Retrieving Contextual Information for Long-Form Question Answering using Weak Supervision

Philipp Christmann[*]
Max Planck Institute for Informatics
pchristm@mpi-inf.mpg.de

Svitlana Vakulenko
Amazon AGI
svvakul@amazon.com

Ionut Teodor Sorodoc
Amazon AGI
csorionu@amazon.com

Bill Byrne
Amazon AGI
willbyrn@amazon.co.uk

Adrià de Gispert
Amazon AGI
agispert@amazon.com

Abstract

Long-form question answering (LFQA) aims at generating in-depth answers to end-user questions, providing relevant information beyond the direct answer. However, existing retrievers are typically optimized towards information that directly targets the question, missing out on such contextual information. Furthermore, there is a lack of training data for relevant context. To this end, we propose and compare different weak supervision techniques to optimize retrieval for contextual information. Experiments demonstrate improvements on the end-to-end QA performance on ASQA, a dataset for long-form question answering. Importantly, as more contextual information is retrieved, we improve the relevant page recall for LFQA by 14.7% and the groundedness of generated long-form answers by 12.5%. Finally, we show that long-form answers often anticipate likely follow-up questions, via experiments on a conversational QA dataset.

1 Introduction

The goal of long-form question answering (LFQA) is to provide in-depth answers to end-user questions (Stelmakh et al., 2022; Fan et al., 2019). For example, for the user question

> *"When did Lionel Messi start his career?"*

a direct answer would be:

```
16 November 2003
```

which is the date of his first team debut.

However, this answer naturally sparks a series of follow-up questions to obtain more contextual details (Kumar and Joshi, 2017):

> *"For which club?"*
> *"In which match?"*
> *"What about his La Liga debut?"*
> *"How did his career develop?"*

A long-form answer aims to proactively supply a more complete and detailed response beyond the succinct direct answer, essentially anticipating such follow-up questions:

> *Lionel Messi started his career as a professional football player with FC Barcelona. He made his first-team debut in a friendly against Porto on 16 November 2003, at the age of 16. . . .*

We thus draw parallels between LFQA and conversational question answering (ConvQA) (Voskarides et al., 2020; Qu et al., 2020; Vakulenko et al., 2021; Christmann et al., 2023; Coman et al., 2023) and hypothesize that they can be treated as complementary tasks.

Existing approaches for LFQA based on retrieval-augmented generation (Lewis et al., 2020; Izacard and Grave, 2021; Guu et al., 2020) typically utilize retrievers that are optimized to obtain direct answers to questions. Such retrieval systems thus often fail to retrieve relevant context, which is needed to generate faithful and comprehensive long-form answers.

Approach. We propose to train a specialised retriever for the task of LFQA that not only retrieves direct answers for a question, but also retrieves additional context required for grounding long-form answers. Our goal is to retrieve both direct answers and contextual information in one shot.

The major bottleneck for training a retriever for LFQA is the absence of training data. LFQA datasets (Fan et al., 2019; Stelmakh et al., 2022) contain questions and long-form answers but not the ground-truth passages required to produce those long-form answers. In this work, we propose a mechanism to automatically infer *silver passages*, designed to ground both (i) the direct answers, **and** (ii) the contextual information. These passages should provide sufficient evidence to support information in the long-form answer. This is different from previous work on factoid short-form QA,

[*] Work was done during an internship at Amazon AGI.

Findings of the Association for Computational Linguistics: EACL 2024, pages 14301–14310
November 12-16, 2024 ©2024 Association for Computational Linguistics

which identified such passages by matching only against the direct answers (Shen et al., 2023).

Based on these silver passages, we train BERT-based re-ranking models (Nogueira and Cho, 2019). These re-rankers are applied to the initial retrieval results, to enhance recall of contextual information in the top-ranked passages. These top-ranked passages are then provided as input to an LLM to generate the long-form answer.

We conduct experiments on ASQA (Stelmakh et al., 2022), a dataset for LFQA, and show that we substantially improve end-to-end QA performance, while also increasing the groundedness of the long-form answer w.r.t. the retrieved passages.

