DynRank: Improving Passage Retrieval with Dynamic Zero-Shot Prompting Based on Question Classification

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani,
Mohammed M. Abdelgwad, Adam Jatowt
University of Innsbruck
{abdelrahman.abdallah, jamshid.mozafari, bhawna.piryani,
mohammed.ali, adam.jatowt}@uibk.ac.at

Abstract

This paper presents DynRank, a novel framework for enhancing passage retrieval in open-domain question-answering systems through dynamic zero-shot question classification. Traditional approaches rely on static prompts and pre-defined templates, which may limit model adaptability across different questions and contexts. In contrast, DynRank introduces a dynamic prompting mechanism, leveraging a pre-trained question classification model that categorizes questions into fine-grained types. Based on these classifications, contextually relevant prompts are generated, enabling more effective passage retrieval. We integrate DynRank into existing retrieval frameworks and conduct extensive experiments on multiple QA benchmark datasets.

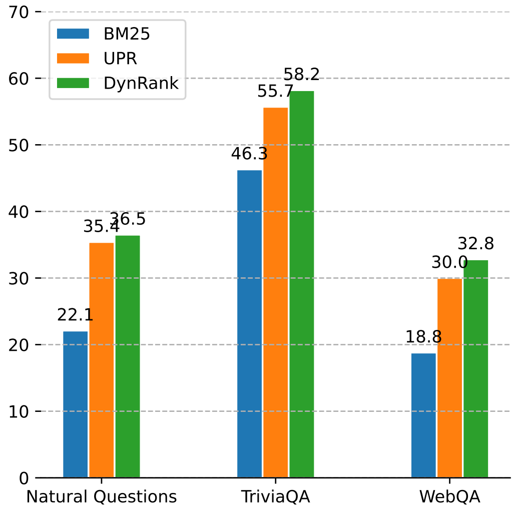

Figure 1: Top-1 Accuracy after re-ranking the top 1,000 passages retrieved by BM25 with DynRank comparing with UPR (Sachan et al., 2022) on the Natural Questions TriviaQA and WebQA datasets.

1 Introduction

Document retrieval plays a crucial role in many NLP tasks, particularly in open-domain question-answering (ODQA) systems (Gruber et al., 2024). In ODQA, a passage is first retrieved and then processed to answer a question. In these systems, the retriever component initially retrieves the most relevant passages from document resources like Wikipedia relevant to the posed question. Subsequently, the reader component examines the retrieved passages to detect an answer. This pipeline causes that the effectiveness of ODQA systems heavily relies on the quality and efficiency of the retriever. Recent advancements in NLP have focused on enhancing the performance of the retriever component by reranking the retrieved passages to improve the likelihood of retrieving the most relevant content (Pradeep et al., 2023a; Sachan et al., 2022; Sun et al., 2023; Zhuang et al., 2024; Pradeep et al., 2023b).

With the development of large language models (LLMs), many researchers focused on employing LLMs to rerank retrieved passages using diverse strategies and prompting techniques (Sun et al., 2023). Despite the success of LLM-based reranking methods such as UPR (Sachan et al., 2022) which generate a question from on each retrieved passage and then re-rank the passages based on how similar their generated questions are to the original question, one significant limitation remains. These models often rely on static and pre-defined instructions during reranking, which can reduce their adaptability across diverse queries and contexts. This static nature restricts the dynamic interaction between the query and retrieved passages, potentially leaving relevant information unused. While methods like RankGPT (Sun et al., 2023) have shown improved performance by generating permutations of passages based on relevance, they still follow a static instruction approach. RankGPT excels in zero-shot passage re-ranking tasks by em-

Proceedings of the 31st International Conference on Computational Linguistics, pages 4768–4778
January 19–24, 2025. ©2025 Association for Computational Linguistics

ploying a fixed set of instructions, but it is not dynamically adjusted based on the query context. In contrast, our approach aims to overcome this limitation by incorporating more dynamic interactions between the queries and retrieved passages. Similar to UPR (Sachan et al., 2022) it generates questions from passages to be used for passage reranking; yet this generation is conditioned by the inferred categories of target questions. A detailed discussion of related works is provided in Appendix B. Figure 1 illustrates the Top-1 accuracy achieved by DYNRANK, UPR, and BM25 after re-ranking the top 1,000 passages retrieved by BM25 across the Natural Questions, TriviaQA, and WebQA datasets.

