The impact of varying knowledge on Question-Answering system

Anh Nguyen Thach Ha
Department of Information Technology
FPT University
Ho Chi Minh City, VietNam
anhnthse173147@fpt.edu.vn

Trung Nguyen Quoc
Department of Information Technology
FPT University
Ho Chi Minh City, VietNam
trungng46@fpt.edu.vn

Tien Nguyen Van
Pythera AI
tien.nguyen@pythera.ai

Hieu Pham Trung
Pythera AI
hieu.pham@pythera.ai

Vinh Truong Hoang
Faculty of Information Technology
Ho Chi Minh City Open University
Ho Chi Minh City, VietNam
vinh.th@ou.edu.vn

Tuan Le-Viet
Faculty of Information Technology
Ho Chi Minh City Open University
Ho Chi Minh City, VietNam
tuan.lv@ou.edu.vn

***Abstract*—Scaling up huge language models to store massive amounts of knowledge in their parameters results in increased costs and training durations. Thus, in this work, we investigate the impacts of language models on increasing external knowledge and compare the performance of extractive and abstractive generating tasks when developing a question-and-answering system. To maintain consistency in our assessments, we update the MS MARCO and MASH-QA datasets by removing unnecessary support documents and increasing contextual relevance by mapping input questions to the nearest supported documents in our database structure. Finally, we evaluate performance in the health domain; our experience yields encouraging results not only for information retrieval but also for retrieval augmentation tasks, with the objective of boosting performance in future work.**

***Index Terms*—Extractive generation, Abstractive generation, Knowledge-based Question-Answering**

I. Introduction

Recent advances in chatbots and other language modeling approaches have resulted in outstanding performance in Natural Language Processing problems. However, producing long and meaningful phrases suffers from repetition, truncation, and hallucination [1], [2], which are typical in a generation of not only language modeling (LM) but also large language modeling. To mitigate this risk, [3] introduced ground-based information retrieval from external knowledge sources, and [4], [5] improved knowledge through LLM to increase the quality of replies with explicit query statements. Several approaches have been successfully compared to humans in terms of answering questions and picking up knowledge [6], [7]. However, their achievement is not practical for a case study because solutions emerge in many contexts when looking for external knowledge [8]. Therefore, the purpose of this research was to use several linguistic models to understand the context, extract external knowledge, and rewrite responses more smoothly with fewer hallucinations.

Long-Form Question Answering (LFQA) [9] introduces a task that generates detailed and explained answers to open-ended questions from Reddit forums 'Explain like I am five years old'. However, Krishna et al. [1] evaluated that at least 81% of the validation questions overlapped with the training/validation data, which led to training and inference bias. A survey on approaches to challenges in medical health introduces automatic question-answering, which has been successfully applied to various domains, such as search engines and chatbots [10]. The MASH-QA [11] is a publicly available large-scale benchmark dataset that differs from existing machine reading comprehension datasets with short single-span answers for question-answering. The MASH-QA answers were extracted from multiple spans within a long context document containing questions and knowledge articles from the consumer health domain. Recently, transformer encoder models, such as BERT [12] and RoBERTa [13] trained from a large corpus, have shown the best performance when adapted to specific tasks using transfer learning, among which is dense passage retrieval (DPR) [3], [14]. It has been demonstrated to outperform traditional sparse vector space models on several benchmarks by allowing the capture of semantic similarity and handling of lexical variations easily integrated with existing retrieval readers or retrieval generators to achieve state-of-the-art performance [3], [15]–[17]. In addition, the encoder-decoder models BART [18] and T5 [17] can be used for seq2seq tasks, such as machine translation, text summarization and question-answering tasks, to obtain promising results.

In this paper, we present a knowledge-based question-answering system using DPR [14] to retrieve external knowledge (in this study called support documents) and the Fusion-in-Decoder [17] takes the responsibility of generating the answer. There are two important observations in our proposal. 1) Large language models store vast amounts of

knowledge within their parameters, increasing the cost and training time. This is unnecessary for language models that require models to depend on support knowledge to control quality and reduce hallucination. 2) Using external document retrieval not only augments intrinsic knowledge, but also grounds model outputs in a knowledge source, providing interpretability in session A. The proposed methods can be summarized as follows.

