Entity Linking for Retrieval-Augmented Factoid Question Answering with Large Language Models

Anonymous ACL submission

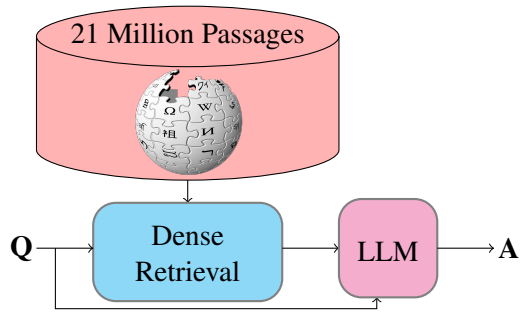
Abstract

Factoid questions seek real world factual knowledge. We explore the application of retrieval augmented question answering with large language models in the context of factoid questions. Our novel approach, *Entity Retrieval*, employs entity linking to identify relevant documents, offering an alternative to dense retrieval techniques. Our findings establish that retrieval-augmented methods are particularly effective for smaller large language models, and that *Entity Retrieval* outperforms other retrieval methods while demanding reduced time and resources[1].

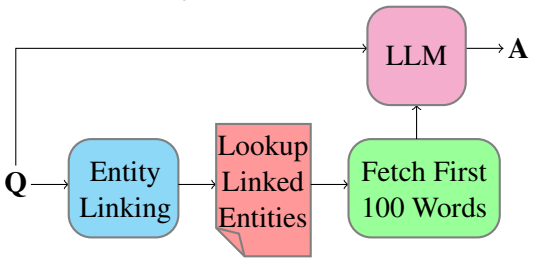
1 Introduction

Factoid questions seek factual information about the real world, and typically have answers that are concise single words or short phrases. These answers often reference or directly stem from a knowledge base entity (Ranjan and Balabantaray, 2016). The literature on factoid question answering is replete with a variety of well-studied methodologies, including rule-based (Shitu et al., 2020), pattern-based (Pala Er and Cicekli, 2013), and neural network-based approaches (Lukovnikov et al., 2017; Mohammed et al., 2018; Lukovnikov et al., 2019).

In recent years, Large Language Models (LLMs; OpenAI, 2023; Touvron et al., 2023; Jiang et al., 2024) have significantly transformed the field of natural language processing. Retrieval-augmented generation (RAG; Lewis et al., 2020b; Izacard and Grave, 2021; Singh et al., 2021) has emerged as a prevalent approach to question answering with LLMs. RAG systems typically utilize the retriever-reader architecture (Chen et al., 2017), where the retriever can be sparse (Peng et al., 2023), dense (Karpukhin et al., 2020), or a hybrid of the two



(a) Retrieval-Augmented QA with Dense Retrieval



(b) Retrieval-Augmented QA with *Entity Retrieval*

Figure 1: *Entity Retrieval* simplifies the process of obtaining augmentation documents by replacing the need to search through large indexed passages with a straightforward lookup.

(Glass et al., 2022). The reader, a generative language model (e.g., BART; Lewis et al., 2020a, T5; Raffel et al., 2020, GPT-3; Brown et al., 2020), then conditions the generated output based on the documents retrieved by the retriever. Recent RAG methodologies leverage the in-context learning capabilities of LLMs to integrate the retrieved documents into the prompt (Shi et al., 2023; Peng et al., 2023; Yu et al., 2023).

Following the prevalent belief that factoid question answering is nearing resolution (Petrochuk and Zettlemoyer, 2018), there has been a shift in the community towards reading comprehension (Joshi et al., 2017), multi-hop (Yang et al., 2018), and commonsense (Talmor et al., 2019) question answering. This shift has relegated factoid question answering to a less central position. Consequently,

[1]We have included our source code implementation of the project, along with the generated model answers, in the Software section of our ARR submission.

a reliable evaluation of the performance of common retrieval-augmented question answering methods for factoid question types is currently lacking.