In response to the above-mentioned challenges, we propose DynRank, a novel re-ranking framework designed to dynamically generate prompts tailored to each specific question. We leverage methods from the Question Classification (QC) task (Cortes et al., 2020), which categorizes questions into two groups: coarse-grained, indicating the major question type, and fine-grained, specifying the minor question type. Table 7 in Appenidx D presents the question types in detail. By incorporating a fine-grained question classification model, DynRank adapts to the context of each query, generating prompts that are more relevant and contextually appropriate. Our contributions are threefold:

1. We introduce a dynamic prompting mechanism that leverages question classification to tailor prompts based on the specific context of each query.

2. We integrate DynRank into existing retrieval framework and demonstrate its compatibility and ease of implementation.

3. We conduct extensive experiments on benchmark QA datasets, demonstrating that DynRank outperforms both traditional and state-of-the-art re-ranking methods.

2 Method

Figure 2 presents an overview of our approach for open-domain retrieval, introducing a novel dynamic zero-shot question generation method (DYNRANK) for re-ranking passages retrieved by any existing retriever.

2.1 Retriever

Let $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$ be a collection of evidence documents. Given a question $\boldsymbol{q}$, the retriever selects a subset of relevant passages $\mathcal{Z} \subset \mathcal{D}$, one or more of which ideally contains the answer to $\boldsymbol{q}$. Our method works with passages obtained from any retriever — either based on sparse representations like BM25 or dense representations like DPR. We assume that the retriever provides the K most relevant passages, denoted as $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$.

2.2 Question Classification

We fine-tune the RoBERTa model (Liu et al., 2019) on the UIUC dataset (Li and Roth, 2002) to build a question classifier. This fine-tuned model categorizes questions based on different levels of granularity, including 5 coarse-grained classes and 50 fine-grained classes. Given an input question $\boldsymbol{q}$, the question classification model assigns it to a major type l_{maj} and a minor subtype l_{min}. Formally, let Q be the set of all possible questions, and L_{maj} and L_{min} be the sets of major and minor types, respectively. The classification function $\mathcal{C} : Q \to L_{maj} \times L_{min}$ maps each question to its corresponding type and subtype:

$$(l_{maj}, l_{min}) = \mathcal{C}(\boldsymbol{q}) \tag{1}$$

2.3 Dynamic Prompt Generation

The dynamic prompt generator constructs a prompt p tailored to each question based on its classification result (l_{maj}, l_{min}). Let $\mathcal{T}$ be a template function that generates prompts. The prompt p is generated as follows:

$$p = \mathcal{T}(l_{maj}, l_{min}) \tag{2}$$

For instance, Given the major type l_{maj} as "human" and the minor type l_{min} as "individual," the generated prompt for the passage could be: *Document: [passage]. The above Document is about humans, specially on individuals, please write a question based on humans.*

2.4 Re-ranking with Pre-trained Language Models

Given the dynamically generated prompt p and a set of retrieved passages $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$ for the question $\boldsymbol{q}$, the goal of the re-ranking module is to reorder the passages such that the ones containing the correct answer are ranked higher. We define the relevance score s_i for each passage $\boldsymbol{z}_i$ as the

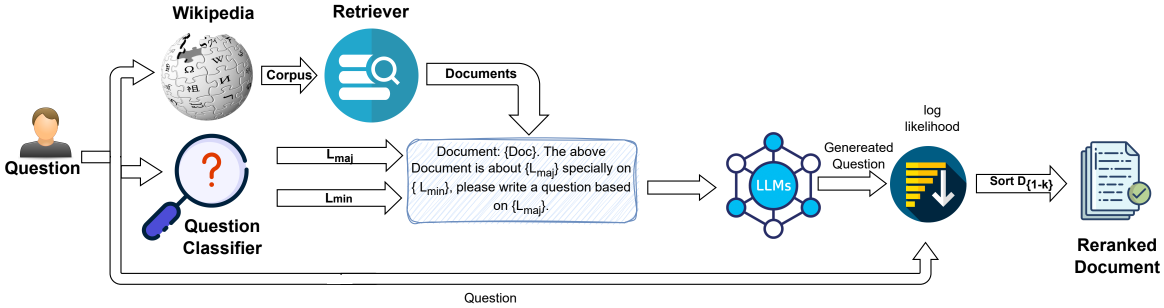

Figure 2: The architecture of the proposed DYNRANK framework. First, retriever retrieves relevant documents. The retrieved documents are then re-ranked based on a dynamically generated prompt, tailored to the question's classification result into major (L_{maj}) and minor (L_{min}) question types. A large language model (LLM) processes this prompt to re-rank the top-k documents, improving retrieval accuracy.

