SUMMER SCHOOL 2023

Forward Looking Active Retrieval Augmented Generation

Lukas Zeit-Altpeter

Friedrich Schiller University Jena

lukas.zeit-altpeter@uni-jena.de, Matrix: @lukaszett:ag-link.xyz

ai.uni-jena.de

INTRODUCTION

Question Answering (QA) is an active field of research in the disciplines of Natural Language Processing and Information Retrieval. QA-systems are tasked with providing a factually correct answer for a given question. Common approaches for QA can be grouped into three categories [1]:

1. Systems that make use of structured data (e.g. knowledge graphs): Abacha and Zweigenbaum [2] parse medical abstracts to extract semantic relations such as "cures", "causes" etc. before parsing a question and retrieving a fitting answer from the constructed knowledge graph.
2. Extraction based systems that extract answer spans from a given context. Chen et al. [3] train a recurrent neural network to identify answer-spans for given questions from Wikipedia articles.
3. Generative large language models (LLMs) that generate an answer when given the question as prompt. Google's MED-PALM (and MED-PALM 2) [4] finetune an instruction-tuned large language model on trustworthy medical information.

Long-form answers are generally preferable for open-ended questions, and Lewis et al. [5] note that using generative systems can outperform extractive methods due to knowledge being encoded within a model's weights. With this paper I want to present a pre-print of a paper that combines retrieval of relevant documents to the given question with generative LLMs. The main question that this paper aims to answer can be condensed into: **At which point during the answer generation does the system retrieve which documents?** With my poster I present the work presented in Jiang et al. [6]: All figures, experiments and results are taken directly from the paper and only changed for the sake of layout.

METHOD

A common problem when using generative LLMs is that of 'hallucination': A confident sounding answer is generated that states misinformation. One approach to mitigate the risk of this happening is retrieval augmented generation. Alongside the prompt, the model is presented with relevant context.

Usually, the context is retrieved using a search engine using the question as query. While the possibility of hallucination still exists, this method has been empirically shown to significantly increase factual accuracy for QA tasks [5]. Especially for long form questions retrieving relevant information once might not be sufficient. Jiang et al. [6] introduce the method of Forward Looking Active Retrieval (FLARE) based answer generation. Figure 1 provides an overview on how they approach works.

Generative language models work by predicting following (or intermediate) tokens based on a given prompt. Alongside the prediction the model also outputs its confidence in this specific decision. For QA-tasks it has been shown [7] that a low token confidence is a good indicator for missing factual knowledge. FLARE makes use of this to identify the need for re-retrieval of context: Initially, the model prompts a generative LLM (the approach itself is agnostic to the specific model being used) alongside context retrieved for the given question. The model then generates text and is stopped when a sentence with a token-confidence lower than a threshold θ is generated. The last generated sentence (including the low-confidence token) is then used to generate a query to retrieve new context.