Experiments on the ConvQA dataset ConvMix (Christmann et al., 2022) demonstrate that our method generates long-form answers that often also contain answers to the follow-up questions, when provided only with the first question of the conversation as input. This indicates that LFQA can indeed anticipate likely follow-up questions.

Contributions.

- A novel mechanism using the target long-form answers to identify silver passages expressing relevant contextual information.
- Improving end-to-end QA performance, achieving state-of-the-art performance on the competitive ASQA benchmark.
- An investigation into the relationship between the LFQA and ConvQA tasks as alternatives for satisfying the same information needs.

2 Identifying silver passages

The key idea for obtaining such silver passages is to utilize both the long-form answers (LFAs) **and** direct answers (DAs), as annotated in the ASQA dataset, jointly as a weak supervision signal for passage relevance. We retrieve a large set of passages first (say 100), using first-stage retrieval (Karpukhin et al., 2020), and then choose up to k silver passages from this candidate pool.

We considered three techniques for matching candidate passages against an LFA: (i) lexical matching, (ii) semantic similarity, and (iii) LLM perplexity. Our proposed approach matches against a combination of both LFAs and DAs, and we also compare these against matching only with DAs.

Lexical matching with LFA. We initially evaluated (i) token recall, (ii) Jaccard similarity between token sets, and (iii) ROUGE-L (Lin, 2004), and found that plain token recall works best. We thus compute the matching score for a pair of a candidate passage p and the LFA as:

$$match(p, LFA) = \frac{|tokens(p) \cap tokens(LFA)|}{|tokens(LFA)|} \tag{1}$$

where $tokens$ are sets of words produced by a tokenizer with stopword removal.

Semantic similarity with LFA. We use a pre-trained Sentence-Transformer model[1] (Reimers and Gurevych, 2019; Vaswani et al., 2017), to compute the semantic similarity between a candidate p and the LFA as follows (where Enc is the text encoder, and $\cdot$ is the dot product):

$$match(p, LFA) = Enc(p) \cdot Enc(LFA) \tag{2}$$

LLM perplexity of LFA. Inspired by the approach used in Toolformer (Schick et al., 2023), we compute the LLM perplexity of the target LFA (of length n), given a candidate passage p, as follows:

$$\begin{aligned} &match(p, LFA) \\ &\quad = -\sum\nolimits_{j=1}^{n} \log P(t_j|C, p, t_1...t_{j-1}) \end{aligned} \tag{3}$$

where P is the probability function of the LLM, t_j is the j-th token of the LFA, and C is the same prompt as the one applied during LLM training and inference. C includes a random sample of k-1 candidate passages. We observed that adding this random sample substantially improves the performance as it makes the samples closer to the input seen during LLM inference/training, i.e. a context with k passages.

Matching with DA. All candidate passages that contain one of the DAs are considered as relevant. Here we consider exact lexical matches since the DAs are relatively short. This is the typical approach when the goal is to provide crisp answers (Shen et al., 2023). Since there might be multiple passages matching the answer, and answers may also be matched out-of-context, we sort all answer-matching passages by their token-recall with the question.

Silver: Matching with LFA & DA. Finally, we consider the combination of matching both the LFA and the set of DAs, for selecting k silver passages. First, we ensure that each DA is matched by at least one of the candidate passages. In case there are multiple candidate passages matching the same

[1] https://huggingface.co/sentence-transformers/nli-roberta-base-v2

DA, the one with the highest matching score with the LFA is chosen as relevant. The remaining passages are chosen based on their LFA matching score, to obtain a total of k silver passages. We utilize lexical matching with the LFA, which showed strong results (see Sec. 3) and is computationally inexpensive. This combined variant is the SILVER approach proposed in this work, as it jointly optimizes towards information for directly answering the question and relevant contextual information.

Pseudo-code for SILVER approach. Algorithm 1 in the Appendix illustrates the end-to-end workflow of deriving silver passages, fine-tuning the re-ranker, and fine-tuning the LLM on LFQA data.

