**80.22** (80.16)	76.41	76.24	80.68 (**80.70**)	80.66 (**80.70**)	67.18	67.44	76.34 (76.28)	**76.32** (76.27)	79.77	79.66	83.13 (83.07)	83.15 (**83.16**)
	nDCG@1	34.16	25.17	51.77 (51.17)	**54.95** (53.99)	46.30	35.10	55.85 (57.76)	58.70 (**59.80**)	30.76	21.19	52.84 (52.74)	**54.35** (53.83)	57.74	37.16	62.55 (66.77)	63.47 (**67.23**)
	nDCG@5	30.46	23.63	49.27 (48.52)	**51.21** (50.23)	41.60	32.77	53.55 (56.65)	55.28 (**55.89**)	27.78	19.76	50.47 (50.42)	**51.60** (51.09)	49.69	34.42	59.53 (61.93)	59.99 (**62.16**)
	nDCG@20	31.78	28.69	50.92 (50.36)	**52.06** (51.43)	40.68	36.65	54.93 (55.53)	55.66 (**55.93**)	29.25	25.80	53.12 (52.80)	**53.62** (53.12)	46.33	39.64	59.90 (61.14)	60.06 (**61.18**)
	MAP@100	26.61	23.85	44.69 (43.93)	**45.58** (44.81)	34.85	30.98	49.50 (49.91)	49.93 (**50.11**)	24.02	20.83	47.02 (46.47)	**47.45** (46.71)	39.40	33.01	54.21 (55.34)	54.31 (**55.42**)

* Orig. indicates the original results from the retriever, where no re-ranking method is employed.

Table 1: Re-ranking results on the test set of datasets of the Top-100 retrieved documents with the LLaMA2-7B model. The best results are highlighted in bold. The higher scores of original retriever compared with UR^3 is highlighted in red. The results on other models (Mistral-7B and GPT-Neo) are detailed in Appendix F.



(a) Relevant Document Distribution (b) Rank Shift of Relevant Document (c) Document Distribution on Top-1

Figure 3: Visualization of Analysis on the Enhanced Performance in the Re-ranking task

Unstable performance on Interpolation. The degree of increase (or decrease) brought by the Interpolation method primarily depends on the performance of the initial retriever. For the supervised DPR retriever, its exposure to relevant paragraphs during training yields substantially higher Top-1 accuracy on the NQ and WebQ datasets. With results from the DPR, the Interpolation method usually brings a significant enhancement in ranking. When combined with our UR^3 method, this can lead to maximal improvement. Conversely, when based on an unsupervised retriever, our method predominantly outperforms the Interpolation method.

4.2.2 Analysis on the Enhanced Performance

To further explore why UR^3 demonstrates greater enhancements for the ranks close to the top, we conduct empirical analyses on the NQ dataset with BM25 retriever.

As illustrated in Figure 3a, we analyze the distributional shift of relevant document positions before and after re-ranking. The histogram represents the proportion of relevant documents at different ranks, and a curve fitting illustrates the trend of this distributional change. Overall, it is evident that UR^3 tends to shift the distribution of relevant documents towards higher ranks compared to UPR.

Then we explore the reason behind the forward shift in this distribution. In Figure 3b, we quantify the magnitude of rank shifts for each relevant document after re-ranking. The blue solid (shadowed) histogram represents the positional change in rank by UR^3 compared to UPR (BM25), while the green shadowed histogram indicates the change by UPR compared to BM25. The bandwidth of the line graph represents the variance of these changes. The figure clearly shows that UR^3 induces smaller shifts in each position; in other words, our method tunes the rankings of relevant documents within a narrower range, thereby obtaining greater benefits to the ranks closer to the top (as the distribution of relevant documents is higher in Figure 3a).

Figure 3c presents a scatter plot that statistically categorizes the relevant documents at the Top-1 rank, comparing UPR and UR^3. Each green dot

Model	LLaMA2-13B						Mistral-7B						Gemma-7B					
	NQ		WebQ		TriviaQA		NQ		WebQ		TriviaQA		NQ		WebQ		TriviaQA	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Contriever	22.02	29.11	19.69	30.21	49.90	57.08	20.69	26.61	14.37	24.39	49.89	56.94	17.40	25.13	14.71	26.05	45.54	53.66
+ Inference with UPR	28.06	35.99	22.00	33.48	58.89	67.05	24.68	31.81	15.90	25.80	59.20	66.83	21.33	29.72	15.29	27.05	55.26	64.05
+ Inference with UR^3	**28.45**	**37.05**	**23.18**	**34.92**	**59.45**	**67.72**	**25.51**	**32.69**	**17.13**	**27.10**	**59.26**	**67.09**	**22.13**	**30.78**	**15.99**	**27.65**	**55.47**	**64.53**
BM25	20.20	27.08	16.39	26.60	55.21	62.91	19.14	25.31	13.29	23.03	54.91	62.15	16.40	23.84	12.35	22.84	52.34	60.89
+ Inference with UPR	27.23	34.92	19.98	31.27	61.86	69.96	24.79	31.71	15.21	25.43	61.68	69.16	21.02	29.47	14.67	25.32	58.53	67.36
+ Inference with UR^3	**28.37**	**36.69**	**21.46**	**32.83**	**62.34**	**70.62**	**25.73**	**32.87**	**16.49**	**26.69**	**61.81**	**69.50**	**22.30**	**30.96**	**15.40**	**25.80**	**58.99**	**67.83**
MSS	19.86	26.16	16.83	27.94	49.29	56.03	18.20	24.28	13.98	23.79	48.41	55.22	14.38	21.44	12.20	22.91	43.56	51.54
+ Inference with UPR	26.81	34.63	20.37	31.77	57.87	65.69	23.35	30.04	16.04	26.24	57.46	65.11	19.83	28.65	14.16	26.20	53.84	62.67
+ Inference with UR^3	**27.26**	**35.35**	**21.16**	**32.81**	**58.28**	**66.15**	**24.29**	**31.22**	**16.58**	**26.72**	**57.69**	**65.34**	**21.27**	**29.69**	**15.06**	**26.29**	**53.97**	**62.77**
DPR	30.30	38.42	22.79	34.36	55.33	62.96	**28.17**	34.9	18.21	**28.55**	55.03	62.24	**24.43**	**33.14**	**16.68**	28.00	50.76	59.59
+ Inference with UPR	29.09	37.21	23.18	34.09	60.87	69.02	26.51	34.13	16.98	26.78	60.95	68.71	22.33	31.36	15.99	27.44	57.23	66.19
+ Inference with UR^3	**30.80**	**39.23**	**24.36**	**36.15**	**61.21**	**69.35**	27.84	**35.36**	**18.31**	28.13	**61.12**	**68.94**	23.91	32.80	16.54	**28.16**	**57.43**	**66.40**

Table 2: EM and F1 scores for the open-domain QA task. We perform inference with the re-ranked Top-1 results of Table 1. The best performing models are highlighted in bold. We highlight the best scores obtained by original retriever in red. We also conduct inference on the re-ranking results of Mistral-7B in Table 11.