In this paper, we examine the suitability of retrieval-augmented question answering methods for factoid questions. In addition, following the established methodologies in statistical factoid question answering systems (Aghaebrahimian and Jurčíček, 2016; Li et al., 2021; Adebisi et al., 2022), we study the role of entity linking as a critical component in retrieval-augmented factoid question answering with LLMs. Our contributions in this paper are as follows:

(1) We examine the performance of retrieval- and non-retrieval-augmented LLMs for answering factoid questions.

(2) We propose *Entity Retrieval*, a simple yet effective approach to employ entity linking in retrieval-augmented factoid question answering with LLMs.

Figure 1 presents a schematic comparison between *Entity Retrieval* and the common dense retrieval approach (e.g. Karpukhin et al., 2020) in identifying retrieval documents to enhance question answering with LLMs.

2 Retrieval-Augmentation Techniques

Retrieval-augmentation (Lewis et al., 2020b) is a method of converting closed-book question answering[2] (Roberts et al., 2020) into extractive question answering (Abney et al., 2000; Rajpurkar et al., 2016), where the answers can be directly extracted from the retrieved documents. Despite the abundance of effective retrieval-augmentation techniques for question answering in existing literature, this section will concentrate on a select few methods utilized to study the factoid question answering capabilities of LLMs in this paper.

Dense Passage Retrieval (DPR; Karpukhin et al., 2020) leverages a bi-encoder architecture, wherein the initial encoder processes the question and the subsequent encoder handles the passages to be retrieved. The similarity scores between the two encoded representations are computed using a dot product. The encoded representations of the second encoder are fixed and indexed in FAISS (Johnson et al., 2019; Douze et al., 2024), while the first encoder is optimized to maximize the dot-product scores based on positive and negative examples. The performance of DPR solidifies its position as a superior retriever compared to BM25-based (Robertson et al., 1994) sparse retrieval methods for question answering.

REPLUG (Shi et al., 2023) views LLM as a black-box, encoding each of the k most relevant retrieved documents along with the input query to generate k probability distributions over the forthcoming token. These distributions are then weighted averaged, considering the similarity of each retrieved document to the original input query. The strength of REPLUG lies in its ability to infuse the knowledge from the retrieved documents while generating the answer. This makes REPLUG a compelling candidate for studying retrieval-augmented question answering.

3 Entity Linking for Question Answering

While quite powerful, most retrieval-augmented systems are notably time and resource-intensive, necessitating the storage of extensive lookup indices and the need to attend to all retrieved documents to generate a response.

Entity linking has been an integral component of statistical factoid question answering systems (Aghaebrahimian and Jurčíček, 2016, *inter alia*). Additionally, the extensively studied field of Knowledge Base Question Answering (Cui et al., 2017, *inter alia*) has underscored the significance of entity information from knowledge bases in question answering (Salnikov et al., 2023).

A traditional neural question answering pipeline may contain entity detection, entity linking, relation prediction, and evidence integration (Mohammed et al., 2018; Lukovnikov et al., 2019), where entity detection can employ LSTM-based (Hochreiter and Schmidhuber, 1997) or BERT-based (Devlin et al., 2019) encoders. Inspired by this body of work, we investigate the relevance of entity linking as an alternative strategy to dense retrieval methods for augmenting factoid question answering with LLMs. We propose *Entity Retrieval*, a method employing a simple heuristic for implementing entity linking-based document retrieval. *Entity Retrieval* leverages entity linking to identify entities within the question and retrieves corresponding knowledge base articles, providing the first 100 words of each article as the retrieved documents (see Figure 1.b).

[2]Closed-book QA focuses on answering questions without additional context during inference.

