Retrieval-Augmented Generation for Enhancing Open-Domain Question Answering Systems with External Knowledge

Zhuo Chen
University of California, Berkeley
zhuochen@berkeley.edu

Abstract

This paper investigates the impact of retrieval-augmented generation on the performance of large language models (LLMs) in Open-Domain Question Answering (ODQA). We incorporate an external knowledge augmentation system leveraging the PAQ dataset, which enriches LLMs like Flan-T5 and Llama-2-13B-chat with real-time access to over 14 million question-answer pairs. By enabling dynamic retrieval of relevant information, we aim to overcome the inherent limitations of LLMs related to fixed knowledge bases and context length constraints. The study evaluates enhancements in accuracy, employing metrics such as Exact Match, F1, and Rouge scores. Our results demonstrate that retrieval-augmented techniques significantly improve the answer quality across various LLMs (e.g., $8.45 \rightarrow 35.48$ EM on Flan-T5-large), highlighting the potential of integrating these systems to extend the capabilities and applicability of traditional neural language models in complex question answering tasks.[1]

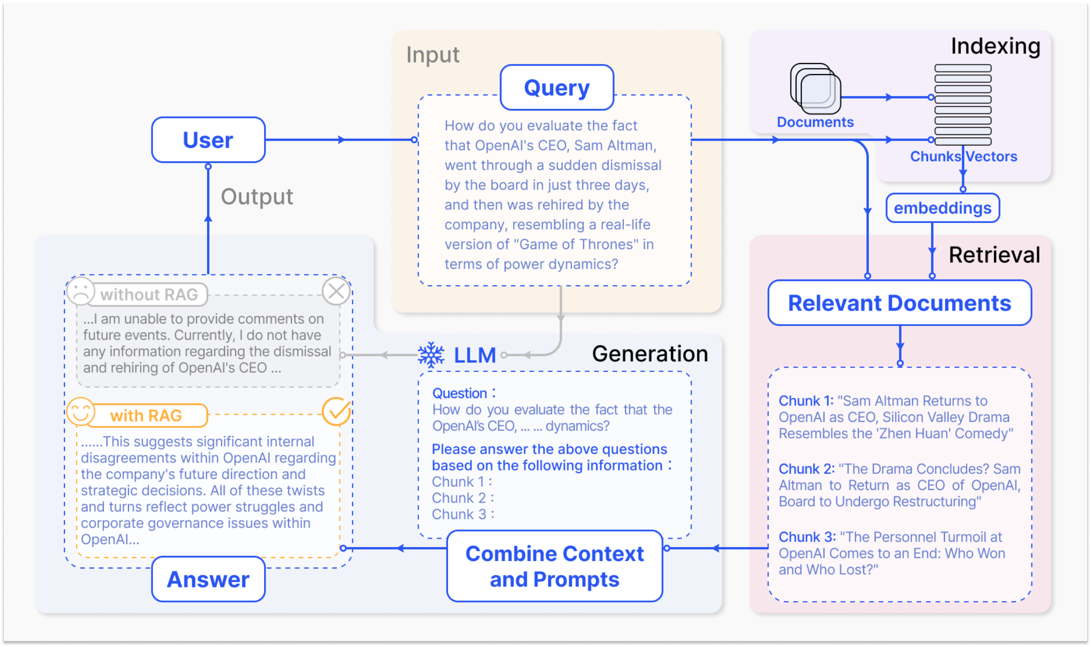
1 Introduction

The emergence of ChatGPT and the widespread use of transformer based large-language-models (LLMs) have significantly improved many areas of NLP tasks. LLMs can converse, answer questions, reason logically, and retrieve information from its parametric stored human knowledge. Despite its impressive capabilities, LLMs such as GPTs series have notable limitations. First, it is restricted by context length, with *GPT-3.5-Turbo-0613* supporting a 4096-token window (gpt3.5) and *GPT-4-0613* supporting an 8192-token window (gpt4). This limitation affects the model's ability to understand and remember longer text contexts. Second, LLMs are trained through pretraining and fine-tuning, which compresses knowledge into the model's parameters. Studies show that these models capture only partial knowledge (Petroni et al. (2019)), and while increasing the model size can improve coverage (Raffel et al., 2020; Roberts et al., 2020; Brown et al., 2020a), it requires more computational resources and time. Lastly, the models are limited by their knowledge cutoff, with both *GPT-3.5-Turbo-0613* and *GPT-4-0613*'s knowledge being up-to-date only until September 2021 (gpt3.5; gpt4). Updating the models to include the latest knowledge requires retraining and releasing newer versions, and the stored knowledge may not cover new or unseen data, especially in specific expert domains, leading to potential inaccuracies and hallucinations (Maynez et al., 2020).