NQ Dataset	Top-1		Top-3		Top-5	
	EM	F1	EM	F1	EM	F1
Contriever	15.09	22.00	14.93	20.36	18.50	23.80
+ Inference with UPR	20.97	27.90	19.31	25.51	22.33	28.61
+ Inference with UR^3	**21.93**	**29.06**	**19.98**	**25.77**	**22.55**	**28.72**
BM25	15.65	21.38	14.35	20.04	16.45	22.05
+ Inference with UPR	20.75	27.71	19.56	25.90	21.91	28.28
+ Inference with UR^3	**21.75**	**28.91**	**20.02**	**26.90**	**22.27**	**28.70**
MSS	13.60	19.34	13.63	19.48	16.59	22.22
+ Inference with UPR	19.78	26.98	18.70	24.96	20.72	26.87
+ Inference with UR^3	**21.69**	**28.66**	**19.47**	**26.20**	**21.27**	**27.68**
DPR	23.38	30.69	19.61	26.21	22.60	28.84
+ Inference with UPR	22.13	29.58	20.42	27.19	23.74	30.21
+ Inference with UR^3	**24.29**	**31.16**	**22.08**	**29.24**	**24.54**	**30.96**

Table 3: EM and F1 scores for the open-domain QA task with different number of input documents on the NQ dataset with LLaMA2-7B model. The best performing models are highlighted in bold.

represents a correct calibration by the UR^3 method, where an irrelevant document ranked by UPR is adjusted to a relevant one at the Top-1 rank. Conversely, each blue dot indicates a incorrect calibration by UR^3, where a previously Top-1 relevant document is replaced with an irrelevant one. The axes values denote the respective query/document generation negative log-likelihood loss (nll) discussed in Formula 12. The density distribution of the scatter plot reveals that the positive gains brought about by UR^3 significantly outweigh the negative impacts, which substantiates the improvement observed at the Top-1.

4.3 Application in Question Answering

As discussed above, we have demonstrated that UR^3 significantly enhances ranking performance. In this section, we apply the results of the re-ranking (Table 1) to apply in current RAG models for the evaluation in open-domain QA tasks. Specifically, we utilize the Top-n (n $\leq$ 5) items as inputs to achieve higher scores with fewer documents.

4.3.1 Overall Performance

Not More is Better. We utilize different number of document inputs on NQ dataset to evaluate the QA performance in Table 3. Expanding the documents from Top-1 to Top-3 does not invariably enhance performance; in fact, it occasionally results in a decline in both EM and F1 scores. This trend suggests that increasing the number of documents beyond the most relevant one may introduce noise or less pertinent information. Furthermore, the marginal gains observed when moving from Top-1 to Top-5 are minimal, which underscores the diminishing returns of adding more documents. In summary, utilizing the Top-1 document emerges as the most cost-effective approach, offering a balance between computational efficiency and accuracy.

Superiority Over UPR. As illustrated in the Table 2 and 3, the UR^3 method substantially enhances the performance of QA tasks, achieving superior EM and F1 scores compared to the UPR method. Furthermore, this improvement trend is consistent with the enhancements observed during the re-ranking phase.

Outliers on DPR. In Table 1, it is noteworthy that the highest scores (indicated in red) are achieved by the original DPR method on the Mistral and Gemma models. This is explainable given that both UPR and UR^3 exhibit inferior re-ranking performance compared to DPR for the Top-1 results. However, when employing the LLaMA2-13B model, it demonstrates superior QA performance relative to the DPR. This improvement can be attributed to the strategic use of a generative model with a distribution similar to that of the re-ranker (e.g., within the same LLaMA2 series) in a QA task. In UR^3, maximizing the probability of document

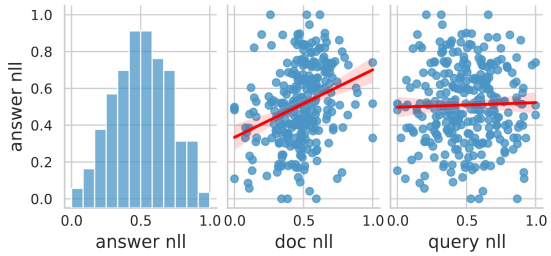
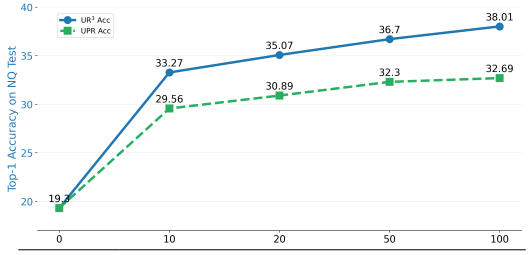


Figure 4: Distributed correlation in answer generation with normalized NLL in the QA task.



Time / Query / A100 GPU (in seconds)	Number of Re-ranked Documents			
	10	20	50	100
UPR	0.2233	0.4765	1.1140	2.3694
UR^3	0.2285	0.4872	1.1282	2.3839

Figure 5: Effect of the number of document candidates on Top-1 accuracy and calculation efficiency when re-ranked with LLaMA2-7B model. Evaluation is done on the NQ test set using BM25 retrieved documents.

generation has a benefit to selecting documents that closely align with the model's distribution. Such alignment significantly enhances the model's reliance on external documents, thereby boosting the overall performance of the QA task.

4.3.2 Analysis on the DPR results

We conduct empirical analysis for the improved performance on DPR retriever with LLaMA2-13B. Figure 4 presents the distributed correlation in answer generation with normalized negative log-likelihood loss of the QA task. When the generation probability is high, the corresponding loss is low. The left panel displays the distribution of answer NLL values. The middle and right panels feature scatter plots that illustrate the relationships between document generation NLL (doc NLL) and query generation NLL (query NLL) during the re-ranking phase. Both scatter plots include a regression line, indicating that, compared to query loss, document loss shows a positive correlation. This suggests that higher generation probabilities for documents increase the likelihood of generating the correct answer. This finding aligns with our understanding that selecting documents closely matching the model's distribution can enhance the model's receptivity to external documents.