log-likelihood of generating the question $\boldsymbol{q}$ given $\boldsymbol{z}_i$ and p. We estimate the relevance score s_i as the conditional probability of the question $\boldsymbol{q}$ given the passage $\boldsymbol{z}_i$ and the prompt p using a pre-trained language model. The scoring function $\mathcal{S}$ is defined as:

$$s_i = \mathcal{S}(\boldsymbol{q} \mid \boldsymbol{z}_i, p) \tag{3}$$

The log-likelihood $\log P(\boldsymbol{q} \mid \boldsymbol{z}_i, p)$ is then computed as:

$$\log P(\boldsymbol{q} \mid \boldsymbol{z}_i, p) = \frac{1}{|\boldsymbol{q}|} \sum_{t=1}^{|\boldsymbol{q}|} \log P(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}_i, p; \Theta) \tag{4}$$

where q_t is the t-th token of the question $\boldsymbol{q}$, $\boldsymbol{q}_{<t}$ represents all tokens before q_t, and Θ denotes the parameters of the pre-trained language model. The passages are then re-ranked based on their computed scores s_i. The re-ranking function $\mathcal{R}$ sorts the passages by descending order of s_i:

$$\mathcal{R}(\mathcal{Z}) = \text{sort}(\mathcal{Z}, \text{by } s_i) \tag{5}$$

This approach ensures that the most relevant passages, as determined by the dynamically generated prompts, are ranked higher, thereby improving the accuracy of passage retrieval in open-domain question-answering systems.

3 Experiments

3.1 Datasets

Our experiments utilize several benchmark datasets including Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (WebQ) (Berant et al., 2013), the BEIR benchmark (Thakur et al., 2021), and the Question Classification dataset (Li and Roth, 2002). These datasets are crucial for evaluating the performance of question-answering systems across different types of queries and tasks. Table 1 summarizes the number of training, development, and test examples for the TriviaQA, NQ, and WebQ datasets. These datasets vary in size and complexity, providing diverse challenges for evaluating open-domain question answering (ODQA) models.

Dataset	Train	Dev	Test
TriviaQA	78,785	8,837	11,313
NQ	79,168	8,757	3,610
WebQ	3,417	361	2,032

Table 1: Statistics of the ODQA datasets: TriviaQA, NQ, and WebQ.

Natural Questions (NQ) The Natural Questions dataset contains 79,168 training examples, 8,757 development examples, and 3,610 test examples. It comprises questions derived from Google searches, paired with Wikipedia pages that provide the context necessary for answering the questions.

TriviaQA TriviaQA consists of 78,785 training examples, 8,837 development examples, and 11,313 test examples. It is made up of trivia questions sourced from trivia and quiz-league websites, annotated with evidence documents that support answers.

WebQuestions (WebQ) WebQuestions features 5,810 question-answer pairs, based on Freebase data, developed to facilitate research in question answering using structured knowledge sources.

Retriever	NQ				TriviaQA				WebQ			
	Top-1	Top-20	Top-100	Avg	Top-1	Top-20	Top-100	Avg	Top-10	Top-20	Top-100	Avg
Unsupervised Retrievers												
BM25	22.1	62.9	78.2	54.4	46.3	76.4	83.1	68.6	18.8	62.4	75.4	52.2
BM25 + UPR	35.4	78.4	85.2	66.3	55.7	82.8	86.4	74.9	30.0	72.8	81.6	61.4
BM25 + DynRank	**36.5**	**78.7**	**85.5**	**66.9**	**58.2**	**83.1**	**86.6**	**75.9**	**32.8**	**73.8**	**81.9**	**62.8**
Contriever	22.1	67.8	80.5	56.8	34.1	73.9	82.9	63.6	19.9	65.6	80.1	55.2
Contriever + UPR	36.3	79.6	86.6	67.5	56.7	82.8	86.4	75.3	30.0	74.6	82.9	62.5
Contriever + DynRank	**37.0**	**80.1**	**86.9**	**68.0**	**58.5**	**82.7**	**86.4**	**75.8**	**32.8**	**75.1**	**83.4**	**63.7**
Supervised Retrievers												
DPR	48.6	79.1	85.7	71.1	57.4	79.7	85.0	74.0	44.8	74.6	81.6	67.0
DPR + UPR	42.5	83.3	88.5	71.4	61.3	84.2	87.2	77.5	34.9	77.1	83.8	62.2
DPR + DynRank	**42.5**	**83.9**	**89.0**	**71.8**	**61.9**	**85.2**	**88.1**	**78.4**	**37.2**	**77.7**	**84.0**	**66.3**

Table 2: Top-{1, 20, 100} retrieval accuracy on the test set of datasets before and after re-ranking the top 1000 retrieved passages.