In addition to encoding knowledge directly into model parameters, *retrieval-augmented models* (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Das et al., 2022), can access external sources like Wikipedia and retrieve relevant information and incorporate it into LLM's output generation process. Fig. 1 shows general workflows of *retrieval-augmented generations*. This approach has significantly improved the accuracy of answering open-domain questions, where the model must respond without specific context (Chen et al., 2017). Furthermore, RAG can be tailored to specialized fields by creating domain-specific knowledge embeddings. These embeddings are stored in vector databases and can be retrieved based on queries, enriching the model's context for more accurate responses.

To explore the potential of retrieval-augmented approaches to enhance the Open-Domain Question Answering (ODQA) task of large language models, this study devises an external knowledge augmentation system. This system incorporates approximately 14 million Question-Answer pairs sourced from the *PAQ* dataset (Lewis et al., 2021), a large collection of question-answering generated from Wikipedia, as the system's knowledge source. Uti-

[1] Code and datasets are available at `https://github.com/chenzhuo1005/datasci-w266-nlp-project`.



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question：
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?
Please answer the above questions based on the following information：
Chunk 1 :
Chunk 2 :
Chunk 3 :

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

Figure 1: Architecture of Retrieval-Augmented Generation (RAG) (Gao et al., 2024): factual knowledge is stored in a key-value memory where keys and values correspond to questions and answers, respectively; during inference, the model retrieves information from the memory via similarity search and uses it to condition the generation process.

lizing the *LlamaIndex* APIs, key-value embeddings for each Question-Answer pair were generated and stored within *Pinecone*, a managed vector database service renowned for vector similarity search. The evaluation process uses the *nq_open* (Kwiatkowski et al., 2019a; Lee et al., 2019) validation dataset, an open domain question answering benchmark that is derived from Natural Questions (Lee et al., 2019). For the inference tasks pertaining to question answering, this study employs a selection of large language models: *Flan-T5-base*, *Flan-T5-large*, *Llama-2-13B-chat*, and *GPT-3.5-Turbo*, comparing the standard versions of these models with versions enhanced by *PAQ* dataset knowledge. Performance is measured using *Exact Match (EM)*, *F1*, and *Rouge scores* to evaluate the effectiveness of knowledge augmentation.

2 Project Overview

2.1 Datasets

The dataset utilized in this study is the *nq_open* validation dataset, hosted on HuggingFace (https://huggingface.co/datasets/nq_open/viewer/nq_open/validation). This validation dataset, introduced by (Lee et al., 2019), comprises 3,610 question-answer pairs and serves as a benchmark for open-domain question answering, derived from *NaturalQuestions*. The objective is to predict an English answer string for a given English question.