4.4 Accuracy and Efficiency Comparison

In this section, we evaluate the impact of the number of document candidates to be re-ranked on both retrieval performance and computational efficiency. The evaluation is conducted using the NQ test set. We re-rank up to the Top-100 documents obtained from the BM25 retriever and measure performance using Top-1 accuracy.

In Figure 5, as the number of re-ranked documents increases, both UR^3 and UPR exhibit improvements in Top-1 accuracy. UR^3 consistently outperforms UPR across all document counts, achieving higher Top-1 accuracy. On the other hand, the computational time per query increases linearly with the number of re-ranked documents for both methods. Despite the increase in computational time, UR^3 maintains a similar computational demand compared to UPR.

In conclusion, the results clearly show that UR^3 significantly enhances performance without incurring additional computational time, which shows the superiority of our method.

5 Conclusion

In this study, we introduced the UR^3 framework, which utilizes Bayesian decision theory to address the estimation bias in QLMs based on LLMs. The novelty of UR^3 lies in its approach to unify the probabilities of query and document generation under a risk minimization framework, thereby enhancing the efficiency of document ranking and question answering.

Our experimental results demonstrate that UR^3 significantly improves re-ranking performance, especially in terms of Top-1 accuracy. In open-domain question-answering tasks, UR^3 contributes to achieving higher accuracy with reduced reliance on the number of input documents.

Limitations

- This paper observes relatively minor improvements when ranking is extended to Top-20 or Top-50, marking a principal limitation. However, a longer context does not necessarily equate to superior performance for the LLMs (Liu et al., 2024), which have also been discussed in Section 4.3.1. Our method achieves a substantial improvement in Top-1

accuracy with comprehensive analysis, which provides optimal cost-effectiveness.

- A limitation of UR^3 is that re-ranking a large pool of document can have a high latency as it involves performing cross-attention whose complexity is proportional to the product of the question and document tokens and the number of layers of the LLM. We have also discussed this quantitatively in Section 4.4.
- UR^3 also shares the inherent limitation associated with all the re-ranking approaches in that its maximum possible performance is dependent on the first-stage retrieval.

Acknowledgments

This work is supported by the National Science and Technology Major Project (No. 2022ZD0116300) and the National Science Foundation of China (No. 62106249).

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Sukmin Cho, Soyeong Jeong, Jeong yeon Seo, and Jong Park. 2023. Discrete prompt optimization via constrained generation for zero-shot re-ranker. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 960–971, Toronto, Canada. Association for Computational Linguistics.

Andrew Drozdov, Honglei Zhuang, Zhuyun Dai, Zhen Qin, Razieh Rahimi, Xuanhui Wang, Dana Alon, Mohit Iyyer, Andrew McCallum, Donald Metzler, and Kai Hui. 2023. PaRaDe: Passage ranking using demonstrations with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14242–14252, Singapore. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Djoerd Hiemstra. 2001. *Using Language Models for Information Retrieval*. Ph.D. thesis, University of Twente, Enschede, Netherlands.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John W. Paisley. 2013. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *CoRR*, abs/2112.09118.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

John D. Lafferty and ChengXiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 111–119. ACM.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295.

Iurii Mokrii, Leonid Boytsov, and Pavel Braslavski. 2021. A systematic evaluation of transfer learning and pseudo-labeling with bert-based ranking models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2081–2085, New York, NY, USA. Association for Computing Machinery.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 275–281. ACM.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Abraham Wald. 1950. *Statistical Decision Functions*. John Wiley.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3170–3179. Association for Computational Linguistics.

ChengXiang Zhai and John Lafferty. 2006a. A risk minimization framework for information retrieval. *Information Processing & Management*, 42(1):31–55.

ChengXiang Zhai and John D. Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 334–342. ACM.

ChengXiang Zhai and John D. Lafferty. 2006b. A risk minimization framework for information retrieval. *Inf. Process. Manag.*, 42(1):31–55.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

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. *CoRR*, abs/2308.07107.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8807–8817. Association for Computational Linguistics.

A Bayes Decision Theory

A possible action of the re-ranking process involves reordering the document subset $\mathcal{C}$ to ensure that a document containing the correct answer is ranked as highly as possible.

In the general framework of Bayesian decision theory, each action a is associated with a loss $L(a, \theta)$, which depends upon $\theta \equiv (\theta_{Q_e}, \{\theta_D\}^k, \{\theta'_D\}^k)$, including the query language model, document language models and estimated models based on a LLM. Based on the Figure 2, a possible action is to return a single document $a = d$, and the loss function depends on θ_{Q_e}, θ_D and θ'_D, the **expected risk of action** a can be formulated as:

$$R(\mathbf{d};\mathbf{q}) \overset{\text{def}}{=} R(a = \mathbf{d} \mid U, \mathbf{q}, \mathcal{S}, C, \theta') = \int_{\theta_{Q_e}} \int_{\Theta'_D} \int_{\Theta_D} L(\theta_{Q_e}, \theta'_D, \theta_D) p(\theta_{Q_e} \mid \mathbf{q}, \mathcal{U}) \times p(\theta'_D \mid \mathbf{d}, \theta')\, p(\theta_D \mid \mathbf{d}, \mathcal{S}), d\theta_{Q_e} d\theta'_D\, d\theta_D$$

Instead of explicitly computing the parameter distributions, the value can be approximated at the posterior mode as follows:

$$R(\mathbf{d};\mathbf{q}) \propto L(\hat{\theta}_{\mathbf{q}}, \theta'_D, \hat{\theta}_{\mathbf{d}}) p(\hat{\theta}_{\mathbf{q}} \mid \mathbf{q}, \mathcal{U}) (\hat{\theta}_{\mathbf{d}} \mid \mathbf{d}, \mathcal{S})$$

where the distribution of θ'_D is determined by the document $\mathbf{d}$ with LLM θ' as $p(\mathbf{d}; \theta')$, thereby the posterior of $p(\theta'_D)$ is a point mass distribution. And

$$\hat{\theta}_{\mathbf{q}} = \arg\max_{\theta_{Q_e}} p(\theta_{Q_e} \mid \mathbf{q}, \mathcal{U})$$
$$\hat{\theta}_{\mathbf{d}} = \arg\max_{\theta_D} p(\theta_D \mid \mathbf{d}, \mathcal{S})$$

Based on the prior assumption in the QLM that the document prior $p(\mathbf{d})$ is uniform, we can infer that $p(\hat{\theta}_d | \mathbf{d}, \mathcal{S})$ is the same for all $\mathbf{d}$. For a specific query, the posterior distribution of the query model can also be dropped, because it is unrelated to the ranking of documents.