4 Experiments

4.1 Setup

We focus on Wikipedia as the knowledge base and utilize the pre-existing Wikipedia passages and the dense retrieval model available in the `wiki_dpr`[3] repository from huggingface. `wiki_dpr` follows established practices (Chen et al., 2017; Karpukhin et al., 2020) and segments the articles into non-overlapping text blocks of 100 words, resulting in 21,015,300 passages. These passages are processed with a pre-trained DPR context encoder, generating fixed embedding vectors stored in a FAISS index (Douze et al., 2024). Factoid questions are encoded using the DPR question encoder, and the top k relevant passages to the encoded question are retrieved from the FAISS index. We use the `exact` FAISS index storage, `single-nq` DPR question encoder, and retrieve the top 4 documents for each question, in our experiments. As well for better time efficiency, following Ram et al. (2023), we treat document retrieval as a pre-processing step, caching the most relevant passages for each question before conducting the question answering experiments.

For entity linking in *Entity Retrieval*, we select SpEL (Shavarani and Sarkar, 2023) mainly due to its near-perfect linking precision. Architecturally, SpEL comprises an entity knowledge fine-tuned RoBERTa (Liu et al., 2019) model as the encoder and a classification layer atop the encoder which maps the encoded representations to the space of predicted entities. SpEL models entity linking as structured prediction which enables it to be fast and minimal resource demanding. In this study, we employ the fine-tuned SpEL-large model with an entity vocabulary of 500K, enabling identification of entity mentions referencing the 500K most hyperlinked Wikipedia pages.

Given the proven effectiveness of utilizing initial sentences from Wikipedia pages for entities in tasks such as document classification (Shavarani and Sekine, 2020) and question answering (Choi et al., 2018), we propose employing the first 100 words of Wikipedia articles corresponding to the identified entities in questions as retrieved documents for *Entity Retrieval* settings. We consider two such settings: (1) using SpEL for question annotation and utilizing its suggested linked entities to retrieve Wikipedia articles, (2) using gold entity link annotations for dataset questions to retrieve the Wikipedia articles.

For LLMs, we consider the open weight LLaMA 2 (Touvron et al., 2023) model in all three available sizes (7B, 13B, and 70B). However, due to hardware constraints — limited to 2 RTX A6000s with 49GB GPU memory each — we utilize the 8-bit quantized version of the 70B model. In all our experiments with LLaMA 2, we prevent it from generating sequences longer than 10 subwords. Additionally, we evaluate GPT 4 (`0613` version) from OpenAI (2023).

As a public implementation of REPLUG is not available, we implement it with the haystack[4] library, employing our cached DPR passages for each question to autoregressively generate answers.

To verify the capacity of LLMs in utilizing the retrieved documents without additional fine-tuning or further in-context examples, we do not use any training question-answer pairs in the prompts of our models. Aside from a simple instruction for answering the question, in the closed-book setting, the prompt solely comprises the question, while in the DPR and REPLUG settings, it includes the retrieved documents from the DPR cache along with the question. Similarly, for the *Entity Retrieval* settings, the prompt consists of the first 100 words of the Wikipedia pages corresponding to the identified or gold entities in the question. We follow Ram et al. (2023) for question normalization and prompt formulation.

4.2 Data

We use the following datasets in our experiments:

FactoidQA (Smith et al., 2008) contains 2203 hand crafted factoid question-answer pairs derived from Wikipedia articles, with each pair accompanied by its corresponding Wikipedia source article included in the dataset. We use `OpenQA-eval` (Kamalloo et al., 2023) scripts to evaluate model performance, reporting exact match (EM) and F1 scores by comparing expected answers to model responses for FactoidQA questions.

StrategyQA (Geva et al., 2021) is a complex boolean question answering dataset, constructed by presenting individual terms from Wikipedia to annotators. Its questions contain references to more than one Wikipedia entity, and necessitate implicit reasoning for binary responses. The dataset comprises 5111 answered questions which are split into two subsets: `train` and `train_filtered` subsets

[3] https://huggingface.co/datasets/wiki_dpr, created on a Wikipedia dump from December 20, 2018.