2.2 Knowledge Source

The *PAQ* dataset (Lewis et al., 2021) serves as the knowledge source and is available on GitHub (https://github.com/facebookresearch/PAQ?tab=readme-ov-file#paq-qa-pairs). The QA pairs in PAQ were trained and retrieved from *NaturalQuestions* and *TriviaQA*, with the full version containing 64.9M QA pairs. For this project, We utilized *PAQ-L1*, a lighter version of *PAQ* with 14.1M QA pairs. Although *PAQ-L1* has more than four times fewer data rows, it still maintains substantial coverage of *NQ* (88.3%) and *TQA* (90.2%) compared to the full *PAQ* version, which covers *NQ* (90.2%) and *TQA* (91.1%) (Lewis et al., 2021). Since the *PAQ* data is derived from *NaturalQuestions*, it encompasses the coverage of the *nq_open* dataset mentioned earlier and is thus a suitable candidate for a knowledge source in knowledge augmentation generation.

2.3 Build Knowledge

The process of building knowledge is crucial for knowledge retrieval and augmentation. The aim is to transform the raw *PAQ-L1* data, which consists of question-answer text pairs, into key-value knowledge embeddings, using the question as the embed-

Natural Question (*PAQ*)	
Q: where is conan meriadoc from?	**A**: *british*
Retrieved Key-Values	
Q: who is conan meriadoc mentioned alongside in armes prydein?	**A**: cadwaladr
Q: who argued that conan meriadoc dates back to the mid 12th century?	**A**: hubert guillotelr
Q: conan meriadoc was the roman name for whom?	**A**: magnus maximus
Q: who is the compiler of conan meriadoc?	**A**: gurheden
Q: what nationality was conan meriadoc?	**A**: *british*

Table 1: Example question from *PAQ* and its associated answer, along with the top-5 question-answer pairs retrieved from the vector store by similarity search.

ding key. The choice of the embedding model is also significant, as it impacts the precision of knowledge compression and retrieval, as well as the speed of embedding computation. After thorough research and considering both accuracy and efficiency, *BAAI/bge-small-en* (Xiao et al., 2023) was chosen for its lower computational demand yet competitive performance. This model is based on the English language and has an embedding size of 384. Given that both the knowledge source and the evaluation dataset consist of brief questions and answers, an embedding size of 384 is deemed sufficient to encapsulate the relevant information.

Given the limited GPU computing power, a batch size of 500 for the embedding computation step has been determined. The embedded QA pairs are then stored in *Pinecone* vector database and are integrated into a node retrieval object implemented by *LlamaIndex* APIs. This retrieval object takes a question as the query text, computes its embedding using the same *BAAI/bge-small-en* embedding model, and performs a vector similarity search in *Pinecone*. The top-k embedded QA pairs are returned with their associated scores. Table 1 shows an example question and the top-5 retrieved answers from *PAQ*.

Last, the question associated with the retrieved knowledge QA pairs with scores will be passed to LLMs for generation inference.

2.4 Model Selection

Flan-T5-base: A variant of *T5-base* with the same number of parameters (220 million) but has been fine-tuned on over 1,000 additional tasks (Chung et al., 2022). It outperforms *T5-base* in many tasks, including question-answering. It has also been fine-tuned with instruction-based prompt texts, allowing it to utilize external knowledge to provide answers. Therefore, *Flan-T5-base* has been selected as the baseline model. HuggingFace: `https://huggingface.co/google/flan-t5-base`

Flan-T5-large: With an increased number of parameters (770 million), *Flan-T5-large* retains all the features of *Flan-T5-base* and represents a slight advancement over the baseline model. HuggingFace: `https://huggingface.co/google/flan-t5-large`

Llama-2-13B-chat-GGUF: Similar to *Llama-2-13B-chat-hf* except for its format, introduced by the *llama.cpp* team. Llama 2 consists of a series of pretrained and fine-tuned generative text models developed by Meta (Touvron et al., 2023). In this study, we use the 13 billion parameter version of Llama 2, optimized for dialogue applications. HuggingFace: `https://huggingface.co/TheBloke/Llama-2-13B-chat-GGUF`