3 Experiments

Benchmarks. We conduct experiments on two datasets: (i) ASQA (Stelmakh et al., 2022), a dataset for LFQA, and (ii) CONVMIX (Christmann et al., 2022), a dataset for ConvQA. We use ASQA for fine-tuning the re-rankers and LLMs. Experiments on CONVMIX use the same models, thus also test the generalizability of the approach.

Retrieval metrics. For evaluating retrieval, we measure recall in the top-5 retrieved passages. On ASQA, we compute *Direct Answer Recall* as the fraction of DAs appearing in the retrieved passages. As a proxy for recall of contextual information, we measure *Wikipage Recall* as the fraction of the ground-truth relevant Wikipages matched with our retrieval results. A Wikipage is matched if we retrieve a passage from the respective page. On CONVMIX, we measure *Direct Answer Recall* / *Follow-up Answer Recall* as the fraction of DAs for the first question / follow-up questions appearing in the retrieved passages.

Metrics. On ASQA, we keep the metrics used in the original work, and use their evaluation code[2]. This includes *ROUGE-L* (Lin, 2004) for measuring the overlap of the generated text with one of the two reference LFAs. For answer correctness, the Disambig-F1 (*D-F1*) metric is used, measuring the fraction of question interpretations answerable from the generated LFA. A pre-trained machine reading comprehension (MRC) (Hermann et al., 2015) model is used to this end. The *DR* metric combines the two metrics via the geometric mean.

For CONVMIX, inspired by the D-F1 metric, we compute *C-F1* as the fraction of conversational questions answerable from the generated LFA. We use the same MRC model as for the D-F1 metric.

In addition, we measure *Groundedness* as the fraction of tokens in the answer that is also present in the retrieved passages. Similar to lexical matching in Sec. 2, stopwords are not considered.

Configuration. We use DPR (Karpukhin et al., 2020)[3] for first-stage retrieval, which was shown to outperform BM25 (Robertson and Zaragoza, 2009) on ASQA in previous work (Sun et al., 2023), and is still considered state-of-the-art on NaturalQuestions (Kwiatkowski et al., 2019), which is a superset of ASQA.

We use Vicuna 13B 1.5 (Zheng et al., 2023) as LLM for generating LFQAs and fine-tune it for 1 epoch using the long-form answers in the ASQA dataset as the target output. The input to the model are the question and the top-5 passages retrieved either by DPR or by our SILVER re-rankers. We use the same LLM prompt as in the original ASQA paper to combine the question and the retrieved passages for consistency.

Further details on the setup in Appendix A.

3.1 Results

Recall of contextual information is enhanced. Table 1 shows our main results. The foremost take-away is that recall of contextual information is greatly improved when incorporating our SILVER re-rankers, compared to DPR. On ASQA, Wikipage recall increases from 0.450 to 0.516. On CONVMIX, follow-up answer recall is improved from 0.260 to 0.306, indicating that our SILVER re-rankers aid the LLM to anticipate and successfully answer follow-up questions.

Lexical matching shows strong performance. An interesting finding is that simple lexical matching achieves better performance (DR of 35.0) compared to the more complex and computationally expensive variants based on semantic similarity (DR of 34.1) or LLM perplexity (DR of 34.0). This result demonstrates that lexical matching is sufficiently robust for the relatively long answers.

Improving end-to-end QA performance. Combining LFA and DA to provide supervision signal for training the re-ranker leads to the best performance (DR of 35.5), substantially improving over the baseline results with DPR (DR of 34.3). Both improvements in answer formulation (ROUGE-L of 43.4 vs. 42.5) and provision of the right answer

[2] https://github.com/google-research/language/tree/master/language/asqa

[3] DPR is trained on NQ (Kwiatkowski et al., 2019) and used along the corresponding Wikipedia dump.