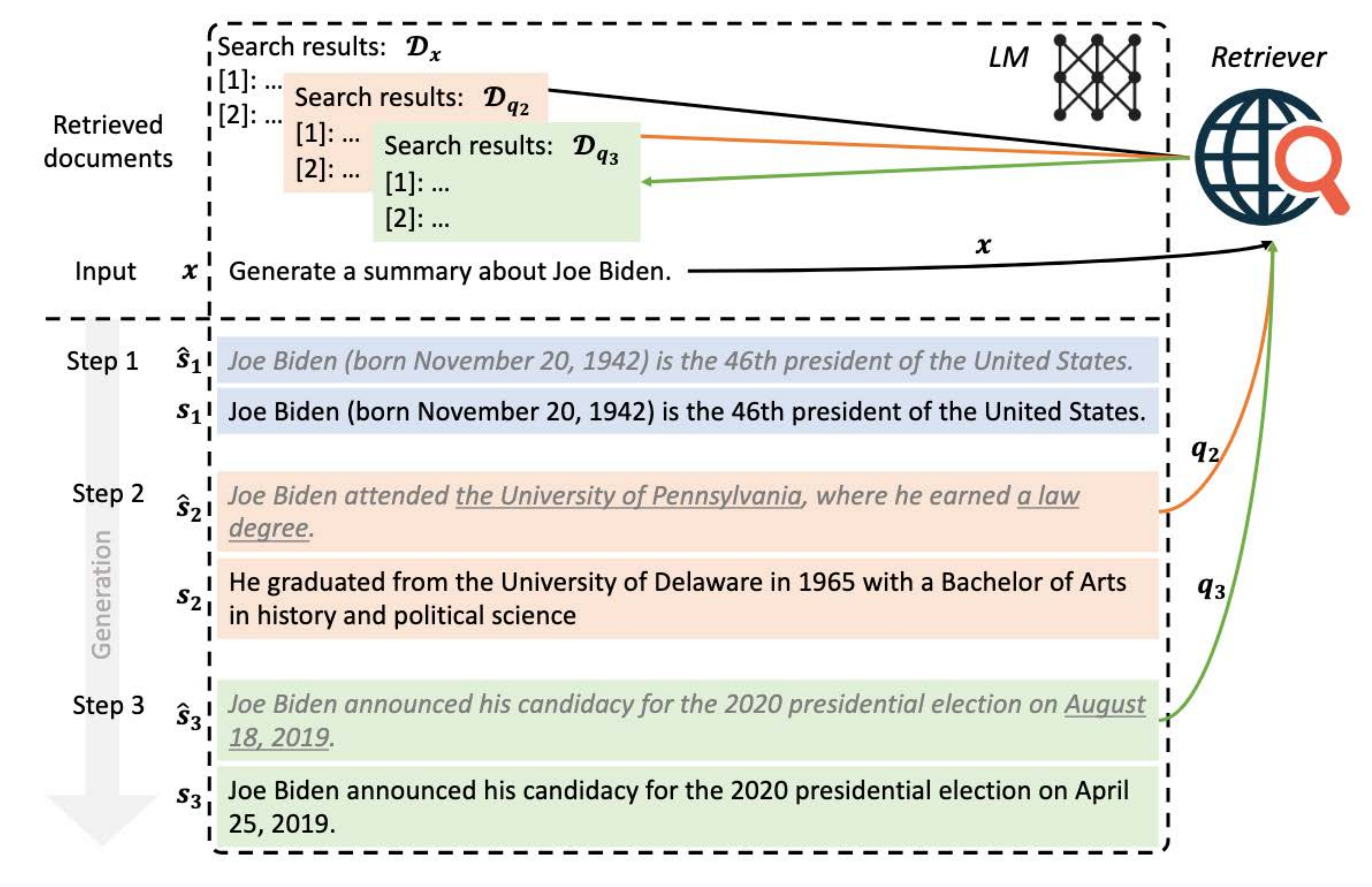
Search results: $\mathcal{D}_x$
[1]: …
[2]: …
Search results: $\mathcal{D}_{q_2}$
[1]: …
[2]: …
Search results: $\mathcal{D}_{q_3}$
[1]: …
[2]: …
LM
Retriever
Retrieved documents
Input
x
Generate a summary about Joe Biden.
x
Step 1
$\hat{s}_1$
Joe Biden (born November 20, 1942) is the 46th president of the United States.
s_1
Joe Biden (born November 20, 1942) is the 46th president of the United States.
q_2
Step 2
$\hat{s}_2$
Joe Biden attended the University of Pennsylvania, where he earned a law degree.
s_2
He graduated from the University of Delaware in 1965 with a Bachelor of Arts in history and political science
q_3
Generation
Step 3
$\hat{s}_3$
Joe Biden announced his candidacy for the 2020 presidential election on August 18, 2019.
s_3
Joe Biden announced his candidacy for the 2020 presidential election on April 25, 2019.