Hence, the formula of risk can be simplified as:

$$R(\mathbf{d};\mathbf{q}) \propto L(\hat{\theta}_{\mathbf{q}}, \theta'_D, \hat{\theta}_{\mathbf{d}})$$

To summarize, the document set $\mathcal{C}$ is represented through a series of k sequential decisions. This process yields a list of documents ranked in ascending order according to the $R(\mathbf{d};\mathbf{q})$. A smaller loss L means a better ranking for the document.

B Distance-based Loss Function

KL framework of QLM. The loss is calculated as:

$$L(\theta_{Q_e}, \theta_D) \propto \text{KL}[P(\theta_{Q_e}) || P(\theta_D)]$$

The relevance value of a document with respect to a query is measured by the distance between two models. It is calculated by the KL divergence from the document model distribution $P(\theta_D)$ to the query model distribution $P(\theta_{Q_e})$.

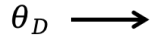


Figure 6: The estimation in QLM

KL framwork of UR³. Based on the QLM framework, the calculation of $L(\theta_{Q_e}, \theta_D, \theta'_D)$ aims to measure the distance between the actual query and document model distributions through a LLM. It can be interpreted as the proportional sum of the distance between the document model θ_D and the estimated model θ'_D, and the distance from the estimated model θ'_D to the query model θ_{Q_e}. We consider the two estimations are independent (left and right items in Figure 7), then we approximate the distance in QLM as the sum of the following items:

$$\begin{aligned} & L(\theta_{Q_e}, \theta_D, \theta'_D) \\ &= c \cdot \text{KL}[P(\theta_{Q_e} || P(\theta_D)] \\ &\approx c_1 \cdot \text{KL}[P(\theta_Q) \parallel P(\theta'_D)] + c_1 \cdot \text{KL}[P(\theta_D) \parallel P(\theta'_D))] \end{aligned}$$

where c, c_1 and c_2 are scale factors.

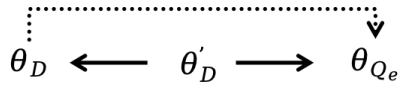


Figure 7: The estimations in UR^3

C Detailed Derivation

C.1 Probability of Query Generation

Following the work of Lafferty and Zhai (2001), when the $\hat{\theta}_{\mathbf{q}}$ is considered as the empirical distribution of the query $\mathbf{q} = q_1 q_2 \ldots q_m$; that is,

$$p(w \mid \hat{\theta}_{\mathbf{q}}) = -\frac{1}{m} \sum_{i=1}^{m} \delta(w, q_i)$$

where, δ is the indicator function, then we obtain

$$
\begin{aligned}
\Delta(\hat{\theta}_{\mathbf{q}}, \theta'_D) &\stackrel{\text{def}}{=} \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{q}}) \parallel p(\theta'_D)\,] \\
&= \sum_w p(w \mid \hat{\theta}_{\mathbf{q}}) \log \frac{p(w \mid \hat{\theta}_{\mathbf{q}})}{p(w \mid \theta'_D)} \\
&\propto -\sum_w p(w \mid \hat{\theta}_{\mathbf{q}}) \log p(w \mid \theta'_D) + c_{\mathbf{q}} \\
&\propto -(\sum_{w \in \mathbf{q} \cap D} \log p(w|\theta'_D) \\
&+ \sum_{w \in \mathbf{q}, w \notin D} \log p(w|\theta'_D)) + c_q \\
&\propto -\sum_{w \in \mathbf{q}} \log p(w|\theta'_D) + c_q \\
&\propto -\log p(\mathbf{q} \mid \theta'_D) + c_{\mathbf{q}} \\
&\propto -\frac{1}{m} \sum_{i=1}^{m} \log p(q_i \mid \mathbf{q}_{<i}, \mathbf{d}; \theta')
\end{aligned}
$$

where the constant $c_{\mathbf{q}}$ presents the entropy of the query model. This is precisely the log-likelihood criterion that has been in used in all work on the language modeling of query generation approach.

C.2 Evidence Lower Bound (ELBO)

Here we view $p(\hat{\theta}_{\mathbf{d}})$ and $p(\theta' \mid \mathbf{d})$ as two distributions across the space of θ. And we denote the distribution $p(\hat{\theta}_{\mathbf{d}})$ as $q(\theta)$ and $p(\theta' \mid \mathbf{d})$ as $p(\theta' \mid \mathbf{d})$, thus

$$
\begin{aligned}
&\mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\theta' \mid \mathbf{d})] \\
&= \mathrm{KL}(q(\theta) \parallel p(\theta \mid \mathbf{d})) \\
&= -\int q(\theta) \log \frac{p(\theta \mid \mathbf{d})}{q(\theta)}\, d\theta \\
&= \int q(\theta) \log q(\theta)\, d\theta - \int q(\theta) \log p(\theta \mid \mathbf{d})\, d\theta \\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\theta \mid \mathbf{d})] \\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q\left[\log \frac{p(\mathbf{d}, \theta)}{p(\mathbf{d})}\right] \\
&= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p(\mathbf{d}, \theta)] + \log p(\mathbf{d}) \\
&= -\mathrm{ELBO}(\theta) + \log p(\mathbf{d})
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathrm{ELBO}(\theta) \\
&= \mathbb{E}_q[\log p(\mathbf{d}, \theta)] - \mathbb{E}_q[\log q(\theta)] \\
&= \mathbb{E}_q[\log p(\mathbf{d} \mid \theta) p(\theta)] - \mathbb{E}_q[\log q(\theta)] \\
&= \mathbb{E}_q[\log p(\mathbf{d} \mid \theta)] + \mathbb{E}_q[\log p(\theta)] - \mathbb{E}_q[\log q(\theta)] \\
&= \mathbb{E}_q[\log p(\mathbf{d} \mid \theta)] + \mathbb{E}_q\left[\frac{\log p(\theta)}{\log q(\theta)}\right] \\
&= \mathbb{E}_q[\log p(\mathbf{d} \mid \theta)] + \int q(\theta) \frac{\log p(\theta)}{\log q(\theta)}\, d\theta \\
&= \mathbb{E}_q[\log p(\mathbf{d} \mid \theta)] - \mathrm{KL}[\, q(\theta) \| p(\theta)] \\
&\approx \log p(\mathbf{d} \mid \theta') - \mathrm{KL}[\, p(\hat{\theta}_{\mathbf{d}}) \parallel p(\theta')]
\end{aligned}
$$

where the expectation of the term $\mathbb{E}_q[\log p(\mathbf{d} \mid \theta)]$ employs the generation probability on LLM θ' as $\log p(\mathbf{d} \mid \theta')$ to minimize computational costs.