	ASQA (Stelmakh et al., 2022)						CONVMIX (Christmann et al., 2022)			
Metric → Retrieval Method ↓	Recall: Direct Ans.	Recall: Wikipage	Ground.	ROUGE-L	D-F1	DR	Recall: Direct Ans.	Recall: Follow-up Ans.	Ground.	C-F1
Baseline retriever	0.489	0.450	0.763	42.5	27.7	34.3	0.558	0.260	0.726	20.2
+ pre-trained re-ranker	0.336	0.345	0.664	40.1	23.7	30.8	0.514	0.235	0.683	19.9
Matching with DA	**0.498**	0.483	0.810	42.9	28.0	34.7	0.622	0.292	0.803	**20.8**
Lexical matching with LFA	0.489	0.501	0.851	43.3	28.3	35.0	**0.635**	**0.306**	0.828	**20.8**
Semantic similarity with LFA	0.482	0.500	0.829	42.8	27.2	34.1	0.619	0.303	0.817	20.4
LLM perplexity of LFA	0.483	0.495	0.845	43.3	26.7	34.0	0.625	0.305	0.838	20.7
SILVER: Matching with LFA & DA	0.491	**0.516**	**0.858**	**43.4**	**29.0**	**35.5**	**0.635**	**0.306**	**0.839**	**20.8**

Table 1: Main results comparing retrieval and end-to-end QA performance on ASQA and CONVMIX. We use DPR (Karpukhin et al., 2020) as our retrieval baseline, which has been commonly used on the ASQA dataset.

Method	Length	ROUGE-L	D-F1	DR
FLARE (Jiang et al., 2023)	–	34.3	28.2	31.1
PaLM 540B (Amplayo et al., 2023)	64.1	40.7	27.8	33.5
JPR + T5-large (Stelmakh et al., 2022)	71.6	43.0	26.4	33.7
SIXPAQ (Sun et al., 2023)	63.5	43.8	28.9	35.6
DPR + VICUNA 13B	70.5	43.0	28.3	34.9
SILVER + VICUNA 13B (proposed)	70.0	**44.1**	**30.8**	**36.9**

Table 2: Comparison with state-of-the-art on ASQA.

in an appropriate context (D-F1 of 29.0 vs. 27.7) contribute to this overall increase in performance.

Groundedness is substantially enhanced. Another key take-away is the effect of our SILVER re-rankers on the groundedness of the generated answers. The groundedness is dramatically improved on both datasets compared to DPR retrieval (0.763 to 0.858 on ASQA and 0.726 to 0.839 on CONVMIX). This result indicates that our generated LFAs are more likely to be based on the retrieved passages rather than hallucinated by the LLM.

Comparison against a pre-trained re-ranker. We also conducted an experiment with a pre-trained re-ranker replacing our proposed SILVER re-rankers (results are shown in Table 1). We used the same Sentence-Transformer as for our semantic similarity variant[4]. As can be expected, recall drops substantially, which leads to a much lower DR score. Interestingly, since relevant information is missing from the retrieval results, the groundedness of answers is greatly reduced in comparison to our proposed approach. Note that our re-rankers, once trained on the ASQA dataset, can be successfully applied to a different dataset (CONVMIX in our experiments).

3.2 Analysis

Anecdotal examples. Table 3 demonstrates how our approach can improve the groundedness of generated answers: the fraction of un-grounded tokens (i.e. the ones not present in the retrieved passages), as depicted in orange, is much higher when conditioned on DPR retrieval. When conditioned on our SILVER retrieval, the LLM can mostly rely on the information in the provided passages for generating LFAs. This can reduce factual hallucinations compared to the DPR-based variant, as illustrated in the second example: as DPR retrieval is insufficient, the LLM hallucinates incorrect contextual information (e.g., that Sopranos started on *January 15, 1999*).

Comparison with state-of-the-art. Our approach achieves state-of-the-art performance on the ASQA dataset, as shown in Table 2.

Answer recall per turn. We plot the average answer recall per turn on CONVMIX in Fig. 1. As expected, the answer recall drops as the conversation drifts away from the initial topic. Recall with our SILVER re-ranker remains consistently higher than for DPR.

Effect of LLM size. We further investigate the effect of the LLM size, to verify that our improvements still hold for smaller/larger LLMs, using the 7B and 33B versions of Vicuna. Results are shown in Table 4. In general, the DR metric decreases (SILVER: 33.3; DPR: 32.5) using the 7B version. Further, we found that the effect of scaling up the LLM to 33B parameters is negligible compared to enhancements on the retrieval side, observing very similar results as for the 13B version (DR metric for SILVER: 35.6; DPR: 34.3).