GPT-3.5-Turbo: Provided by *OpenAI*, *GPT-3.5-Turbo* is considered state-of-the-art for many NLP tasks due to its highly adaptable prompt capability (Brown et al., 2020b). It is used as the most advanced LLM in this study. The current version, *GPT-3.5-Turbo-0125*, features a 16,385-token context window and training data up to September 2021. OpenAI: `https://platform.openai.com/docs/models/gpt-3-5-turbo`

3 Experiments

The main focus of the experiment was to evaluate the performance of LLMs by comparing the accuracy of answers generated with and without knowledge retrieval augmentation. Each LLM mentioned in the Model Selection section was tested in pairs. The first module simply generated text based on each question in the *nq_open* dataset. Given the different pre-training and fine-tuning of the LLMs, a specific prompt was chosen for each to conduct question-answering inference tasks (see Appendix A). The second module added a knowledge retrieval step by setting up a RAG module using *LlamaIndex* APIs. The input question was first embedded by the embedding model and then a vector search was conducted for similar embedded questions in the *Pinecone* vector database, which stores *PAQ* knowledge. The top-k most similar question-answer pairs, along with their similarity scores, were retrieved. The hyper-parameter **K** was set to 10 for this experiment. Finally, the embedded questions and retrieved knowledge question-answer

Natural Questions (*nq_open*)
Question: when was the last time anyone was on the moon? **Ground Truth**: ["14 December 1972 UTC", "December 1972"] **Flan-T5-base**: *1969 (False)* **Flan-T5-base+RAG**: *14 November 2016 (False)*
Question: who wrote he ain't heavy he's my brother lyrics? **Ground Truth**: ["Bobby Scott", "Bob Russell"] **Flan-T5-base**: *Bobby Mccartney (False)* **Flan-T5-base+RAG**: *Bobby Scott (True)*
Question: how many seasons of the bastard executioner are there? **Ground Truth**: ["one", "one season"] **Flan-T5-base**: *three (False)* **Flan-T5-base+RAG**: *two (False)*
Question: when did the eagles win last super bowl? **Ground Truth**: ["2017"] **Flan-T5-base**: *Super Bowl XXX (False)* **Flan-T5-base+RAG**: *2007 (False)*
Question: who won last year's ncaa women's basketball? **Ground Truth**: ["South Carolina"] **Flan-T5-base**: *North Carolina (False)* **Flan-T5-base+RAG**: *Tennessee (False)*

Table 2: Top 5 question-answer pairs from the *nq_open* validation dataset, associated with ground truth answers and answers generated by *Flan-T5-base* and *Flan-T5-base+RAG*.

pairs were fed into the LLMs to generate answers.

Evaluation Metrics *Exact Match (EM)* scores are the primary metric reported in this experiment, as they indicate a complete match with the generated answer. For questions in the validation dataset that have multiple answers (which is quite common, as there are often different ways to answer the same question), a generated answer is considered a successful exact match if it completely matches any of the answers in the list of multiple answers. In addition to *EM*, *F1 scores* and *Rouge scores* are also reported as supplementary metrics.

Baselines *Flan-T5-base* was selected as the baseline language model for this experiment due to its minimal number of parameters and simplest structure. The original *T5-base* model was not chosen because it failed to generate reasonable answers, with or without *RAG*, despite various prompt engineering trials. *Flan-T5-base*, fine-tuned with instructions, provided reasonable baseline metrics after experimentation and was therefore chosen. Table 2 shows top 5 questions from the validation dataset answered by the *Flan-T5-base* model, with and without *PAQ* knowledge.