[4] https://haystack.deepset.ai

Setting	LLaMA 2						GPT 4	
	7B		13B		70B-8bQ			
	EM	F1	EM	F1	EM	F1	EM	F1
Closed-book	30.5	38.3	33.7	42.9	37.0	45.9	**42.4**	**55.1**
DPR	33.6	42.7	37.1	45.5	35.6	42.0	35.9	47.1
REPLUG	15.8	22.0	27.7	33.4	22.0	25.3	-	-
Entity Retrieval w/ SPEL	**38.1**	**47.3**	**40.5**	**49.2**	**40.6**	**49.0**	37.3	48.0
Entity Retrieval w/ oracle entities†	37.2	46.9	40.2	49.3	41.4	49.7	38.5	48.2

Table 1: FactoidQA evaluation results. EM refers to the exact match between predicted and expected answers, disregarding punctuation and articles (a, an, the). †*Entity Retrieval* with oracle results are not directly comparable to other approaches, as they leverage gold annotated entity links from the dataset.

	Setting	7B		13B		70B-8bQ	
		Acc	Inv #	Acc	Inv #	Acc	Inv #
train	Closed-book	51.8	215	51.4	302	**60.3**	191
	DPR	52.8	212	52.5	280	52.8	336
	Entity Retrieval w/ SPEL	**53.8**	**175**	**52.7**	**200**	58.0	**152**
train_filtered	Closed-book	59.9	286	61.6	337	**67.0**	232
	DPR	59.8	274	63.7	296	62.7	407
	Entity Retrieval w/ SPEL	**60.1**	**233**	**64.9**	**190**	66.6	**206**

Table 2: StrategyQA evaluation with LLaMA 2 results.

containing 2290 and 2821 questions, respectively. For evaluation, we present accuracy scores by comparing model responses to the expected boolean answers in the dataset. As well, to assess model comprehension of the task, we count the number of invalid answers that deviate from Yes or No and report this count in a distinct column labeled "Inv #" for each experiment.

4.3 Results and Analysis

We generate answers to FactoidQA questions for the following settings: (1) closed-book, (2) DPR, (3) REPLUG, (4) *Entity Retrieval* with SPEL-identified entities, and (5) *Entity Retrieval* with oracle entity annotations from the dataset. Table 1 summarizes our evaluation results.

Our experimental results prove that *Entity Retrieval* is a formidable contender among retrieval-augmented techniques for factoid question answering, particularly exhibiting enhanced efficacy with smaller LLMs. The outcomes from our GPT-4[5] experiments substantiate this assertion, revealing a consistent decline in performance across all investigated retrieval-augmentation techniques.

Furthermore, comparing the evaluation results using SPEL identified entities and the oracle entities for *Entity Retrieval*, we realize that despite SPEL's constrained entity lexicon comprising the 500K most hyperlinked entities, its performance remains notably competitive. While acknowledging this observation, we defer a comprehensive evaluation of alternative entity linking methods beyond SPEL to future investigations.

Table 2 presents our evaluation results for the StrategyQA dataset. Notably, *Entity Retrieval* with oracle annotations is excluded due to the absence of oracle entity links for questions in StrategyQA, while the exclusion of REPLUG is attributed to its comparatively inferior performance relative to DPR in FactoidQA experiments. Our results affirm our previous inference that retrieval-augmentation is not beneficial with sufficiently large models. However, despite the complex reasoning demanded by this dataset, *Entity Retrieval* achieves comparable results to other retrieval-augmented methods, while offering better hardware efficiency. Additionally, invalid count values indicate that *Entity Retrieval* is capable of aiding the model in understanding the boolean nature of expected responses without relying on dense retrieval from millions of passages.

5 Conclusion

We highlight the disproportionate benefit of retrieval augmentation for smaller LLMs in the context of factoid question answering, and introduce *Entity Retrieval* as a promising entity linking-based alternative to dense retrieval for augmenting factoid questions in prompting LLMs.