- Building a knowledge-based question-answering system using a sparse transformer-based system comprising both long-form and short-form answers.
- The Fusion-in-Decoder architecture is better for extractive and abstractive generation tasks because it enhances external knowledge, which reduces the risk of hallucinations and smooths the response.
- We focus on mental healthcare using transfer learning entries on the ELI5, MS MARCO, and MASH-QA datasets. MASH-QA covers a broad biomedicine domain that covers clinical, biomedicine, consumer health, and examination.

The remainder of this article is structured as follows. Section II discusses relevant studies. The key concepts of the healthcare system and detailed implementation are explained in Section III. The results and results are then reported in Sections V-VI, respectively.

II. Related works

The key advantages of knowledge-based question-answering tasks include many benefits relative to use such as language modeling [3], augmented language modeling, and large language modeling [4], [5]. Digging deeper, Open Domain Question Answering introduced an architecture called "Retriever-Reader" to analyze the various systems that follow this architecture, as well as the specific techniques adopted in each component as extractive generation methods on the SQuAD, TriviaQA, Natural Questions, and MS MARCO datasets [19]–[22]. Fan et al. [9] introduced the 'Retriever Generation' for the abstractive task by augmenting external knowledge and their variant approach to the efficient transformer architecture [16], [23] in several datasets, such as ELI5, MASH-QA and SaaC [9], [11], [24]. Su et al. [25] designed an end-to-end framework for long-form question answering that combines relevant machine learning documents and extracts salient information prior to generating a paragraph-length answer that is faithful to the facts.

The most fundamental distinction between the SeeKer search module [26], sparse retrieval methods, such as TF-IDF [27], BM25 [28], or rewriting questions made clear context QReCC [29] is focused on exact word matching or/and frequency statistics that do not reflect the correct meaning of the input question when performing task retrieval. The DPR module [3], [14] learned the embeddings of questions and passages using a simple dual encoder framework. This allows the DPR model to capture semantic similarity and handle lexical variations better. Some recent techniques have also attempted to adapt knowledge by editing and tuning language model variants, providing a novel perspective for solving knowledge-intensive tasks by replacing document retrievers with large language model generators. Krishna et al. [1] have demonstrated the effectiveness of the Maximum Inner Product Search to retrieve Wikipedia articles relevant to a question using a transformer model with nearest-neighbor lookup. Borgeaud et al. [15] built large language model retrieval models over a database of trillions of tokens; however, this method has limitations relative to retrieval knowledge because the database is fixed, which means that it is not always up to date with the latest knowledge and current events.

Perhaps the closest to our work through experiments on the proposed method and several baselines is the KILT benchmark ELI5 dataset, which is a long-form question answering a strong abstractive task [30]. We use the 'natural language generation' competition track (NLGen v2.1) [22] of MS MARCO, where each query has a human-generated answer and requires the use of the most relevant given passage to create answers in a way in which it could be read from a smart speaker and make sense without any additional context as an extractive. Finally, the MASH-QA dataset [11] is a publicly available large-scale question-answer dataset with answers extracted from multiple spans in a long context document. The proposed model is based on questions and articles of knowledge in the consumer health domain. Human evaluation confirmed that the proposed framework can improve the quality of generation in terms of relevance and factual correctness. The detailed setup dataset is presented in session A.

III. Methodology

In the following section, we decompose the question-answering system into 2 modules: Retriever, which retrieves support documents using DPR [14]; Generator, which processes each question-document pair using the Fusion-in-Decoder [17] approach and then generates an answer. This structure is similar to the retriever-reader framework that was first introduced in DrQA [31], but instead of using a reader, we train the language model using Fusion-in-Decode as a generator, and these two modules can be developed independently. The framework is illustrated in Figure 1. The top panel shows the process of building a knowledge base with data from a Wikipedia dump in our experiment. The knowledge base comprises the text and embedding representation of each knowledge passage. The FAISS is then used to create an index to speed up querying of support documents. To query this, the input question must be encoded in the embedding vector. The similarity score with support documents in clusters that belong to some knowledge passages. The bottom panel presents how the returned support documents are paired and calculated using the question in the Fusion-in-Decoder approach. An input question is paired with each knowledge returned from the knowledge base. All questions are passed to the encoder of