BEIR Benchmark The BEIR (Benchmark for Evaluating Information Retrieval) benchmark is a diverse set of retrieval tasks and domains that assess the generalization capabilities of retrieval systems. It includes several datasets covering various retrieval tasks such as fact-checking, question answering, and others.

Question Classification Dataset The Question Classification dataset is used for training models to categorize questions into coarse and fine-grained classes, enhancing the targeting of retrieval mechanisms. It consists of questions labeled according to their type, which helps in tuning retrieval systems to the specific needs of the question being asked.

3.2 Experimental Setup:

This section provides comprehensive details on the implementation of our experiments. We aim to ensure reproducibility and provide insights into the specific configurations used across different setups. We use BM25, Contriever, and DPR for initial retrieval. For BM25, the Pyserini toolkit is utilized, while for MSS and DPR, we employ implementations from Singh et al. (Singh et al., 2021). Contriever utilizes checkpoints from Huggingface (Wolf et al., 2020a). The top-1,000 passages retrieved are re-ranked according to our method, and results are reported for top-{1,20,100} retrieval accuracy.

For the BEIR benchmark, the top 100 passages retrieved using BM25 are re-ranked using our methodology. The nDCG@10 metric is used to assess performance.

Model	Accuracy	
	Coarse-grained	Fine-grained
Bert-base-cased	97.2	**91.8**
Bert-large-cased	98.0	83.8
DistilBert-base-cased	97.0	88.2
Albert-base-v2	96.0	87.8
Albert-large-v2	95.8	79.2
RobertA-base	97.2	90.6
RobertA-large	**97.8**	89.4

Table 3: The accuracy of question classifier based on the various models.

For Question Classification, we use the Transformers library (Wolf et al., 2020b) to fine-tune models. Training is conducted over 5 epochs with a batch size of 64 and a dropout rate of 0.1. A learning rate of 2e-5 is applied for base models, while 5e-6 is used for large models. This fine-tuning is essential for enhancing the model's ability to accurately classify questions into coarse and fine-grained categories.

All experiments are conducted using NVIDIA A40 GPUs on a high-performance computing cluster, ensuring significant computational resources.

4 Results

Question Classification We evaluated several models as the question classifier, including BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019). Table 3 shows that while RoBERTa excelled in coarse-grained classification, BERT-base performed optimally in fine-grained scenarios, underscoring the strengths of different models depending on the granularity of the classification task.

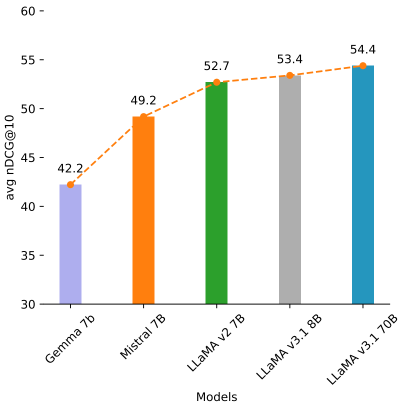

Figure 3: Average nDCG@10 scores for different language models on the BEIR benchmark.

Open-Domain QA As shown in Table 2, DynRank improves retrieval performance across various retrievers and datasets. When applied to BM25, DynRank increased top-20 accuracy by 3.3% on NQ and 3.6% on TriviaQA, and it achieved a 2.7% improvement on average across all datasets with Contriever. These results demonstrate DynRank's effectiveness in enhancing performance, even when working with unsupervised models. The supervised DPR model also benefited from DynRank, with a 4.4% boost in top-10 accuracy and a 4.8% increase in top-20 accuracy on the NQ dataset, along with a 4.5% improvement in top-1 accuracy on TriviaQA. Compared to UPR, a recent unsupervised re-ranking method, DynRank consistently outperformed it, including a 1.4% higher top-10 accuracy on WebQuestions. These gains highlight DynRank 's ability to dynamically tailor prompts, resulting in more effective retrieval than static approaches like UPR.