Model	**EM**	**F1**	**Rouge1**	**RougeL**
Flan-T5-base	4.52	7.84	5.78	5.78
Flan-T5-base+RAG	**27.59**	**32.27**	**24.65**	**24.61**
Flan-T5-large	8.45	12.63	9.72	9.72
Flan-T5-large+RAG	**35.48**	**39.95**	**30.48**	**30.5**
Llama-2-13B-chat-GGUF	35.96	40.16	30.4	30.22
Llama-2-13B-chat-GGUF+RAG	**37.62**	**40.73**	**30.97**	**30.95**
GPT-3.5-Turbo-0125	34.07	39.02	30.99	30.82
GPT-3.5-Turbo-0125+RAG	**37.89**	**42.64**	**33.01**	**32.91**

Table 3: *Exact Match (EM)*, *F1* and *Rouge Scores* results in comparison of models: *Flan-T5-base*, *Flan-T5-large*, *Llama-2-13B-chat-GGUF* and *GPT-3.5-Turbo* with/without RAG enhancement.

4 Results

Table 3 displays the experimental results on the *nq_open* validation datasets. The *EM*, *F1*, *Rouge1*, and *RougeL* scores are reported for each large language model along with its paired RAG-enhanced version. The baseline model, *Flan-T5-base* with RAG enhancement, reported an *EM* of 27.59, which represents a substantial increase of 23.07 percentage points compared to its plain version without knowledge augmentation. The *F1* increased by 24.43 percentage points (7.84 → 32.27), and the increases for *Rouge1* and *RougeL* were 18.87 (5.78 → 24.65) and 18.83 (5.78 → 24.61), respectively.

The next model, *Flan-T5-large*, initially reported higher metrics compared to *Flan-T5-base* without RAG, with *EM*, *F1*, *Rouge1*, and *RougeL* at 8.45, 12.63, 9.72, and 9.72, respectively. This improvement can be attributed to the larger model size (770M vs 220M parameters), which allows for better understanding of the prompted questions and stores more parameterized knowledge. With RAG enhancement, *Flan-T5-large* also demonstrated substantial improvement, with increases of *EM* 27.03 (8.45 → 35.48), *F1* 27.32 (12.63 → 39.95), *Rouge1* 20.76 (9.72 → 30.48), and *RougeL* 20.78 (9.72 → 30.5) percentage points, respectively.

Llama-2-13B-chat-GGUF, a 13-billion parameter variant of the Llama 2 model, exhibited strong performance in answering *nq_open* questions without external knowledge, presenting *EM*, *F1*, *Rouge1*, and *RougeL* scores of 35.96, 40.16, 30.4, and 30.22, respectively. With RAG enhancement, *Llama-2-13B-chat-GGUF* showed slight improvements, with increases in *EM*, *F1*, *Rouge1*, and *RougeL* of 1.66 (35.96 → 37.62), 0.57 (40.16 → 40.73), 0.57 (30.4 → 30.97), and 0.73 (30.22 →

30.95) pencentage points, respectively. These results indicate that the knowledge embedded within *Llama-2-13B-chat-GGUF* model parameters is robust, enabling effective responses to ODQA questions. Additionally, while RAG enhancement still aids accuracy, its impact is not as pronounced as in smaller models like *Flan-T5-base* and *Flan-T5-large*.

GPT-3.5-Turbo showed similar performance to *Llama-2-13B-chat-GGUF*, with increases in *EM*, *F1*, *Rouge1*, and *RougeL* of 3.82 (34.07 → 37.89), 3.62 (39.02 → 42.64), 2.02 (30.99 → 33.01), and 2.09 (30.82 → 32.91). With RAG knowledge enhancement, the metric improvements were slightly higher compared to those generated by *Llama-2-13B-chat-GGUF*. This difference can be attributed to the fact that our ground truth answers in *nq_open* have different training time cutoffs compared to *GPT-3.5-Turbo*. For instance, questions such as `‘who won last year’s NCAA women’s basketball?’` heavily depend on the model’s knowledge cutoff time. Notably, *GPT-3.5-Turbo* recorded the highest scores for *Rouge1* and *RougeL*, indicating the fluency of its responses. However, the complexity of the answers it generated might lead to them being marked as false instances in *Exact Match*, causing it to slightly underperform compared to *Llama-2-13B-chat-GGUF*. A better-designed instruction prompt should help increase the alignment of *GPT-3.5-Turbo*’s answer generations with the format of *nq_open* answers.