4 Related work

Long-form question answering. With the recent advances of LLMs (Ouyang et al., 2022; Devlin et al., 2019; Zheng et al., 2023) question answering has evolved beyond crisp and direct answers (Yahya et al., 2013; Bast and Haussmann, 2015; Sun et al., 2018; Roy and Anand, 2022) toward supplying more in-depth and comprehensive passage-length responses. There has been exten-

[4] https://huggingface.co/sentence-transformers/nli-roberta-base-v2

Question	*"Who played grandpa joe in charlie and the chocolate factory?"* (from ASQA)
DPR + VICUNA 13B	*Willy Wonka and the Chocolate Factory is a 1970 children's book by Roald Dahl. It has been adapted into various media, including a 1971 film and a 2005 film. In the 1971 film, Grandpa Joe is played by Wilbur Robinson. In the 2005 film, Grandpa Joe is played by David Morris.* – **Groundedness:** 0.61
SILVER + VICUNA 13B (proposed)	*Charlie and the Chocolate Factory is a 2005 musical fantasy comedy film directed by Tim Burton and written by John August, based on the 1964 British novel of the same name by Roald Dahl. The film stars Johnny Depp as Willy Wonka and Freddie Highmore as Charlie Bucket. The storyline follows Charlie, who wins a contest and, along with four other contest winners, is led by Wonka on a tour of his chocolate factory, the most magnificent in the world. In the film, Grandpa Joe is played by David Kelly. In the 2010 musical adaptation of Charlie and the Chocolate Factory, Grandpa Joe is played by Richard Dreyfuss.* – **Groundedness:** 0.94
Question	*"Number of episodes in The Sopranos?"* (from CONVMIX)
DPR + VICUNA 13B	*The Sopranos is an American crime drama television series created by David Chase. The series ran for 73 episodes over six seasons, from January 15, 1999, to June 11, 2007.* – **Groundedness:** 0.23
SILVER + VICUNA 13B (proposed)	*The Sopranos is an American crime drama television series created by David Chase. The series portrays the difficulties that fictional New Jersey-based, Italian American mobster Tony Soprano (James Gandolfini) faces as he tries to balance his home life and his criminal organization. The series features Tony's family members, mafia colleagues, and rivals in prominent roles and story arcs, most notably his wife Carmela (Edie Falco) and protégé Christopher Moltisanti (Michael Imperioli). The series ran for 86 episodes over six seasons.* – **Groundedness:** 1.00

Table 3: Anecdotal examples from both datasets (randomly sampled) illustrating the benefits of explicitly retrieving contextual information. Un-grounded tokens (i.e., the ones not appearing in the retrieved passages) are highlighted in orange. The fraction of grounded tokens is greatly improved with our proposed approach in both cases.

Method	ROUGE-L	D-F1	DR
DPR + VICUNA 7B	40.6	26.0	32.5
DPR + VICUNA 13B	42.5	27.7	34.3
DPR + VICUNA 33B	42.6	27.8	34.3
SILVER + VICUNA 7B	42.3	26.3	33.3
SILVER + VICUNA 13B	43.4	29.0	35.5
SILVER + VICUNA 33B	43.6	29.0	35.6

Table 4: Comparison of end-to-end QA performance with different model sizes of the VICUNA model family.

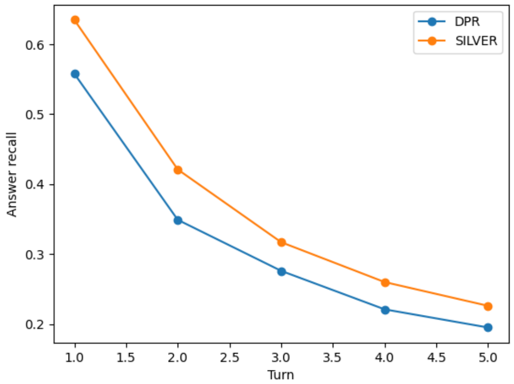

Figure 1: Answer recall per turn on CONVMIX.

sive research on LFQA recently (Nakano et al., 2021; Stelmakh et al., 2022; Amplayo et al., 2023; Jiang et al., 2023; Fan et al., 2019; Su et al., 2022; Wang et al., 2022; Krishna et al., 2021; Gao et al., 2023; Sun et al., 2023), which has mostly built upon retrieval systems optimized for retrieving direct answers. More details and discussion in the Appendix B.