BEIR Benchmark Table 4 presents the nDCG@10 results for four BEIR datasets. DynRank shows substantial improvements over the BM25 baseline, achieving an average nDCG@10 score of 54.1 across the BEIR datasets, compared to 47.7 with BM25. Specifically, DynRank outperforms BM25 by 6.0% on the Covid dataset and by 4.5% on SciFact. Compared to other state-of-the-art methods such as RankGPT, DynRank achieves competitive results. While RankGPT slightly outperforms DynRank on Covid and SciFact, DynRank shows stronger performance on NFCorpus and Signal, highlighting its versatility and effectiveness across various domains.

Method	Covid	NFCorpus	SciFact	Signal	BEIR (Avg)
BM25	59.47	30.75	67.89	33.05	47.7
UPR (Sachan et al., 2022)	68.11	35.04	72.69	31.91	51.9
SGPT-CE (Muennighoff, 2022)	23.4	37.0	46.6	48.0	38.7
HyDE (Gao et al., 2022)	27.3	-	46.6	-	-
RankGPT (Sun et al., 2023)	76.67	35.62	70.43	32.12	53.7
DynRank	76.06	35.80	72.80	32.93	**54.39**

Table 4: Results (nDCG@10) on BEIR.

Impact of Different Language Models To assess the impact of different language models on retrieval performance, we evaluated several pre-trained models on the BEIR benchmark, specifically measuring nDCG@10 across different datasets. The models compared in this analysis include Gemma 7B, Mistral 7B, LLaMA v2 7B, LLaMA v3.1 8B, and LLaMA v3.1 70B. Fig 3 presents the average nDCG@10 scores for each model. Our findings show that LLaMA v3.1 70B achieved the highest performance with an nDCG@10 of 54.39, followed closely by LLaMA v3.1 8B with a score of 53.4. Both outperformed Mistral 7B and LLaMA v2 7B, which achieved scores of 49.18 and 52.7, respectively.

5 Conclusion

We introduced DynRank, a framework that improves open-domain question answering by generating tailored prompts using a question classification model. Experiments across multiple datasets demonstrate that DynRank outperforms static prompt methods, significantly enhancing retrieval performance and accuracy when integrated with both supervised and unsupervised retrievers.

6 Limitations

Despite the promising results, DynRank has several limitations that warrant further investigation:

- **Computational Complexity:** The dynamic generation of prompts, while beneficial for accuracy, introduces additional computational overhead.
- **Dependence on Pre-trained Models:** DynRank's performance is heavily dependent on the quality and size of the pre-trained language models used for generating the question based on the dynamic prompt.

References

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*.

Eduardo Cortes, Vinicius Woloszyn, Arne Binder, Tilo Himmelsbach, Dante Barone, and Sebastian Möller. 2020. An empirical comparison of question classification methods for question answering systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5408–5416, Marseille, France. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496*.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2024. Complextempqa: A large-scale dataset for complex temporal question answering. *arXiv preprint arXiv:2406.04866*.

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers)*.

Bernhard Kratzwald, Anna Eigenmann, and Stefan Feuerriegel. 2019. Rankqa: Neural question answering with answer re-ranking. *arXiv preprint arXiv:1906.03008*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Dingcheng Li, Jingyuan Zhang, and Ping Li. 2018. Representation learning for question classification via topic sparse autoencoder and entity embedding. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 126–133.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Yaqing Liu, Xiaokai Yi, Rong Chen, Zhengguo Zhai, and Jingxuan Gu. 2018. Feature extraction based on information gain and sequential pattern for english question classification. *IET Software*, 12(6):520–526.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.

Mudasir Mohd and Rana Hashmy. 2018. Question classification using a knowledge-based semantic kernel. In *Soft Computing: Theories and Applications*, pages 599–606, Singapore. Springer Singapore.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Dinesh Nagumothu, Bahadorreza Ofoghi, and Peter W Eklund. 2023. Semantic triple-assisted learning for question answering passage re-ranking. In *International Conference on Document Analysis and Recognition*, pages 249–264. Springer.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Harshith Padigela, Hamed Zamani, and W Bruce Croft. 2019. Investigating the successes and failures of bert for passage re-ranking. *arXiv preprint arXiv:1905.01758*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

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. *arXiv preprint arXiv:2204.07496*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv e-prints*, arXiv:1910.01108.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

Harish Tayyar Madabushi and Mark Lee. 2016. High accuracy rule-based question classification using question syntax and semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1220–1230, Osaka, Japan. The COLING 2016 Organizing Committee.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020a. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020b. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xuyang Wu, Zhiyuan Peng, Sravanthi Rajanala, Hsin-Tai Wu, and Yi Fang. 2024. Passage-specific prompt tuning for passage reranking in question answering with large language models. *arXiv preprint arXiv:2405.20654*.