D Query Generation Instruction

The query generation instruction (Sachan et al., 2021) uses the log-probability of the query.

> Please write a question based on this passage.
> Passage: `{{passage}}`
> Question: `{{query}}`

Document Generation. Specifically, when calculating the probability of document generation, we compute the negative log loss using the document portion prior to the output query under the current prompt. This approach synchronizes the computation of the query and the document within the same output, significantly reducing computational costs.

E Performance on BEIR Benchmark

To evaluate the generalization of our method, we conducted experiments a popular subset of the BEIR benchmark dataset (Thakur et al., 2021). The evaluation metrics employed are Top-1 accuracy and nDCG@10, the official metric for the BEIR benchmark. For a fair comparison, all the rerankers consider the Top 100 documents retrieved by Contriever. The results are shown in Table 4.

In summary, the results demonstrate the effectiveness of UR^3 as the average Top-1 accuracy improves by 4.39% and the NDCG@10 scores improve by 2.37%. Due to the diversity in datasets, there is a considerable variation in performance gains across them. UR^3 achieves the highest

Dataset	Top-1			NDCG@10		
	Original	UPR	UR^3	Original	UPR	UR^3
NQ	22.16	32.38	37.67	23.19	35.58	38.64
HotpotQA	53.37	84.35	86.02	60.60	85.74	87.43
FIQA	21.14	40.43	39.66	29.16	48.50	48.16
MS-Marco	8.70	11.92	12.16	20.68	27.26	27.41
Trec-Covid	44.00	64.00	72.00	33.43	62.22	66.77
Touche-2020	22.49	10.23	30.64	23.89	19.68	27.03
ArguAna	0.00	9.72	7.33	0.31	28.38	22.91
DBpedia	48.25	43.00	50.75	37.64	40.37	44.65
Fever	52.51	41.13	48.23	70.33	53.32	60.14
Climate-Fever	12.64	7.88	12.18	20.68	15.18	21.04
Scifact	51.71	54.33	55.72	65.04	64.78	65.43
Scidocs	18.67	21.32	21.04	23.81	32.04	31.80
Average	29.63	35.06	**39.45**	34.06	42.75	**45.12**

Table 4: Re-ranking results on the Top100 documents retrieved by Contriever on BEIR benchmark (Thakur et al., 2021). On average, the performances improve both on the Top-1 accuracy and NDCG@10 metrics. The drop in some datasets is highlighted in red.

relative performance improvements on datasets such as Trec-Covid, NQ, Touche-2020 and DBpedia. These datasets typically involve information-seeking questions, which benefit significantly from our advanced re-ranking method.

We also observe a decline in performance on the FIQA, ArguAna, and Scidocs datasets, each characterized by high average document lengths. This suggests that these datasets contain more complex and extensive information, which could be challenging for UR^3 to process effectively, since the UR^3 might struggle with effectively calculating the estimation bias among documents with such complexity, causing a drop in performance. Additionally, the Finance and Science domains might pose specific challenges that UR^3 is not optimized for.

F Re-ranking on Mistral-7B and GPT-Neo-2.7B models

As illustrated in Table 9 and 10, the results demonstrate that UR^3 enhances the overall rankings of the Top-100 documents both on the Mistral and GPT-Neo models. The improvements of our method are observed across all nDCG@K metrics, indicating that UR^3 prioritizes relevant documents more effectively compared to the UPR method.

G Document Distribution on Top-1

The Figure 8 presents the complete results of the scatter plot in Figure 3c, which that statistically categorizes the relevant documents at the Top-1 rank, comparing UPR and UR3. The class 0 denotes a correct calibration where an irrelevant document ranked by UPR is adjusted to a relevant one at the Top-1 rank. The class 1 denotes a incorrect calibration from relevant document to irrelevant one. The class 2 denotes the correct results on both methods, while class 3 representes the incorrect results on both models.

H Discussion on Hyperparameters

Here we display the detailed results about hyperparameter α analysis on different datasets of LLaMA2-7B model with nDCG@K metrics.

As depicted in Figure 9a and 9b, our analysis reveals that a hyperparameter setting of $\alpha = 0.25$ consistently yields robust enhancements compared to other evaluated values. While the highest observed NDCG@1 on the WebQ dataset exceeds 0.25, the overall performance metrics substantiate that 0.25 remains the optimal choice. Consequently, this hyperparameter configuration is adopted across all experimental evaluations.

I Performance on Paraphrased Query

To evaluate the retrieval outcomes on paraphrased queries, we utilized the GPT-3.5 Turbo model to paraphrase the queries from the WebQ dataset. The paraphrasing process was guided by the prompt, *"Please paraphrase the following question: question"*. Below are examples of the paraphrased questions we generated:

- Original Question: "What happened after Mr. Sugihara died?"-> Paraphrased: "What occurred following Mr. Sugihara's passing?"
- Original Question: "Where was George Washington Carver from?"-> Paraphrased: "What was George Washington Carver's place of origin?"

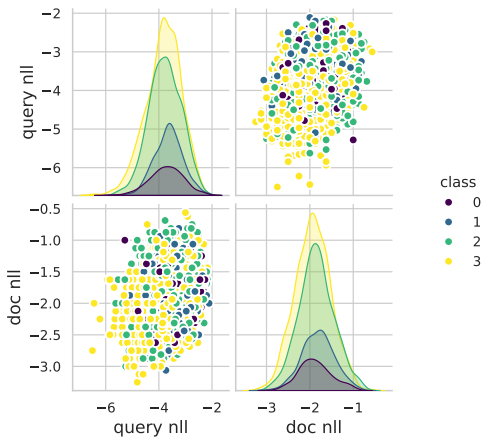
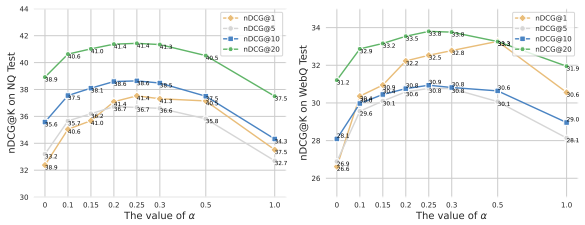


Figure 8: Complete results of Distribution on Top-1



(a) NQ (b) WebQ

Figure 9: Comparative effects of the hyperparameter on nDCG@K across different datasets. Evaluation is done on the Contriever retrieved documents.