Weak supervision for training retrieval systems. Obtaining training data for retrieval in QA has been a long-standing challenge (Shen et al., 2023). The most common approach to obtain training samples is to consider all passages matching the direct answer as relevant (Karpukhin et al., 2020; Sachan et al., 2021; Joshi et al., 2017). We extend this approach to the task of LFQA showing how to use both long-form and direct answers for optimizing retrieval toward contextual information.

5 Conclusion

The retrieval of contextual information is often neglected in LFQA, while being an important ingredient for generating and grounding comprehensive long-form answers. We investigate techniques to obtain training samples providing such contextual information for training re-ranking models. We show that incorporating our re-rankers improves retrieval and QA performance on a LFQA dataset, yielding state-of-the-art performance on ASQA. Notably, our method enhances groundedness of generated texts by 12.5%, which can reduce factual hallucinations in answers. Experiments on CONVMIX show that our method, trained on ASQA, is able to generalize to an unseen ConvQA dataset.

Acknowledgements. We thank Rexhina Blloshmi from Amazon AGI for useful inputs at various stages of this work.

Bill Byrne holds concurrent appointments as an Amazon Scholar and as Professor of Information Engineering at the University of Cambridge. This paper describes work performed at Amazon.

6 Limitations

Our experimental setup with the ASQA dataset, which has reference long-form answers and short-form answers, allows us to investigate the duality of the LFQA and ConvQA tasks. However, we only evaluated our approach on data that is publicly available. We leave it for future work to run experiments with the approach in the wild.

In this work, we showed improvements on the retrieval side of a RAG pipeline based on Vicuna models of different sizes. RAG pipelines based on other language model architectures might be affected differently by the enhanced retrieval recall provided by our approach, which is not investigated in this paper.

7 Ethical considerations

We did not collect or release any private data or user data in this work. All experiments are based on static datasets.

We make use of LLMs, which are known to generate factually incorrect texts or hallucinations. In this work, we aim to enhance the grounding of long-form answers by explicitly retrieving passages for contextual information. Experiments indicate that with our approach the groundedness of answers can be improved, which is a promising direction of reducing hallucinations in LLMs.

References

Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2023. Query refinement prompts for closed-book long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Hannah Bast and Elmar Haussmann. 2015. More accurate question answering on Freebase. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*.

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. In *Advances in Neural Information Processing Systems*.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Conversational question answering on heterogeneous sources. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2023. Explainable conversational question answering over heterogeneous sources via iterative graph neural networks. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Andrei C Coman, Gianni Barlacchi, and Adrià de Gispert. 2023. Strong and efficient baselines for open domain conversational question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

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*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*.

Gautier Izacard and Édouard 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*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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)*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Vineet Kumar and Sachindra Joshi. 2017. Incomplete follow-up question resolution using retrieval based sequence to sequence learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, and Kenton Lee. 2019. Natural Questions: A benchmark for question answering research. In *Transactions of the Association for Computational Linguistics*. MIT Press.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. In *arXiv*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, et al. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. In *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in neural information processing systems*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*.

Rishiraj Saha Roy and Avishek Anand. 2022. *Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections*. Springer.