Min Yang, Wei Zhao, Lei Chen, Qiang Qu, Zhou Zhao, and Ying Shen. 2019. Investigating the transferring capability of capsule networks for text classification. *Neural Networks*, 118:247–261.

Chunting Zhou, Chonglin Sun, Zhiyuan Liu, and Francis C. M. Lau. 2015. A C-LSTM Neural Network for Text Classification. *arXiv e-prints*, arXiv:1511.08630.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–47.

A Hyperparameters

For our experiments in open-domain question answering (ODQA) and BEIR retrieval tasks, we employed various large language models (LLMs) and configurations tailored to each dataset.

For the open-domain question answering tasks, we utilized the T0-3B model from Hugging Face, which is known for its ability to perform zero-shot learning based on prompts. Tokenized input passages were at 512 tokens and the generated question was 128 tokens. The top 1,000 passages retrieved were re-ranked for each query. For the BEIR benchmark, which involves a diverse set of retrieval tasks, we used larger models to accommodate the complexity of the retrieval tasks. The following models and settings were employed:

- **Models:** LLaMA v3.1 (70B, 8B), LLaMA v2 7B, Mistral 7B, and Gemma 7B.
- **Batch Size:** 16 per GPU.
- **Evaluation Metric:** The primary evaluation metric used was nDCG@10, a common metric for evaluating the relevance of ranked lists in retrieval tasks.

These configurations were chosen to balance retrieval performance, model size, and computational efficiency, ensuring that our approach was scalable across different datasets.

B Related Work

Passage re-ranking in question answering (QA) systems has been a focus of considerable research, as it significantly impacts the effectiveness of retrieving the most relevant information from a large corpus. Traditional approaches often rely on unsupervised methods like BM25 and dense retrievers such as DPR, which have laid the foundation for initial retrieval steps (Wu et al., 2024).

Recent advancements have shifted towards leveraging large pre-trained language models (LLMs) like BERT for re-ranking. Nogueira and Cho's work on BERT-based re-ranking demonstrated substantial improvements in passage retrieval accuracy by fine-tuning BERT on retrieval-specific tasks, showing that neural models could outperform traditional retrieval methods (Nogueira and Cho, 2019;

Padigela et al., 2019). This approach set a new benchmark, particularly in datasets such as MS MARCO and TREC-CAR, highlighting the effectiveness of transformer-based architectures in understanding query-passage relevance.

Question Classification is a task that requires determining the type of answer for a given question. Given the task's dependency on the language of the questions, the most effective methods leverage external linguistic resources (Tayyar Madabushi and Lee, 2016; Liu et al., 2018; Mohd and Hashmy, 2018). However, some methods have been developed to address this task without relying on the language of the questions (Yang et al., 2019; Zhou et al., 2015; Li et al., 2018). In this study, we adopt the first approach, as we focus on English, the most widely used language.

Further innovations include techniques like Passage-specific Prompt Tuning (PSPT), which fine-tunes a small number of parameters while keeping the core LLM parameters fixed. PSPT dynamically adjusts prompts based on individual passages, enhancing the model's adaptability across various contexts and improving re-ranking performance by incorporating passage-specific knowledge (Wu et al., 2024). This method stands out for its ability to tailor the retrieval process more closely to the nuances of the input data. Another significant contribution is the work on RankQA, which extends the conventional two-stage QA process with an additional re-ranking stage. This method integrates retrieval and comprehension features to re-rank answers, thereby improving the accuracy and robustness of QA systems. RankQA's simple yet effective design underscores the importance of combining multiple features for enhanced passage retrieval (Kratzwald et al., 2019). Moreover, generative models have also been explored for passage retrieval. These models, while computationally intensive, show promise by generating candidate passages that are then ranked for relevance. However, their high computational cost remains a challenge for practical deployment in large-scale systems (Nagumothu et al., 2023). Our proposed method, Dynamic Zero-Shot Question Generation for Re-ranking (DynRank), builds on these advancements by introducing dynamic prompt generation tailored to each question. By classifying questions into fine-grained categories and generating contextually relevant prompts, DynRank aims to enhance the adaptability and accuracy of passage retrieval, addressing some of the limitations of static prompt-based approaches.