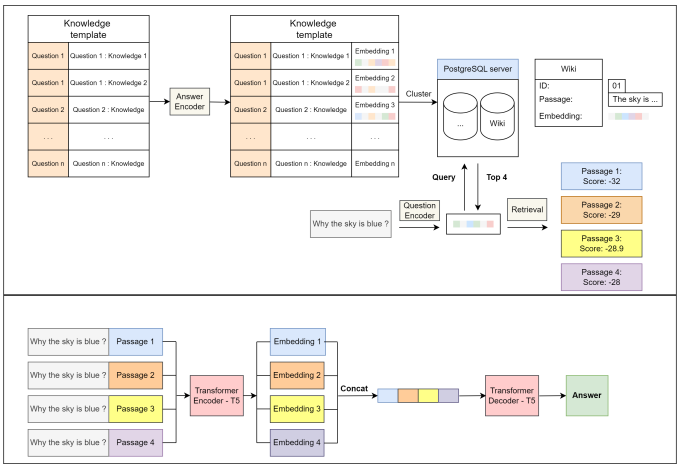

Fig. 1. FiD architecture.

TABLE I
THE DETAILS OF QUESTION-ANSWERING EXTRACTIVE AND ABSTRACTIVE DATASETS.

Dataset	Average Length			Size			
	Question	Answer	Document	Train	Dev	Test	Total
ELI5	42.2	130.6	97.6	272,634	1,507	600	274,741
MS MARCO	6.0	13.1	93.5	453,033	2,540	1,669	457,242
MASH-QA	8.8	61.7	98.7	12,115	1,596	1,644	15,355

the backbone model's encoder, and corresponding semantic embedding vectors are generated. These embeddings are concatenated into one, passed to the decoder, and then generated as the response.

A. Retriever

To access support documents, we use DPR to manage the Retriever module. The documents and questions are encoded as dense vector representations that are calculated using two BERT models-one for encoding questions called $DPR_q(.)$, and another for encoding documents called $DPR_d(.)$. We have N knowledge passages stored in the database, denoted as $\{d_1, d_2, ..., d_N\}$, and represented by allow-dimensional vector embedding as $E_d \in R^{N \times D_r}$ where D_r is the hidden dimension

$$E_{di} = DPR_d(d_i) \tag{1}$$

where $i \in \{1, 2, ..., N\}$.

For the input question q, $DPR_q(.)$ converts the string-based question to a low-dimensional vector embedding as well, E_q:

$$E_q = DPR_q(q) \tag{2}$$

We use FAISS [32] to speed up the retrieval of support documents. The retrieval process was performed using Approximate Nearest Neighbors. We rank and select the most relevant knowledge passages by calculating the dot product of the two vector representations E_q and E_{di}, which serve as the retrieval scores.

B. Generator

The retriever-reader approach comprises two stages. However, in the second stage, the Retriever generator generates free text directly to answer the question. We train the Generator to produce a human-like response with clear grammar and expression. The generator does not simply extract start or end positions from a retrieved passage, nor does it generate an answer based on pieces of information already available in knowledge passages like the original FiD.

The Fusion-in-Decoder approach is also based on a pre-trained T5 [33]. With the Fusion-in-Decoder, we can use more knowledge passages without truncating them due to the token limit of any language model when concatenating all knowledge passages and a question to a single input. For each knowledge passage recovered by the retriever, a question is paired, processed independently, combined, and then pushed to the encoder. Processing passages independently in the decoder allows us to parallelize the computation. A question and relevant knowledge passages are separated by special prefixes such as 'question:' and 'context:'.