Table 4 provides additional examples of answers generated by different LLMs with knowledge augmentation.

5 Conclusions

This study has systematically explored the enhancement of Open-Domain Question Answering (ODQA) capabilities through the integration of retrieval-augmented generation with large language models (LLMs). The deployment of an external knowledge augmentation system utilizing approximately 14 million Question-Answer pairs from the *PAQ* dataset has demonstrated significant advancements in model performance across various metrics.

Our findings reveal that the augmentation of traditional LLMs like *Flan-T5-base* and *Flan-T5-large* with Retrieval-Augmented Generation (RAG) not only substantially improves their *Exact Match (EM)*, *F1*, and *Rouge scores* but also addresses inherent limitations due to the static nature of their trained parameters. For instance, *Flan-T5-base*, enhanced with RAG, exhibited a remarkable improvement, with *EM scores* increasing by over 23 percentage points, highlighting the effectiveness of leveraging external, dynamically retrieved knowledge.

Moreover, the comparison between models of different capacities, from the 220 million parameter *Flan-T5-base* to the 13 billion parameter *Llama-2-13B-chat-GGUF*, has underscored the scale’s impact on knowledge comprehension and retrieval capabilities. Interestingly, while larger models inherently performed better in baseline settings, the marginal gains provided by RAG were proportionally greater in smaller models, suggesting a significant enhancement in their ability to contextualize and reason with external information.

The *Llama-2-13B-chat-GGUF* model, with its vast parameter count, initially displayed robust performance without external aids. However, the application of RAG still offered slight improvements, emphasizing that even the most powerful models can benefit from access to expanded knowledge bases. This is particularly evident in the nuanced domain of question answering, where contextual relevance and up-to-date information are crucial.

Furthermore, the nuanced differences in performance increments between models also point to the varying effects of knowledge augmentation depending on the model architecture and initial training data. For example, *GPT-3.5-Turbo*, despite its sophisticated capabilities, showcased a unique dependency on the RAG for optimizing its outputs, particularly when faced with questions tied closely to recent data.

In conclusion, this study not only reinforces the utility of integrating retrieval-augmented techniques into existing LLM frameworks but also highlights the critical need for continuous adaptation and enhancement of these models to maintain relevance over time. As the field of NLP progresses, strategies that merge static knowledge with dynamic retrieval mechanisms will be pivotal in addressing the evolving complexities of language understanding and generation. The promising results from various configurations of model and augmentation setups offer a clear pathway for future research focused on refining these integrations, potentially setting new benchmarks in the ODQA landscape.

6 Limitations

Due to limited experimental time and restricted access to computational resources, this study primarily utilized Google Colab Pro with a T4 GPU. This infrastructure offers limited CPU and GPU memory and computational speed compared to more advanced systems such as the A100 or T100 GPUs. This limitation affects the choice of embedding models, as larger embedding sizes—such as OpenAI's *text-embedding-3-small* (embedding size 1536) and *text-embedding-3-large* (embedding size 3072) (openai-embed)—require significantly higher computing power and storage capacity in vector databases. However, they provide more accurate embedding searches based on similarity. This is crucial for Retrieval-Augmented Generation (RAG), as the model must retrieve the most relevant knowledge from the *PAQ* dataset to accurately answer questions.

Another limitation concerns the quality of the *PAQ* dataset knowledge, which may not cover all questions in the *nq_open* dataset and could potentially mislead LLMs if the retrieved answers are incorrect. An example of this issue is demonstrated in the appendix (Table 4) under the third question: `"How many seasons of The Bastard Executioner are there?"` The correct answer is `"one"` or `"one season"`. Initially, *Llama-2-13B-chat* and *GPT-3.5-Turbo* answered this question correctly using only their internally parameterized stored knowledge. However, the answers retrieved from *PAQ* suggested that there were `"two"` seasons, leading the LLMs to incorrectly answer this question when using RAG. This highlights the significance of the knowledge embedding step in the retrieval augmentation process, and how poorly embedded knowledge can occasionally degrade the performance of LLMs.