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

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Xiaoyu Shen, Svitlana Vakulenko, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. 2023. Neural ranking with weak supervision for open-domain question answering: A survey. In *Findings of the Association for Computational Linguistics: EACL 2023*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid Questions Meet Long-Form Answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before Generate! Faithful Long Form Question Answering with Machine Reading. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Haitian Sun, William W Cohen, and Ruslan Salakhutdinov. 2023. Answering ambiguous questions with a database of questions, answers, and revisions. In *arXiv*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W. Cohen. 2018. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mohamed Yahya, Klaus Berberich, Shady Elbassuoni, and Gerhard Weikum. 2013. Robust question answering over the Web of linked data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Algorithm 1 Pseudo-code for SILVER approach

```
Inputs:
    $\mathcal{D}$: Dataset with questions, LFAs and DAs;
    $\mathcal{R}$: pre-trained BERT;
    $\mathcal{M}$: pre-trained causal LLM;
    DPR: first-stage retriever;
    $k$: number of passages in context;
Outputs:
    $\mathcal{R}_{SFT}$: fine-tuned re-ranker;
    $\mathcal{M}_{SFT}$: fine-tuned causal LLM;

# Identify silver passages
$\mathcal{D}_P \leftarrow \{\}$;
for all $(q, LFA, DA) \in \mathcal{D}$ do
    # First-stage retrieval (top-100)
    $P_{DPR} \leftarrow \text{DPR}(q, 100)$;
    # Compute matching with LFA
    for all $p \in P_{DPR}$ do
        $score(p) \leftarrow match(p, LFA)$;
    end for
    # Compute silver passages $P^*$
    $P^*_{DA} \leftarrow \{p | DA \in p\}$;
    $P^*_{LFA} \leftarrow sort(\{p\}, score)$;
    $P^* \leftarrow top_k(P^*_{DA}, P^*_{LFA}, score)$;
    $P^- \leftarrow sample(P - P^*, 50)$;
    # Add to training data
    $\mathcal{D}_P \leftarrow \mathcal{D}_P \cup \{(q, P^*, P^-)\}$;
end for
# Fine-tune re-ranking model
$\mathcal{R}_{SFT} \leftarrow finetune(\mathcal{R}, \mathcal{D}_P)$;
# Fine-tune LLM
$\mathcal{D}_{LFA} \leftarrow \{\}$;
for all $(q, LFA, DA) \in \mathcal{D}$ do
    $P_{DPR} \leftarrow \text{DPR}(q, 100)$;
    # Apply re-ranker
    $P_{\mathcal{R}} \leftarrow \mathcal{R}_{SFT}(P_{DPR}, k)$;
    $\mathcal{D}_{LFA} \leftarrow \mathcal{D}_{LFA} \cup \{(q, P_{\mathcal{R}}, LFA)\}$;
end for
$\mathcal{M}_{SFT} \leftarrow finetune(\mathcal{M}, \mathcal{D}_{LFA})$;
```

A Additional details on experiments

Datasets. ASQA has 6,316 ambiguous questions that originate from the Google search log (Kwiatkowski et al., 2019; Min et al., 2020) paired with LFAs written by crowdworkers. Every sample contains a set of alternative question interpretations and a DA corresponding to each of them. An example ambiguous question is *"Who played bonnie in gone with the wind?"*, with the two interpretations being *"Who played bonnie in the gone with the wind film?"* and *"Who played bonnie in the gone with the wind musical?"*. The corresponding DAs are `Cammie King` and `Leilah de Meza` in this case. The dataset provides one LFA for each question in the train set, and two LFAs for each question in the dev set, with an average of 64.8 words per LFA (dev set). Since the test set is hidden, we split the train set, using 95% for training and 5% for development, and the original dev set as our test set. We used the official ASQA evaluation code[5] to obtain the ROUGE-L, D-F1 and DR metrics. The dataset is licensed under an Apache License 2.0, thereby permitting its use for research purposes.

We use CONVMIX only for the evaluation of the models trained on the ASQA dataset, since CONVMIX does not provide LFAs to train on. We input only the first question from each conversation (3,000 questions, in total) and evaluate whether the generated LFA provides answers to the (first 4) follow-up questions from the conversation. The dataset is licensed under a CC BY 4.0, thereby permitting its use for research purposes.

Implementation details. We implement SILVER re-rankers as cross-encoders based on BERT models (Devlin et al., 2019)[6] with 110M parameters. The input format is the following: `"[CLS] question [SEP] passage_title [SEP] passage_text"`, and the output is a scalar indicating the relevance of the respective passage. We apply the re-ranker on top-100 DPR results (better performance than for top-1,000). For training the re-ranker, for each question, we randomly sampled 50 negatives (non-silver passages) along with 5 positives (top-5 silver passages) from the top-100 DPR passages. We used AdamW as optimizer with a learning rate of 10^{-5}, batch size of 16, weight decay of 0.01, and warm-up ratio of 0.04. Binary cross-entropy is used as loss function.