Figure 1: The process of FLARE augmented generation: A user inputs x and retrieves Documents D_x. $\hat{d}_i$ are generated answers with low-confidence tokens underlined. The system then retrieves documents D_{q_i} and generates the new sentence s_i.

EXPERIMENTS AND RESULTS

The approach is evaluated using four existing datasets for QA or QA-related tasks. Baselines methods use either (1) no retrieval, (2) context only given once at the start of generation or (3) retrieval at fixed intervals using the last generated sentence as query. All inference was done using OpenAI's "off-the-shelve" blackbox text-davinci-003 model. The BM25-method was used to retrieve relevant context from Wikipedia Performance as shown above is measured using the exact match (EM) metric. This indicated whether the provided short answer from the ground truth dataset (usually only 1-3 words) is present in the generated answer.

Four different datasets are used for evaluation:
2WikiMultiHopQA contains questions such as "When did the director of film Hypocrite (Film) die?" that require multiple follow-up questions to be answered. In this case, the system should first find out who is the film's director before retrieving their date of death.
StrategyQA is a dataset that tests a system's common sense reasoning. It features yes/no questions such as "Would a pear sink in water?".
ASQA consists of questions that have more than one possible interpretation e.g. "What is the criminal's name in the breakfast club?". This question could have both the actor's as well as the character's name as an answer. Additionally the ASQA-hint dataset is used that adds a hint on how the question might be ambiguous.
WikiASP is used to evaluate summarization skills. It contains tasks such as "Generate a summary about Echo School (Oregon) including the following aspects: academics, history" alongside relevant Wikipedia articles (for these experiments, the provided articles where not used).

Figure 2 shows that FLARE outperforms all baseline methods on all tested datasets by a varying margin. Table 1 offers a comparison of two tested query formulation methods for the retrieval step.

Table 1: Evaluation of query formulation methods. The implicit method takes the last generated sentence, removes low confidence tokens and uses the result as query. The explicit method masks low confidence tokens and asks an LLM to generate a question that yields the masked tokens as answer.

	ASQA-hint			
	EM	D-F$_1$	R-L	DR
Implicit	45.7	36.9	37.7	37.3
Explicit	46.2	36.7	37.7	37.2