Given a set of retrieved knowledge passages $\{k_1, k_2, ..., k_{N_\alpha}\}$ where $N_\alpha << N$. D_g is the hidden dimension of each token embedding, and the proposed T5 backbone has L Encoder and Decoder layers. Each input string is calculated by:

$$I = f_{tokenize}(q + k_i) \tag{3}$$

$$W_i^{(0)} = f_{emb}(I) \tag{4}$$

, where $I \in R^S$

$$W_i^{(l)} = f_{t5-enc}(W_i^{(0)}) \tag{5}$$

where $W_i \in R^{S \times D_g}$, S is the maximum length of the input sequence. The tokenized input vector is an embedding representation computed by $f_emb(.)$ and encoded by multiple t5 encoder layers, f_{t5-enc}. Then, the output of the last layer of the encoder is input to the decoder, $f_{t5-dec}(.)$, to compute cross attention and generate the hidden state of the answer V:

$$V = f_{t5-dec}([W_1^{(l)}; W_2^{(l)}; ...; W_{N_\alpha}^{(l)}]) \tag{6}$$

IV. Experiments

A. Datasets

ELI5: We used the ELI5 dataset in the KILT benchmark [30], the KILT version changed the knowledge source from common Crawl to a fixed Wikipedia snapshot on August 01, 2019, which includes 5.9M articles. The total number of samples in the training, validation, and test sets was 272,634, 1,507, and 600, respectively, with an average of 42.2 and 130.6 questions and answers.

MS MARCO: provides a realistic setting for natural language-understanding research and covers diverse topics and domains [22]. The Question Answering and Natural Language Generation task requires using the most relevant passage to create answers "in a way in which it could be read from a smart speaker and make sense without any additional context". The method used to create this dataset included real Bing questions and human-generated answers. The queries were sampled from anonymized user logs, and the answers were generated by human annotators based on relevant web passages. In this case, we modified the entire dataset to satisfy our constraint with training, validation, and testing sets of 453,033, 2,540, and 1,669 samples, respectively.

MASH-QA: The dataset was created by collecting consumer healthcare queries from a commercial search engine and matching these queries with relevant knowledge articles from a health website [11]. The dataset contains 34,808 question-answer pairs and 5,574 documents with answers of 67.2 words on average length, which demonstrates that this dataset serves well for long-form question-answering tasks. MASH-QA was originally used for extractive question-answering tasks. However, we approached this differently by generating answers to the questions using supporting documents from the system's external knowledge.

B. System and parameter settings

We used the ELI5 dataset following the instructions of the KILT benchmark kit and processed MS MARCO and MASH-QA, as mentioned in Section A which uses Wikipedia dump as the source of the documents. We then used DPR to retrieve passages in all datasets. For each question, we retrieved 30 documents and set their maximum length to 250 words. Our main training started with the pre-trained T5 model weight available in the Hugging Face Transformers library, following previous research and fine-tuning the models on each dataset independently using the Adam optimizer and a learning rate of 10^{-4} (effective batch size is 64 and 128, respectively, on abstractive and extractive generation tasks). We evaluated the models every 500 steps using beam search with a beam size of 4 and set a maximum answer length of 200 words.

V. Experiment Results

TABLE II
COMPARISON OF OUR MODEL WITH SEVERAL METHODS ON A DEV AND TEST SET OF KILT ELI5 (THE BOLD DENOTES THE OVERALL BEST PERFORMANCE).

Models	dev		test	
	ROUGE-L	**F1**	**ROUGE-L**	**F1**
T5 [30]	21.02	18.36	19.08	16.10
BART [30]	22.69	22.19	20.55	19.23
DPR+BART [30]	17.41	17.88	17.41	17.88
RAG [3]	16.11	17.24	15.50	17.10
RT+c-REAM [1]	24.40	25.60	23.20	22.90
RBG [25]	24.46	**29.04**	**24.72**	**27.52**
Ours	**29.37**	27.78	22.68	24.48

This paper evaluates the performance of a text generation system using ROUGE-L, BLEU-1, and F1 scores. ROUGE-L measures the similarity between sequences, with higher scores indicating better results. BLEU-1 measures the percentage of words that match machine output and a reference answer, with higher scores indicating greater similarity. F1 measures the accuracy of the generated answer compared to normalized Unigrams. These metrics provide a comprehensive evaluation of the system's effectiveness in generating high-quality text.