Comparison with state-of-the-art. For this comparison, we used the full ASQA train set, same as in related work. Note that the LLM here is trained for one epoch, without optimization on the dev set. As ASQA is a rather small-scale dataset with only 4,353 instances in the original train set, the additional 5% (compared with our new split) make quite an impact on the LLM performance: the DR

[5] https://github.com/google-research/language/tree/master/language/asqa

[6] https://huggingface.co/google-bert/bert-base-uncased

metric improves from 35.5 to 36.9.

Computational costs. Fine-tuning the BERT-based re-ranker took 50 minutes on an AWS EC2 P3 instance. The LLM fine-tuning (Vicuna 13B) took 110 minutes on an AWS EC2 P4 instance.

B Details on related work

Long-form question answering. The ELI5 dataset (Fan et al., 2019) facilitated initial research on LFQA, but due to evaluation problems (Krishna et al., 2021) recent work used the ASQA dataset with factoid long-form answers for fairer comparison (Stelmakh et al., 2022; Amplayo et al., 2023; Jiang et al., 2023; Sun et al., 2023). Yehudai et al. (2024) propose Genie, an approach to create a synthetic LFQA dataset from Wikipedia, similar to ASQA. The state-of-the-art methods for LFQA built upon retrieval systems that are optimized for retrieving direct answers to questions (Stelmakh et al., 2022; Krishna et al., 2021; Sun et al., 2023; Gao et al., 2023).

Sun et al. (2023) proposed SIXPAQ, which constructs a database of potential questions paired with their direct answers, to augment the information obtained from the dense retriever (Ni et al., 2022). Stelmakh et al. (2022) used JPR (Min et al., 2021), an out-of-the-box re-ranker operating on top of DPR results. JPR is optimized for diversifying retrieval of direct answers, and thus for targeting ambiguous questions (the model is not publicly available). This is different from our approach, which aims to retrieve both direct answers and contextual information for a question. This allows our approach to produce more faithful long-form answers, enhancing the groundedness to the retrieval results. Our experiments on a ConvQA dataset further show that our method works on non-ambiguous factoid questions without further training.

To the best of our knowledge, there is no existing work that optimizes the retrieval system towards contextual information, as required for generating comprehensive long-form answers.

Iterative retrieval-augmentation. A new line of work extends RAG pipelines (Lewis et al., 2020) to multiple rounds of retrieval and generation (Jiang et al., 2023; Shao et al., 2023; Yao et al., 2022).

FLARE (Jiang et al., 2023) iteratively generates an upcoming next sentence, uses the generated sentence as query for retrieval, and then generates the actual next sentence based on the retrieval results. We compare against FLARE in Table 2, and show that our approach can produce more suitable long-form answers. IterRetGen (Shao et al., 2023) first follows the standard RAG approach, but then iteratively adds rounds of retrieval and generation to refine the initially generated text. Their experiments are conducted on datasets with crisp short-form answers only. ReAct (Yao et al., 2022) and Toolformer (Schick et al., 2023) provide the LLM with specific actions that can trigger retrieval for a LLM-generated query. The LLM itself can then generate relevant queries, and ground subsequent generations on the retrieval results. ReAct is implemented based on in-context learning (Brown et al., 2020), assuming strong instruction-following capabilities for the LLM. Their main target are reasoning tasks, in which the LLM interacts with an environment to predict a sequence of actions. The Toolformer makes use of LLM-perplexity for identifying relevant calls of tools (such as retrieval with a specific query). We investigate the underlying idea in this work (Table 1) for identifying relevant silver passages.

Note that these approaches, by design, employ multiple rounds of generation and retrieval, making them intractable in many real-world scenarios in which users expect an answer within a few seconds (at most). Further, iterative retrieval and generation can lead to extremely long prompts, as previous retrieval and generation results are often retained as context for the LLM.

Our approach aims to retrieve the most relevant information in *one shot*, and then ground the answer on these one-time retrieval results.