7 Future Works

Exploring Advanced Embedding Models To improve the effectiveness of our knowledge-building process, exploring various embedding models could be beneficial. For instance, evaluating the performance of models such as *sentence-transformers/paraphrase-MiniLM-L6-v2* (Reimers and Gurevych, 2019) or *openai/text-embedding-3-small* (openai-embed) may enhance the quality of knowledge retrieval. Utilizing more sophisticated embedding techniques can potentially lead to significant improvements in how information is indexed and accessed.

Diversifying Data Sources We could also enhance the diversity of our data sources or combine multiple datasets to enrich our knowledge base. Potential resources include NaturalQuestions (NQ, Kwiatkowski et al., 2019b), TriviaQA (TQA, Joshi et al., 2017), and WebQuestions (Berant et al., 2013). By integrating various datasets, we can create a more robust and comprehensive repository of information, which may improve the performance and versatility of our retrieval systems.

Optimizing Data Precision Currently, our knowledge embeddings are generated with 32-bit precision. However, studies indicate that using 8-bit embeddings could maintain performance levels while reducing the 75% GPU memory and knowledge database size by 4 times (Dettmers et al., 2023b,a, 2022, 2021). This reduction would be highly beneficial for the RAG framework, particularly as it addresses one of its key challenges: managing and accessing large volumes of data. Additionally, this approach would help in minimizing the computational resources required.

Rethinking Retrieval Mechanisms Finally, the conventional method in RAG for retrieving knowledge involves searching for the top-k highest scored embedding vectors based on cosine similarity, MIPS implementations such as `faiss` (Johnson et al., 2019) enable searching across millions of vectors in milliseconds on a CPU. However, the constructed knowledge base might exhibit dependencies, and answering complex questions may require integrating multiple knowledge segments. Rather than relying solely on independent vector similarity, incorporating conditional probability or exploring sequence-to-sequence searches and attention mechanisms (Vaswani et al., 2023) in the knowledge retrieval process could provide a more nuanced and context-aware approach. This adjustment might significantly enhance the capability of the RAG system to generate more accurate and contextually relevant responses.

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*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Rajarshi Das, Patrick Lewis, Sewon Min, June Thai, and Manzil Zaheer, editors. 2022. *Proceedings of the 1st Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge*. Association for Computational Linguistics, Dublin, Ireland and Online.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023a. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2023b. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

gpt3.5. Gpt-3.5 turbo model. `https://platform.openai.com/docs/models/gpt-3-5-turbo`. Accessed: 2023-04-12.

gpt4. Gpt-4 and gpt-4 turbo models. `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`. Accessed: 2023-04-12.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. 2019b. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

openai-embed. Guide to openai embeddings. `https://platform.openai.com/docs/guides/embeddings`. Accessed: 2023-04-12.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

A Prompt Engineering

Flan-T5-base, Flan-T5-large:

```
Answer the following question:
<question>
```

Flan-T5-base+RAG, Flan-T5-large+RAG:

```
Given the below context:

The context is a list of questions and
answers pairs that could be used to
answer the questions.

"""
  <question>:<answer>
  <question>:<answer>
  <question>:<answer>
  <question>:<answer>
  <question>:<answer>
  <question>:<answer>
  <question>:<answer>
  ...
"""

Answer the following question:
<question>
```

Llama-2-13B-chat, GPT-3.5-Turbo:

```
[INST]
«SYS»
This is a Open Domain Question Answering
Task. You are given a question and you
need to generate the short answer based
on the question. Below are a few examples:

Question:  where did they film hot
tub time machine?
Answer: Fernie Alpine Resort
Question: who has the right of way in
international waters?
Answer: Neither vessel
Question: who does annie work for attack
on titan?
Answer: Marley
Question: when was the immigration reform
and control act passed?
Answer: November 6, 1986
[ENDSYS]
«/SYS»
Question: <prompt>
[/INST]
```

Llama-2-13B-chat+RAG:
GPT-3.5-Turbo+RAG:
Not applicable, since LlamaIndex API is used and the corresponding retrieved answers and scores will be passed to its API for generation.