TABLE III
PERFORMANCE COMPARISON BETWEEN THE EXTRACTIVE (MS MARCO, MASH-QA) AND ABSTRACTIVE (ELI5) TEST SETS WITH 30 SUPPORTING DOCUMENTATION FOR EACH SAMPLE. WE AIM TO PROVIDE A COMPREHENSIVE EVALUATION USING MULTIPLE METRICS TO ASSESS THE SIMILARITY AND QUALITY OF RESPONSE GENERATION. BASED ON THESE RESULTS, THE MS MARCO DATASET SHOWED IMPROVED PERFORMANCE WITH THE MASH-QA ALGORITHM NOT ONLY FOR BLEU1, BUT ALSO FOR ROUGE-L AND F1 SCORES BY MORE THAN 10%. TO BALANCE THE EXTRACTIVE AND ABSTRACTIVE METHODS, WE ALSO EVALUATED THE ELI5 DATASET WITH ROUGE-L AND F1 SCORES OF 22.68% AND 24.48%, RESPECTIVELY.

	ROUGE-L	**F1**	**BLEU1**
MS MARCO	37.21	37.64	27.50
MASH-QA	20.69	24.60	26.63
ELI5	22.68	24.48	-

Our experiment involved the use of three different datasets and we observed varying levels of success across these datasets. First, we compare the abstractive and extractive generation performance of the proposed system in Table II and Table III. Then we compared the effect of the number of documents on the results in Section B and the abstractive and extractive generation with related and unrelated support

documents in Section C. The performance of abstractive generation Table II shows that our approach ROUGE-L 29.37% and the F1-score 27.78% in the dev set improved performance with many methods while allowing a customized database depending on our system using the dummy Wikidataset. Furthermore, in the test set, the proposed approach maintains the original FAISS template format, and the experimental results are lower than RBG [25] in the F1 score 27.52% and 24.48%, and improve performance 1.6% with the [1] F1 score 22.9% and 24.48%. In addition, our strategic focus is on abstraction of expectations gaps and extraction generation, in Table III when transferring a pretrained learning language model to specific tasks with mixed performance training. The results of the MS MARCO dataset show values for 37.21% ROUGE-L, 27.50% BLEU-1, and the F1 score 37.64% compared to other datasets, which is likely due to several factors, including the use of older model architecture, limited computing power, changes in test data, and differences in supporting documents. Similarly, the MASH-QA dataset did not yield the expected favorable outcomes, and this study represents the first attempt to use this dataset for generative question-answering purposes. Here, results 20.69%, 24.60%, and 26.63%, respectively.

We observed more positive results for transfer learning performance, which was characterized by several advantages for the ELI5 dataset. The supporting documents used in this dataset were of the same type as those used in the original model, reducing variability and improving model performance. In addition, we took advantage of information retrieval through several processing steps to enhance the efficiency of the analysis process and ensure greater consistency in the results of MASH-QA and MS MARCO. The proposal can query any number of documents, facilitating data exploration and refinement of the approach.

VI. Conclusion and Future Work

In this study, we evaluated the ability of the LM to perform extractive and abstractive generation tasks by enhancing external knowledge, thus reducing the risk of hallucinations and smoothing responses. In addition, we carefully evaluated the effect of relevant and irrelevant support documents on the question. Table III summarizes the result of all our modified, ELI5 and MASH-QA correspondences between model accuracy and support documents for different generation tasks. MS MARCO trained twice with different numbers of samples improved performance by 10% when the same hyperparameters were used throughout the training phase.

We believe that the experimental results are flawed in several cases. However, the knowledge gained will help the research community better exploit this field. Instruction tuning with meta-learning has demonstrated strong performance on natural language generation tasks, as applied in conversational AI, which trains and evaluates in a dialog context to improve multitasking, which is our future work.

References

[1] K. Krishna, A. Roy, and M. Iyyer, "Hurdles to Progress in Long-form Question Answering," arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.48550/arxiv.2103.06332.

[2] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, Nov. 2022, doi: https://doi.org/10.1145/3571730.

[3] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv.org, Apr. 12, 2021. https://arxiv.org/abs/2005.11401

[4] W. Yu et al., "Generate rather than Retrieve: Large Language Models are Strong Context Generators," arXiv (Cornell University), Jan. 2022, doi: https://doi.org/10.48550/arxiv.2209.10063.