- Original Question: "Who was Richard Nixon married to?"-> Paraphrased: "Who was the spouse of Richard Nixon?"

We then assessed the re-ranking results across various retrievers, including Contriever, BM25, MSS, and DPR, utilizing the paraphrased queries.

Llama2-7b	Contriever	UPR		UR^3	
Metric		w/o para.	w/ para.	w/o para.	w/ para.
top1	19.98	26.62	28.44	32.53	33.71
top5	43.45	54.92	56.35	58.71	59.30
top20	65.70	72.69	71.75(↓)	73.43	73.61
ndcg@1	19.98	26.62	28.44	32.53	33.71
ndcg@5	18.64	26.78	28.27	30.82	31.51
ndcg@20	22.22	31.18	32.24	33.79	34.61
MAP@100	18.79	25.92	27.12	27.82	28.67

Table 5: Performance comparison on Contriever

Llama2-7b	BM25	UPR		UR^3	
Metric		w/o para.	w/ para.	w/o para.	w/ para.
top1	18.90	27.56	28.25	33.91	34.10
top5	41.83	54.13	54.28	55.17	56.35
top20	62.40	68.5	69.05	69.54	69.98
ndcg@1	18.90	27.56	28.25	33.91	34.10
ndcg@5	19.36	27.39	28.26	30.72	31.34
ndcg@20	22.12	31.44	32.18	33.62	34.07
MAP@100	19.15	26.63	27.18	28.09	28.45

Table 6: Performance comparison on BM25

We observed that the queries paraphrased using ChatGPT generally resulted in marginal improvements in the reranking of retrieval results both on UPR and UR^3 for Contriever, BM25, and MSS. Our analysis suggests that this outcome may stem from the lower accuracy of these retrieval results to begin with. Furthermore, the distribution of queries generated by ChatGPT aligns more closely with the model data distribution than the empirical distribution of the original queries. This alignment potentially offers beneficial support for reranking initially poor retrieval outcomes.

However, for supervised retrieval systems like DPR, which already achieve high accuracy, the paraphrased queries deviate from the empirical data distribution, leading to greater negative impacts.

Llama2-7b	MSS	UPR		UR^3	
Metric		w/o para.	w/ para.	w/o para.	w/ para.
top1	11.66	26.38	26.03(↓)	29.38	29.72
top5	29.04	48.67	50.34	49.85	51.48
top20	49.21	63.19	62.80(↓)	62.40	62.80
ndcg@1	11.66	26.38	26.03	29.38	29.72
ndcg@5	11.57	26.67	27.47	28.21	29.04
ndcg@20	14.84	32.46	32.77	33.20	33.61
MAP@100	12.03	26.20	26.67	26.84	27.14

Table 7: Performance comparison on MSS

Llama2-7b	DPR	UPR		UR^3	
Metric		w/o para.	w/ para.	w/o para.	w/ para.
top1	44.83	39.32	37.89(↓)	42.18	42.86
top5	65.01	66.83	64.76(↓)	66.88	66.91
top20	74.61	76.67	76.53(↓)	76.96	76.93
ndcg@1	44.83	39.32	37.89(↓)	42.18	42.86
ndcg@5	39.76	38.66	37.95(↓)	40.34	40.97
ndcg@20	38.95	41.81	41.93	42.65	43.10
MAP@100	33.32	36.46	36.63	36.82	37.44

Table 8: Performance comparison on DPR

Consequently, the performance of the UPR method is noticeably affected.

Nonetheless, our approach involves a bias-corrected estimation, which, by leveraging the document probability values, mitigates the performance decline observed in DPR results and even achieves slight improvements.

J More Experiments on QA Task

We conduct inference on the re-ranking results of Mistral-7B in Table 11. The UR^3 method substantially enhances the performance of QA tasks, achieving superior EM and F1 scores compared to the UPR method on Mistral model. While DPR method has better performances on NQ and WebQ datasets for LLaMA2-13B and Gemma-7B, this trend is consistent with the analysis in Section 4.3.2.