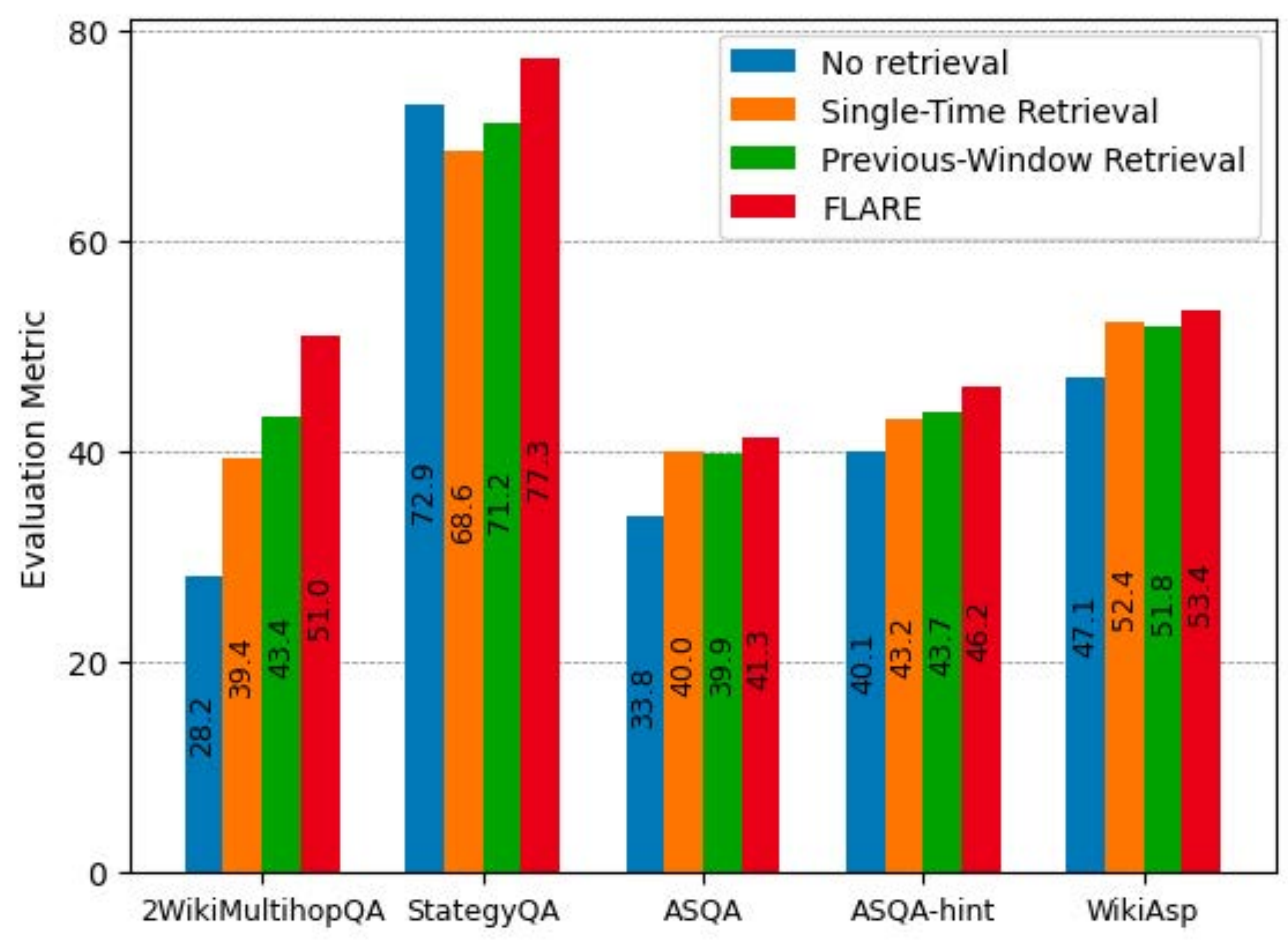

Figure 2: Evaluation of FLARE across different datasets, Exact Match is used for all datasets apart from WikiASP – for this, UniEval is used.

CONCLUSIONS AND LIMITATIONS

- Jiang et al. [6] show that re-retrieving context for generative QA systems benefits from using confidence to determine when and what to retrieve.
- There are still open questions regarding the selection of an adequate confidence threshold θ or different retrieval strategies.
- The approach was evaluated across a wide variety of datasets, performance on a classical long-form QA-dataset was not tested.
- An additional limitation that has to be kept in mind is that the results were generated using an off-the-shelve but closed-down language model and where thus difficult to reproduce.

References

[1] Singh, S., Susan, S. (2023). Healthcare Question–Answering System: Trends and Perspectives. In: Jain, S., Groppe, S., Mihindukulasooriya, N. (eds) Proceedings of the International Health Informatics Conference. Lecture Notes in Electrical Engineering, vol 990. Springer, Singapor
[2] Abacha, A.B., & Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. Inf. Process. Manag., 51, 570-594. [3] Chen, D., Fisch, A., Weston, J., & Bordes, A. (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) (pp. 1870–1879). Association for Computational Linguistics.
[4] Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature (2023).
[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Proceedings of the 34th International Conference on Neural Information Processing Systems.
[6] Jiang, Z., Xu, F.F., Gao, L., Sun, Z., Liu, L.D., Dwivedi-Yu, J., Yang, Y., Callan, J., & Neubig, G. (2023). Active Retrieval Augmented Generation. ArXiv
[7] Zhengbao Jiang, Jun Araki, Haibo Ding, Graham Neubig; How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. Transactions of the Association for Computational Linguistics 2021; 9 962–977.