C Case Study

The case study section presents a comparison between questions generated by two re-ranking methods: UPR and DynRank. The tables showcase how each model generates questions from passages first retrieved by BM25. The examples demonstrate clear differences between UPR and DynRank. UPR tends to generate more generic or passage-level questions, while DynRank generates more specific and contextually aligned questions. This highlights the advantage of the DynRank in producing more relevant and detailed queries, which are better suited for open-domain question-answering tasks. The tables help illustrate the practical impact of these models by showcasing the real output from the datasets (NQ and WebQA) used in the experiments.

D Major and Minor Categories

Table 7 presents the list of major and minor categories used for classification in our experiments. Each major category includes its associated minor types and their descriptions.

Question: who got the first nobel prize in physics?
Title: Nobel Prize in Physics.
Passage: receive a diploma, a medal and a document confirming the prize amount. Nobel Prize in Physics The Nobel Prize in Physics () is a yearly award given by the Royal Swedish Academy of Sciences for those who have made the most outstanding contributions for mankind in the field of physics. It is one of the five Nobel Prizes established by the will of Alfred Nobel in 1895 and awarded since 1901; the others being the Nobel Prize in Chemistry, Nobel Prize in Literature, Nobel Peace Prize, and Nobel Prize in Physiology or Medicine. The first Nobel Prize in Physics was. based on this passage.
UPR Question: What is the Nobel Prize in Physics?
DynRank Question: Who was the first to receive the Nobel Prize in Physics?

Question: when is the next deadpool movie being released?
Title: Deadpool (film).
Passage: Screen Rant called it possibly "the best film marketing campaign in the history of cinema". HostGator's Jeremy Jensen attributed the campaign's success to Reynolds, and to Fox for embracing the film's R rating. "Deadpool"s world premiere was held at the Grand Rex in Paris on February 8, 2016, before its initial theatrical release in Hong Kong the next day. This was followed by releases in 49 other markets over the next few days, including the United States on February 12. The movie was released in several formats, including IMAX, DLP, premium large formats, and D-Box. Kinberg explained that unlike the.
UPR Question: What was the first movie Deadpool was released in?
DynRank Question: When was the world premiere of the film "Deadpool" held?

Question: which mode is used for short wave broadcast service?
Title: Sender Zehlendorf.
Passage: Passage: of Russia, partly in the Simulcast mode. The long wave transmitter changed over on 29 August 2005 as first German large transmitter to Digital Radio Mondiale. The long wave transmitter ceased operation on December 31, 2014 as part of the general shutdown in Germany of AM radio services to the public. The mast continues to support VHF radio antennas providing FM broadcast services. Sender Zehlendorf Sender Zehlendorf is a radio transmission facility which has been in service since 1936, when a short wave transmitter was built in Zehlendorf (a village near Oranienburg) as part of the establishment of permanent radio.
UPR Question: What is the name of the transmitter?
DynRank Question: What digital broadcasting technique did the long wave transmitter at Sender Zehlendorf switch to on August 29, 2005?

Question: the south west wind blows across nigeria between?
Title: Oron people.
Passage: Passage: Civil War. Oron is found in the flood plain of South Eastern Nigeria, with the land mainly intersected by numerous streams and tributaries flowing into Cross River. The entire coastline stretches from Uya Oron to Udung Uko. Oron is in the tropical region and has a uniformly high temperature all the year round. The two main seasons are the dry which spans between October and April and wet season which starts around May and ends in September. There are also two prevailing winds – the South-West onshore winds which brings heavy rains and the North- East trade winds blowing across.
UPR Question: What is the name of the river that flows into Cross River?
DynRank Question: What are the main geographical and climatic features of the Oron region in South Eastern Nigeria?

Table 5: Examples of generated questions, Documents retrieved by BM25 for the Natural Questions (NQ) test dataset, comparing UPR and DynRank-based re-ranking.

Question: who plays ken barlow in coronation street?

Title: Ken Barlow.