[5]Despite undisclosed specifications of GPT-4 models, extrapolating from the known size of GPT-3 (175B parameters; Brown et al., 2020), it is plausible to estimate GPT-4 to surpass 200 billion parameters, with speculations suggesting over 1 trillion parameters implemented as a mixture of experts model.

Limitations and Ethical Considerations

We have not exhaustively explored all potential entity linking methods, which may yield insights enhancing the proposed *Entity Retrieval* approach.

Additionally, due to space constraints and a desire to expedite community engagement, we have not incorporated additional datasets (e.g. 30MFQA; Serban et al., 2016), most of which are annotated with Freebase (Bollacker et al., 2008) and have fallen into disuse following Freebase's discontinuation. We intend to revitalize such neglected factoid question answering datasets, and we posit that revitalizing these datasets could facilitate the development of a benchmark dataset akin to MMLU (Hendrycks et al., 2021), enabling robust evaluations of newly released LLMs in terms of their factual knowledge capabilities.

Our research is on English only, and we acknowledge that factoid question answering in other languages is also relevant and important. We hope to extend our work to cover multiple languages in the future. We inherit the biases that exist in the data used in this project, and we do not explicitly de-bias the data. We are providing our code to the research community and we trust that those who use the model will do so ethically and responsibly.

References

Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. In *Sixth Applied Natural Language Processing Conference*, pages 296–301, Seattle, Washington, USA. Association for Computational Linguistics.

Emmanuel Adebisi, Bolanle Adefowoke Ojokoh, and Folasade Olubusola Isinkaye. 2022. An open domain factoid qa framework with improved validation techniques. *International Journal of Information Science and Management (IJISM)*, 20(1).

Ahmad Aghaebrahimian and Filip Jurčíček. 2016. Open-domain factoid question answering via knowledge graph search. In *Proceedings of the Workshop on Human-Computer Question Answering*, pages 22–28, San Diego, California. Association for Computational Linguistics.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. Kbqa: Learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10(5).

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.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Lin Li, Mengjing Zhang, Zhaohui Chao, and Jianwen Xiang. 2021. Using context information to enhance simple question answering. *World Wide Web*, 24:249–277.

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*, abs/1907.11692.

Denis Lukovnikov, Asja Fischer, and Jens Lehmann. 2019. Pretrained transformers for simple question answering over knowledge graphs. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pages 470–486. Springer.

Denis Lukovnikov, Asja Fischer, Jens Lehmann, and Sören Auer. 2017. Neural network-based question answering over knowledge graphs on word and character level. In *Proceedings of the 26th international conference on World Wide Web*, pages 1211–1220.

Salman Mohammed, Peng Shi, and Jimmy Lin. 2018. Strong baselines for simple question answering over knowledge graphs with and without neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Nagehan Pala Er and Ilyas Cicekli. 2013. A factoid question answering system using answer pattern matching. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 854–858, Nagoya, Japan. Asian Federation of Natural Language Processing.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Michael Petrochuk and Luke Zettlemoyer. 2018. SimpleQuestions nearly solved: A new upperbound and baseline approach. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 554–558, Brussels, Belgium. 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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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.

Prakash Ranjan and Rakesh Chandra Balabantaray. 2016. Question answering system for factoid based question. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 221–224. IEEE.

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.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Hong Kong, China. Association for Computational Linguistics.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Hassan S. Shavarani and Anoop Sarkar. 2023. SpEL: Structured prediction for entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11123–11137, Singapore. Association for Computational Linguistics.

Hassan S. Shavarani and Satoshi Sekine. 2020. Multiclass multilingual classification of Wikipedia articles using extended named entity tag set. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1197–1201, Marseille, France. European Language Resources Association.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Tanzim Tamanna Shitu, Nazia Zaman, and KM Azharul Hasan. 2020. Domain specific factoid question answering by regular expression generation. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 1464–1467. IEEE.

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.

Noah A Smith, Michael Heilman, and Rebecca Hwa. 2008. Question generation as a competitive undergraduate course project. In *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, volume 9.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.