Natural Questions (*nq_open*)	
Question	when was the last time anyone was on the moon?
Answer	[*"14 December 1972 UTC", "December 1972"*]
Flan-T5-base:	*1969 (False)*
Flan-T5-base+RAG:	*14 November 2016 (False)*
Flan-T5-large:	*December 1972 (True)*
Flan-T5-large+RAG:	*Apollo 17 (False)*
Llama-2-13B-chat-GGUF:	*December 1972 (True)*
Llama-2-13B-chat-GGUF+RAG:	*2017 (False)*
GPT-3.5-Turbo-0125:	*The last time anyone was on the moon (False)*
GPT-3.5-Turbo-0125+RAG:	*The last time anyone was on the moon was during (False)*
Question	who wrote he ain't heavy he's my brother lyrics ?
Answer	[*"Bobby Scott", "Bob Russell"*]
Flan-T5-base:	*Bobby Mccartney (False)*
Flan-T5-base+RAG:	*Bobby Scott (True)*
Flan-T5-large:	*John Lennon (False)*
Flan-T5-large+RAG:	*Bobby Scott (True)*
Llama-2-13B-chat-GGUF:	*Bobby Scott (True)*
Llama-2-13B-chat-GGUF+RAG:	*Bobby Scott (True)*
GPT-3.5-Turbo-0125:	*Bobby Scott (True)*
GPT-3.5-Turbo-0125+RAG:	*Bobby Scott (True)*
Question	how many seasons of the bastard executioner are there?
Answer	[*"one", "one season"*]
Flan-T5-base:	*three (False)*
Flan-T5-base+RAG:	*two (False)*
Flan-T5-large:	*6 (False)*
Flan-T5-large+RAG:	*26 (False)*
Llama-2-13B-chat-GGUF:	*one (True*
Llama-2-13B-chat-GGUF+RAG:	*2 (False)*
GPT-3.5-Turbo-0125:	*one (True)*
GPT-3.5-Turbo-0125+RAG:	*two (False)*
Question	when did the eagles win last super bowl?
Answer	[*"2017"*]
Flan-T5-base:	*Super Bowl XXX (False)*
Flan-T5-base+RAG:	*2007 (False)*
Flan-T5-large:	*2017 (True)*
Flan-T5-large+RAG:	*2017 (True)*
Llama-2-13B-chat-GGUF:	*The Eagles won their last Super Bowl in 2018 (False)*
Llama-2-13B-chat-GGUF+RAG:	*2017 (True)*
GPT-3.5-Turbo-0125:	*The Philadelphia Eagles won their last Super Bowl (False)*
GPT-3.5-Turbo-0125+RAG:	*2017 (True)*
Question	who won last year's ncaa women's basketball?
Answer	[*"South Carolina"*]
Flan-T5-base:	*North Carolina (False)*
Flan-T5-base+RAG:	*Tennessee (False)*
Flan-T5-large:	*Michigan State Spartans (False)*
Flan-T5-large+RAG:	*South Carolina (True)*
Llama-2-13B-chat-GGUF:	*Baylor Lady Bears (False)*
Llama-2-13B-chat-GGUF+RAG:	*South Carolina (True)*
GPT-3.5-Turbo-0125:	*Stanford University (False)*
GPT-3.5-Turbo-0125+RAG:	*South Carolina (True)*

Table 4: More examples of LLMs' prediction on *nq_open*.