[5] S. Min et al., "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?," arXiv (Cornell University), Feb. 2022, doi: https://doi.org/10.48550/arxiv.2202.12837.

[6] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[7] M. Li, B. Peng, J. Gao, and Z. Zhang, "OPERA: Harmonizing Task-Oriented Dialogs and Information Seeking Experience," arXiv (Cornell University), Jan. 2022, doi: https://doi.org/10.48550/arxiv.2206.12449.

[8] S. Feng, Siva Sankalp Patel, H. Wan, and S. Joshi, "MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Jan. 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.498.

[9] A. Fan, Yacine Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli, "ELI5: Long Form Question Answering," arXiv (Cornell University), Jul. 2019, doi: https://doi.org/10.48550/arxiv.1907.09190.

[10] Q. Jin et al., "Biomedical Question Answering: A Survey of Approaches and Challenges," ACM Computing Surveys, vol. 55, no. 2, pp. 1–36, Mar. 2023, doi: https://doi.org/10.1145/3490238.

[11] M. Zhu, A. Ahuja, D.-C. Juan, W. Wei, and C. K. Reddy, "Question Answering with Long Multiple-Span Answers," ACLWeb, Nov. 01, 2020. https://aclanthology.org/2020.findings-emnlp.342/

[12] A. Vaswani et al., "Attention Is All You Need," arXiv.org, Dec. 05, 2017. https://arxiv.org/abs/1706.03762

[13] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv (Cornell University), vol. 1, Jul. 2019, doi: https://doi.org/10.48550/arxiv.1907.11692.

[14] Vladimir Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," Apr. 2020, doi: https://doi.org/10.48550/arxiv.2004.04906.

[15] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," arXiv.org, Feb. 07, 2022. https://arxiv.org/abs/2112.04426 (accessed Nov. 23, 2023).

[16] S. Hofstätter, J. Chen, K. Raman, and H. Zamani, "FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation," arXiv (Cornell University), Sep. 2022, doi: https://doi.org/10.48550/arxiv.2209.14290.

[17] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," arXiv.org, Feb. 03, 2021. https://arxiv.org/abs/2007.01282v2 (accessed Aug. 19, 2023).

[18] M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," arXiv (Cornell University), Oct. 2019, doi: https://doi.org/10.48550/arxiv.1910.13461.

[19] Pranav Rajpurkar, J. Zhang, Konstantin Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," Jun. 2016, doi: https://doi.org/10.48550/arxiv.1606.05250.

[20] T. Kwiatkowski et al., "Natural Questions: A Benchmark for Question Answering Research," Transactions of the Association for Computational Linguistics, vol. 7, pp. 453–466, Mar. 2019, doi: https://doi.org/10.1162/tacl_a_00276.

[21] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," arXiv.org, May 13, 2017. https://arxiv.org/abs/1705.03551 (accessed Aug. 19, 2023).

[22] P. Bajaj et al., "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," arXiv (Cornell University), Jan. 2016, doi: https://doi.org/10.48550/arxiv.1611.09268.
[23] de Jong et al., "FiDO: Fusion-in-Decoder optimized for stronger performance and faster inference," arXiv (Cornell University), Jan. 2022, doi: https://doi.org/10.48550/arxiv.2212.08153.
[24] P. Ren, Z. Chen, Z. Ren, Evangelos Kanoulas, C. Monz, and Maarten de Rijke, "Conversations with Search Engines: SERP-based Conversational Response Generation," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2004.14162.
[25] D. Su et al., "Read before Generate! Faithful Long Form Question Answering with Machine Reading," arXiv (Cornell University), Jan. 2022, doi: https://doi.org/10.48550/arxiv.2203.00343.
[26] K. Shuster, Mojtaba Komeili, L. Adolphs, S. Roller, A. Szlam, and J. Weston, "Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion," arXiv (Cornell University), Mar. 2022, doi: https://doi.org/10.48550/arxiv.2203.13224.
[27] B. Das and S. Chakraborty, "An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation," arXiv (Cornell University), Jun. 2018, doi: https://doi.org/10.48550/arxiv.1806.06407.
[28] S. Robertson, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends® in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2010, doi: https://doi.org/10.1561/1500000019.
[29] R. Anantha, S. Vakulenko, Z. Tu, S. Longpre, S. Pulman, and S. Chappidi, "Open-Domain Question Answering Goes Conversational via Question Rewriting," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2010.04898.
[30] F. Petroni et al., "KILT: a Benchmark for Knowledge Intensive Language Tasks," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.48550/arxiv.2009.02252.
[31] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to Answer Open-Domain Questions," Mar. 2017, doi: https://doi.org/10.48550/arxiv.1704.00051.
[32] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," arXiv.org, Feb. 28, 2017. https://arxiv.org/abs/1702.08734.pdf (accessed Nov. 30, 2023).
[33] C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv (Cornell University), Oct. 2019, doi: https://doi.org/10.48550/arxiv.1910.10683.