Datasets	Metric	Contriever			BM25			MSS			DPR		
		Orig.	UPR	UR^3	Orig.	UPR	UR^3	Orig.	UPR	UR^3	Orig.	UPR	UR^3
NQ	Top-1	22.16	32.63	**38.61**	22.11	32.55	**37.89**	19.28	32.60	**37.04**	46.34	37.65	**43.30**
	Top-5	47.26	61.91	**64.82**	43.77	60.36	**63.24**	41.25	59.72	**61.75**	68.28	68.73	**72.27**
	Top-20	67.87	75.57	**76.76**	62.94	72.77	**73.52**	59.97	71.25	**71.61**	80.06	81.99	**82.77**
	nDCG@1	22.16	32.63	**38.61**	22.11	32.55	**37.89**	19.28	32.60	**37.04**	46.34	37.65	**43.30**
	nDCG@5	21.70	33.67	**38.21**	21.63	34.02	**38.03**	18.97	34.53	**37.90**	40.62	38.37	**43.21**
	nDCG@20	26.15	39.02	**42.66**	25.75	39.57	**42.34**	22.88	39.43	**41.81**	42.42	44.32	**47.85**
	MAP@100	20.71	31.65	**35.06**	20.78	32.36	**35.02**	18.11	32.41	**34.77**	34.89	36.34	**39.54**
WebQ	Top-1	19.98	28.44	**33.37**	18.90	29.08	**33.61**	11.66	27.31	**30.12**	**44.83**	39.03	43.06
	Top-5	43.45	56.25	**60.86**	41.83	54.13	**55.95**	29.04	49.56	**51.13**	65.01	66.58	**67.96**
	Top-20	65.70	72.39	**73.67**	62.40	68.80	**69.54**	49.21	**62.89**	62.50	74.61	76.57	**77.17**
	nDCG@1	19.98	28.44	**33.37**	18.90	29.08	**33.61**	11.66	27.31	**30.12**	**44.83**	39.03	43.06
	nDCG@5	18.64	27.88	**31.47**	19.36	28.26	**31.47**	11.57	27.36	**29.53**	39.76	39.14	**41.07**
	nDCG@20	22.22	31.70	**34.59**	22.12	32.01	**34.15**	14.84	32.97	**34.16**	38.95	42.16	**43.39**
	MAP@100	18.79	26.31	**28.49**	19.15	26.98	**28.59**	12.03	26.72	**28.08**	33.32	36.63	**37.53**
TriviaQA	Top-1	34.16	52.63	**56.07**	46.30	55.48	**58.24**	30.76	52.87	**55.00**	57.47	62.48	**63.99**
	Top-5	59.49	73.99	**74.75**	66.28	75.42	**75.89**	52.65	70.64	**71.05**	72.40	**79.08**	79.04
	Top-20	73.91	79.83	**80.25**	76.41	80.77	**80.87**	67.18	76.31	**76.34**	79.77	**83.13**	83.09
	nDCG@1	34.16	52.63	**56.07**	46.30	55.48	**58.24**	30.76	52.87	**55.00**	57.47	62.48	**63.99**
	nDCG@5	30.46	49.63	**51.78**	41.60	53.35	**55.17**	27.78	50.64	**52.04**	49.69	59.60	**60.31**
	nDCG@20	31.78	51.10	**52.61**	40.68	54.72	**55.88**	29.25	53.22	**54.10**	46.33	60.05	**60.27**
	MAP@100	26.61	44.86	**46.06**	34.85	49.36	**50.36**	24.02	47.12	**47.95**	39.40	54.25	**54.46**

Table 9: Re-ranking results on the test set of datasets of the Top-100 retrieved documents with the Mistral-7B model. The best results are highlighted in bold.

Datasets	Metric	Contriever			BM25			MSS			DPR		
		Orig.	UPR	UR^3	Orig.	UPR	UR^3	Orig.	UPR	UR^3	Orig.	UPR	UR^3
NQ	Top-1	22.16	29.86	**34.02**	22.11	29.83	**33.77**	19.28	30.78	**33.63**	**46.34**	36.48	40.64
	Top-5	47.29	57.45	**59.72**	43.77	56.34	**58.31**	41.25	56.34	**57.92**	68.28	66.90	**68.70**
	Top-20	67.87	74.16	**74.65**	62.94	71.63	**71.69**	59.97	69.86	**69.86**	80.06	81.16	**81.99**
	nDCG@1	22.16	29.86	**34.02**	22.11	29.83	**33.77**	19.28	30.78	**33.63**	**46.34**	36.48	40.64
	nDCG@5	21.70	30.93	**33.58**	21.63	31.32	**33.55**	18.97	32.06	**34.07**	**40.62**	37.34	39.87
	nDCG@20	26.15	35.97	**37.99**	25.75	36.65	**38.27**	22.88	36.99	**38.23**	42.42	42.43	**44.39**
	MAP@100	20.71	29.17	**30.92**	20.78	30.11	**31.48**	18.11	30.39	**31.49**	34.89	34.86	**36.43**
WebQ	Top-1	19.98	26.13	**28.40**	18.90	27.21	**29.82**	11.66	25.39	**28.00**	**44.83**	36.86	39.62
	Top-5	43.45	54.53	**57.33**	41.83	52.51	**52.95**	29.04	49.26	**50.49**	65.01	63.73	**65.70**
	Top-20	65.70	71.36	**72.74**	62.40	68.01	**68.26**	49.21	62.16	**62.20**	74.61	75.74	**76.23**
	nDCG@1	19.98	26.13	**28.40**	18.90	27.21	**29.82**	11.66	25.39	**28.00**	**44.83**	36.86	39.62
	nDCG@5	18.64	25.72	**28.28**	19.36	26.51	**28.58**	11.57	25.97	**27.71**	**39.76**	36.46	38.25
	nDCG@20	22.22	29.73	**31.70**	22.12	30.27	**31.68**	14.84	31.40	**32.33**	38.95	39.89	**40.78**
	MAP@100	18.79	24.62	**26.08**	19.15	25.69	**26.79**	12.03	25.18	**25.91**	33.32	34.87	**35.34**
TriviaQA	Top-1	34.16	51.22	**53.50**	46.30	53.81	**55.87**	30.76	50.85	**52.14**	57.47	59.52	**60.02**
	Top-5	59.49	71.74	**71.93**	66.28	74.05	**74.30**	52.65	69.00	**69.39**	72.4	76.93	**77.18**
	Top-20	73.91	**79.02**	78.98	76.41	**80.24**	80.08	67.18	75.48	**75.51**	79.77	82.44	**82.53**
	nDCG@1	34.16	51.22	**53.50**	46.30	53.81	**55.87**	30.76	50.85	**52.14**	57.47	59.52	**60.02**
	nDCG@5	30.46	47.26	**48.61**	41.60	51.36	**52.45**	27.78	48.40	**49.22**	49.69	**59.21**	56.57
	nDCG@20	31.78	48.02	**48.82**	40.68	52.37	**52.71**	29.25	50.54	**51.05**	46.33	56.52	**56.67**
	MAP@100	26.61	41.77	**42.77**	34.85	46.82	**46.86**	24.02	44.33	**44.80**	39.40	50.74	**50.88**

Table 10: Re-ranking results on the test set of datasets of the Top-100 retrieved documents with the GPT-Neo-2.7B model. The best results are highlighted in bold.

