
1 APPENDIX A: DATASETS INFORMATION

For training and evaluating our RAG system, we select eight datasets (five for training and three for evaluation). All the datasets we used are downloaded from the Hugging Face website¹.

1.1 FINE-TUNING DATASETS

We choose five datasets to fine-tune our two-stage architecture. Specifically, in the retrieval stage, we choose two Open-domain Question Answering (ODQA) datasets, *i.e.*, FreebaseQA Yao et al. (2014) and MS-MARCO Bajaj et al. (2016), to fine-tune our model for returning more relevant documents leveraging the powerful semantic understanding capability of LLM while avoiding the variance that happened in the practical situation.

- **FreebaseQA**²: This dataset is designed for open-domain factoid question answering using the Freebase knowledge graph. It contains 28,348 trivia-style questions and over 54,000 question-answer matches from TriviaQA and trivia website.
- **MS-MARCO**³: The MS-MARCO (Microsoft Machine Reading Comprehension) is a large-scale dataset for machine reading comprehension, featuring real-world queries from Bing users and over 1 million passages with query-answer pairs, supporting tasks like question answering and passage ranking.

By using both FreebaseQA and MS-MARCO to fine-tune our retriever, we improve the retrieval part of our Invar-RAG more comprehensively. FreebaseQA provides complex, linguistically diverse queries matched with structured knowledge graph data, enhancing our model’s ability to retrieve fact-based answers. MS-MARCO, with its real-world user queries and passage ranking, improves performance in understanding and ranking everyday information. Together, they enhance the retriever’s precision, relevance, and ability to handle a variety of natural language queries. To fine-tune our model for the generation task, we select three other datasets (two ODQA datasets and one reading comprehension dataset).

- **Wiki QA**⁴: This dataset, introduced by Microsoft, consists of question and sentence pairs collected from Bing search logs. It is designed for research in open-domain question answering. The dataset contains 3,047 questions linked to Wikipedia articles, with 29,258 sentences, 1,473 of which are labeled as correct answers to the questions.
- **Web Question**⁵: Released as part of research on semantic parsing, the WebQuestions dataset contains 5,810 question-answer pairs where the answers are entities from Freebase. The dataset is used to train and evaluate models for answering factual questions using structured data from Freebase.
- **SQuAD v2**⁶: The Stanford Question Answering Dataset (SQuAD) v2 builds upon the original SQuAD dataset by adding over 50,000 unanswerable questions that are similar to answerable ones, totaling 130,319 questions. This dataset challenges models to not only answer questions but also to determine when no answer is available in the context.

1.2 EVALUATION DATASETS

To evaluate both the retrieval and generation performance of Invar-RAG, we import three more ODQA datasets, *i.e.*, TriviaQA, Natural Question, and PopQA, which are all large-scale public datasets that have been widely used as benchmarks in the retrieval or generation tasks Fan et al. (2024).

¹<https://huggingface.co/datasets>

²https://huggingface.co/datasets/microsoft/freebase_qa

³https://huggingface.co/datasets/microsoft/ms_marco

⁴https://huggingface.co/datasets/microsoft/wiki_qa

⁵https://huggingface.co/datasets/Stanford/web_questions

⁶https://huggingface.co/datasets/rajpurkar/squad_v2

- **TriviaQA**⁷: TriviaQA is a reading comprehension dataset that comprises more than 650,000 question-answer-evidence triples. It includes 95,000 question-answer pairs generated by trivia experts, with independent evidence documents provided—around six per question on average—offering distant supervision of high quality for answering the questions.
- **Natural Question**⁸: The NQ dataset consists of questions sourced from real users, requiring QA systems to process and understand full Wikipedia articles that may or may not contain the answer. The use of actual user queries and the need to examine entire pages make this dataset more realistic and challenging compared to earlier QA datasets.
- **PopQA**⁹: PopQA is a large-scale, open-domain QA dataset with 14,000 entity-centric question-answer pairs. The questions are generated by applying templates to knowledge tuples retrieved from Wikidata.

2 APPENDIX B: EVALUATION

In this section, we will detail the information of our evaluation metrics and the implementation of our evaluation process. For retrieval, to provide a comprehensive view of performance, we leverage Accuracy both truncated at five and twenty, to assess the proportion of questions where the correct answers appear in the top 5 or top 20 retrieval results, respectively. For generations, we choose the Exact Match Wongsuphasawat et al. (2012) to evaluate the difference between our prediction and the ground truth. Moreover, in our evaluation process, we adopt 8-shot examples which are randomly selected from the chosen datasets following previous work Izacard & Grave (2020).