Passage: Ken Barlow Kenneth Barlow is a fictional character from the British ITV soap opera C̈oronation Street,̈ played by William Roache. Ken was created by Tony Warren as one of C̈oronation Street'sör̈iginal characters. He debuted in the soap's first episode on 9 December 1960. Having appeared continuously since, December 2010 marked Ken's 50th anniversary on-screen, making Roache the longest-serving actor in a televised soap opera. Roache was honoured at the 2010 Guinness World Records ceremony for the achievement, when he surpassed actor Don Hastings from the American soap opera Äs the World Turns,̈ who previously held this title. Roache stated.

UPR Question: Who is the longest-serving actor in a televised soap opera?

DYNRANK Question: Who created the character Ken Barlow for "Coronation Street"?

Question: what does jamaican people speak?

Title: PhoneShop.

Passage: lives with his mother. He drives an Audi TT with numerous mechanical faults, and it is revealed that he paid "£3,000 for a £30,000 car." Jerwayne is Ashley's best friend as well as close co-worker, and the pair frequently egg each other on in foolish or unrealistic schemes and escapades. Jerwayne has an ongoing romantic interest in Janine, which is highlighted when Janine falls for a seemingly suave and successful Scotsman, provoking a jealous reaction from Jerwayne who vows to disrupt the relationship (which he eventually does with Ashley's help). Jerwayne speaks in a loud, Jamaican English dialect, is of.

UPR Question: What is the name of Ashley's best friend?

DYNRANK Question: What kind of car does Jerwayne drive?

Question: what did james k polk do before he was president?

Title: Washington C. Whitthorne.

Passage: Washington C. Whitthorne Washington Curran Whitthorne (April 19, 1825September 21, 1891) was a Tennessee attorney, Democratic politician, and an Adjutant General in the Confederate Army. Whitthorne was born near Petersburg, Tennessee in Marshall County. One day when Whitthorne was young James K. Polk stayed at his family's home. Polk saw how bright he was and asked, "What are you going to make of this boy?" His father replied "I am going to make him the President of the United States." Polk then told them to send the boy to Columbia and he would make him a lawyer. He attended Campbell.

UPR Question: What was Washington Curran Whitthorne's profession?

DYNRANK Question: What job did Washington C. Whitthorne have before becoming the Adjutant General in the Confederate Army?

Table 6: Examples of generated questions, Documents retrieved by BM25 for the WebQA test dataset, comparing UPR and DYNRANK-based re-ranking

Major Category	**Minor Category**	**Description**
ABBREVIATION (ABBR)	ABBR:abb (0)	Abbreviation
	ABBR:exp (1)	Expression abbreviated
ENTITY (ENTY)	ENTY:animal (2)	Animal
	ENTY:body (3)	Organ of body
	ENTY:color (4)	Color
	ENTY:cremat (5)	Invention, book, and other creative piece
	ENTY:currency (6)	Currency name
	ENTY:dismed (7)	Disease and medicine
	ENTY:event (8)	Event
	ENTY:food (9)	Food
	ENTY:instru (10)	Musical instrument
	ENTY:lang (11)	Language
	ENTY:letter (12)	Letter (a-z)
	ENTY:other (13)	Other entity
	ENTY:plant (14)	Plant
	ENTY:product (15)	Product
	ENTY:religion (16)	Religion
	ENTY:sport (17)	Sport
	ENTY:substance (18)	Element and substance
	ENTY:symbol (19)	Symbols and sign
	ENTY:techmeth (20)	Techniques and method
	ENTY:termeq (21)	Equivalent term
	ENTY:veh (22)	Vehicle
	ENTY:word (23)	Word with a special property
HUMAN (HUM)	HUM:gr (28)	Group or organization of persons
	HUM:ind (29)	Individual
	HUM:title (30)	Title of a person
	HUM:desc (31)	Description of a person
LOCATION (LOC)	LOC:city (32)	City
	LOC:country (33)	Country
	LOC:mount (34)	Mountain
	LOC:other (35)	Other location
	LOC:state (36)	State
NUMERIC (NUM)	NUM:code (37)	Postcode or other code
	NUM:count (38)	Number of something
	NUM:date (39)	Date
	NUM:dist (40)	Distance, linear measure
	NUM:money (41)	Price
	NUM:ord (42)	Order, rank
	NUM:other (43)	Other number
	NUM:period (44)	Duration or time period
	NUM:perc (45)	Percentage, fraction
	NUM:speed (46)	Speed
	NUM:temp (47)	Temperature
	NUM:volsize (48)	Size, area, or volume
	NUM:weight (49)	Weight

Table 7: Major and Minor Categories used for classification in our experiments.