	NQ		WebQ		TriviaQA	
	EM	F1	EM	F1	EM	F1
LLaMA2-13B						
Contriever	22.02	29.11	19.69	30.21	49.90	57.08
+ Inference with UPR	28.56	36.81	21.80	33.18	59.06	67.23
+ Inference with UR^3	**29.00**	**37.39**	**22.54**	**34.47**	**59.43**	**67.69**
BM25	20.20	27.08	16.39	26.60	55.21	62.91
+ Inference with UPR	27.62	35.63	19.59	30.25	62.25	70.27
+ Inference with UR^3	**29.06**	**36.82**	**20.67**	**31.81**	**62.50**	**70.59**
MSS	19.86	26.16	16.83	27.94	49.29	56.03
+ Inference with UPR	26.70	34.61	20.77	31.58	57.94	65.94
+ Inference with UR^3	**27.95**	**35.75**	**21.80**	**32.99**	**58.23**	**66.20**
DPR	30.30	**38.42**	**22.79**	34.36	55.33	62.96
+ Inference with UPR	30.86	37.93	21.78	33.74	60.61	68.79
+ Inference with UR^3	**30.97**	38.38	22.75	**35.50**	**60.98**	**69.21**
Mistral-7B						
Contriever	20.69	26.61	14.37	24.39	49.89	56.94
+ Inference with UPR	25.29	32.30	16.78	26.67	59.35	67.01
+ Inference with UR^3	**25.93**	**32.98**	**17.42**	**27.77**	**59.32**	**67.13**
BM25	19.14	25.31	13.29	23.03	54.91	62.15
+ Inference with UPR	24.32	31.10	15.55	25.30	62.50	69.88
+ Inference with UR^3	**25.37**	**32.24**	**15.85**	**25.71**	**62.61**	**70.02**
MSS	18.20	24.28	13.98	23.79	48.41	55.22
+ Inference with UPR	24.04	30.88	16.04	26.24	58.09	65.76
+ Inference with UR^3	**24.27**	**31.03**	**16.54**	**26.64**	**58.11**	**65.86**
DPR	28.17	34.9	18.21	28.55	55.03	62.24
+ Inference with UPR	26.65	34.01	17.42	27.90	60.53	68.54
+ Inference with UR^3	**28.20**	**34.94**	**18.40**	**28.57**	**60.69**	**68.74**
Gemma-7B						
Contriever	17.40	25.13	14.71	26.05	45.54	53.66
+ Inference with UPR	22.11	30.70	15.00	26.95	55.78	64.67
+ Inference with UR^3	**23.02**	**31.27**	**15.60**	**27.04**	**55.89**	**64.96**
BM25	16.40	23.84	12.35	22.84	52.34	60.89
+ Inference with UPR	22.52	31.09	14.12	**25.22**	59.52	68.23
+ Inference with UR^3	**22.99**	**31.45**	**14.42**	25.05	**59.66**	**68.32**
MSS	14.38	21.44	12.20	22.91	43.56	51.54
+ Inference with UPR	20.94	29.40	14.35	**26.79**	**54.42**	63.08
+ Inference with UR^3	**22.11**	**30.22**	**14.76**	26.58	54.30	**63.28**
DPR	**24.43**	**33.14**	**16.68**	28.00	50.76	59.59
+ Inference with UPR	23.43	32.44	17.22	28.54	57.27	66.16
+ Inference with UR^3	24.10	33.07	16.44	**28.02**	**57.30**	**66.21**

Table 11: EM and F1 scores for the open-domain QA task. We perform inference with the re-ranked Top-1 results of Table 9. The best performing models are highlighted in bold. We highlight the best scores obtained by original retriever in red.

	NQ		WebQ		TriviaQA	
	EM	F1	EM	F1	EM	F1
Top-1						
Contriever	15.90	22.00	14.42	24.46	40.31	47.67
+ Inference with UPR	20.97	27.90	15.26	25.10	52.16	60.40
+ Inference with UR^3	**21.93**	**29.06**	**15.45**	**24.97**	**51.90**	60.34
BM25	15.65	21.38	12.55	21.55	48.41	56.60
+ Inference with UPR	20.75	27.71	14.47	24.59	55.68	64.22
+ Inference with UR^3	**21.75**	**28.91**	**15.70**	**25.27**	**56.73**	**65.13**
MSS	13.60	19.34	11.81	21.63	39.98	47.17
+ Inference with UPR	19.78	26.98	14.76	24.52	50.45	58.75
+ Inference with UR^3	**21.69**	**28.66**	**15.06**	**25.27**	**51.20**	**59.31**
DPR	23.38	30.69	16.20	26.28	46.86	54.87
+ Inference with UPR	22.13	29.58	15.26	25.18	53.84	62.17
+ Inference with UR^3	**24.29**	**31.16**	**17.03**	**27.13**	**54.08**	**62.40**
Top-3						
Contriever	14.93	20.36	12.16	22.03	40.23	49.53
+ Inference with UPR	19.31	25.51	13.44	22.65	49.74	59.38
+ Inference with UR^3	**19.98**	**25.77**	**14.03**	**23.23**	**50.08**	**59.66**
BM25	14.35	20.04	10.73	19.76	49.46	58.58
+ Inference with UPR	19.56	25.90	12.89	21.58	56.17	65.81
+ Inference with UR^3	**20.02**	**26.90**	**13.78**	**22.56**	**56.55**	**65.82**
MSS	13.63	19.48	12.45	21.53	40.36	49.44
+ Inference with UPR	18.70	24.96	13.14	22.82	48.93	58.33
+ Inference with UR^3	**19.47**	**26.20**	**13.93**	**23.01**	**49.21**	**58.51**
DPR	19.61	26.21	13.48	22.55	44.79	54.11
+ Inference with UPR	20.42	27.19	13.24	21.88	**51.55**	**61.24**
+ Inference with UR^3	**22.08**	**29.24**	**15.35**	**24.16**	51.54	61.23
Top-5						
Contriever	18.50	23.80	13.29	23.08	46.53	55.12
+ Inference with UPR	22.33	28.81	15.21	24.14	**55.43**	64.10
+ Inference with UR^3	**22.55**	**28.82**	**15.21**	**24.28**	55.36	**64.26**
BM25	16.45	22.05	12.01	20.69	54.96	63.46
+ Inference with UPR	21.91	28.28	13.88	22.84	60.83	69.51
+ Inference with UR^3	**22.27**	**28.70**	**14.57**	**23.27**	**60.90**	**69.51**
MSS	16.59	22.22	13.39	22.99	46.35	54.79
+ Inference with UPR	20.72	26.87	14.22	23.30	54.04	62.55
+ Inference with UR^3	**21.27**	**27.68**	**14.27**	**23.57**	**54.81**	**62.62**
DPR	22.60	28.84	13.63	22.21	50.61	58.98
+ Inference with UPR	23.74	30.21	15.65	24.25	56.53	65.22
+ Inference with UR^3	**24.54**	**30.96**	**16.34**	**24.72**	**56.63**	**65.29**

Table 12: EM and F1 scores for the open-domain QA task with different number of input documents on the LLaMA2-7B model. The best performing models are highlighted in bold.