3 APPENDIX C: HYPERPARAMETER ILLUSTRATION

In this section, we will present the prime hyperparameter we used and the values we set in our fine-tuning and inferencing progress. Tab.1 shows the hyperparameter for fine-tuning our retrieval stage and generation stage.

Table 1: Hyperparameter for Fine-tuning (Retrieval and Generation)

Stage	lr_scheduler	batch size	seq len	LoRA_rank	LoRA_alpha	LoRA_dropout
Retrieval Stage	cosine	64	4096	16	32	0.05
Generation Stage	cosine	64	4096	16	32	0.05

4 APPENDIX D: BASELINES DESCRIPTION

(a) **BM25** Ram et al. (2023): A classical, sparse information retrieval method based on term frequency and inverse document frequency (TF-IDF). It ranks documents based on the occurrence of query terms, emphasizing exact term matches, making it effective in domains with well-structured data but limited in understanding semantic similarities. (b) **BGE** Izacard et al. (2021): A family of pre-trained embedding models designed for general Chinese text embedding. It achieves superior performance across diverse tasks by using massive datasets and advanced training techniques like contrastive learning, supporting retrieval, ranking, and classification. (c) **Contriever** Xiao et al. (2024): Contriever is an unsupervised dense information retriever based on contrastive learning. It is designed to overcome limitations in traditional sparse retrieval methods like BM25, particularly in zero-shot and multilingual settings. Contriever demonstrates state-of-the-art performance on multiple retrieval benchmarks, such as BEIR, and excels in cross-lingual retrieval. Its architecture relies on a bi-encoder model, which encodes queries and documents independently, and

⁷https://huggingface.co/datasets/mandarjoshi/trivia_qa

⁸https://huggingface.co/datasets/google-research-datasets/natural_questions

⁹<https://huggingface.co/datasets/akariasai/PopQA>

Table 2: Ablation Study on Natural Question and PopQA.

Model Variants	Nature Question			PopQA		
	Retrieval		Generation	Retrieval		Generation
	Acc@5	Acc@20	Exact Match	Acc@5	Acc@20	Exact Match
Default	80.5	88.0	56.2	73.5	82.5	53.6
w/o representation learning	69.0	80.5	54.7	59.5	70.0	51.7
w/o invariance loss	77.5	86.0	55.4	69.5	81.0	52.4
w/o generative fine-tuning	/	/	53.9	/	/	50.7

uses unsupervised contrastive learning to optimize retrieval, even without labeled data. (d) **LLM-embedder** Zhang et al. (2023): A unified embedding model optimized for retrieval augmentation in large language models (LLMs). It leverages multi-task fine-tuning and LLM feedback to retrieve relevant knowledge, tools, and memory, enhancing LLMs’ capabilities in knowledge-intensive tasks. (e) **RepLLaMA** Ma et al. (2024): RepLLaMA is a dense retriever fine-tuned from the LLaMA-2 language model, part of a multi-stage retrieval pipeline that includes rankLLaMA as a reranker. RepLLaMA improves the effectiveness of retrieval tasks, particularly in passage and document retrieval. It outperforms smaller models, showcasing strong generalization and effectiveness in both in-domain and zero-shot settings. The model leverages both hard negatives and in-batch negatives during training.

5 APPENDIX E: ABLATION STUDY

In this section, we will present the ablation study by gradually removing the significant components of our architecture. Specifically, we follow the variant settings in the main body and perform our ablation study on the other two evaluation ODQA datasets (Natural Question¹⁰ and PopQA¹¹), presented in Tab.2. From the experimental results, we can see that the performance of our Invar-RAG, no matter for the retrieval and the generation, shows a similar tendency to the one on TriviaQA, showing the effectiveness of our designed two-stage fine-tuning method, featuring the representation learning, invariance loss and the generation fine-tuning in one single LLM.

REFERENCES

- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.
- Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6491–6501, 2024.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.
- Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2421–2425, 2024.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

¹⁰https://huggingface.co/datasets/google-research-datasets/natural_questions

¹¹<https://huggingface.co/datasets/akariasai/PopQA>

162 Krist Wongsuphasawat, Catherine Plaisant, Meirav Taieb-Maimon, and Ben Shneiderman. Querying
163 event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting*
164 *with computers*, 24(2):55–68, 2012.

165 Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack:
166 Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM*
167 *SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024.

168 Xuchen Yao, Jonathan Berant, and Benjamin Van Durme. Freebase qa: Information extraction or
169 semantic parsing? In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, pp. 82–86,
170 2014.

171 Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to
172 augment large language models. *arXiv preprint arXiv:2310.07554*, 2023.

173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215