APPENDIX

A. Dataset details

The experiment used the original ELI5 benchmark setup for performance comparison on the KILT leaderboard. However, the original MS MARCO and MASH-QA datasets were modified to meet specific constraints. These included filtering MS MARCO from samples with no answers, mapping from input questions to closest supported documents, and enhancing knowledge-based use of DPR system's external knowledge. A 0.05% of the processed training set was randomly split into a new validation set, and the number of samples was halved from 34,808 to 15,355. Each question in both datasets had the same number of supporting documents.

B. Effect of Number of Documents

In this study, we investigate the impact of varying the number of documents on the performance of a generative question-answering system. The experiment presented in this section comprises two distinct parts. During the first part, we altered only the number of supporting documents included in the training loop while keeping the other parameters constant. In the second part, we only varied the number of documents used during the inference phase while keeping all other parameters constant.

TABLE IV
RESULTS OF THE MASH-QA DEV SET WITH DIFFERENT NUMBERS OF DOCUMENTS (THE BOLD DENOTES THE OVERALL BEST PERFORMANCE).

	5	**10**	**15**	**20**	**25**	**30**
F1	22.30	23.70	24.09	24.10	24.46	**24.61**
BLEU-1	24.81	25.43	25.72	25.70	26.10	**26.19**
ROUGE-L	17.83	18.37	18.43	18.47	18.55	**18.74**

To perform our first analysis, we used the MASH-QA dataset and fine-tuned the model with incremental increases of 5 documents, starting from 5 documents and gradually increasing up to 30 documents. Although most of the hyperparameters in both the training and inference processes were kept the same as described in Section B in Experiments, we limited the training loop to 10,000 gradient steps per training and increased the number of documents in each training loop. The experimental results are presented in Table IV.

We employed the best weight from Section III in our second analysis and present the result in Table V. This setup revealed a positive relationship between supporting documents and ROUGE-L, indicating that as the number of documents increased, the ROUGE-L score also increased. This finding is consistent with previous research that also demonstrated a positive correlation between supporting documents and the metric. For example, a study by [17] found that the exact match also increased as the number of documents increased. Similarly, [25] reported a similar positive relationship between the number of documents and the ROUGE-L score. In general, our results add to a growing body of evidence that the number of documents is a critical factor in the value of ROUGE-L, and further research should explore this relationship in depth.

TABLE V
RESULTS OF THE MASH-QA TEST WITH DIFFERENT NUMBERS OF DOCUMENTS (THE BOLD DENOTES THE OVERALL BEST PERFORMANCE).

	5	**10**	**15**	**20**	**25**	**30**
F1	24.59	26.12	24.60	24.60	24.40	**24.60**
BLEU-1	26.27	24.42	26.45	26.14	25.86	**26.63**
ROUGE-L	18.96	18.57	19.03	19.14	18.95	**20.69**

C. Abstraction and extraction task

The FiD output synthesizes supporting documents into a single answer, preserving coherence and logic while generating sentences to make the answer smoother. This binding approach ensures accuracy and yields high results due to the overlap of common subsequences in the generation and labeling sentences. The model's ability to synthesize information, read, and understand context is demonstrated, and some cases use extracted knowledge as keywords combined with training state parameters to generate answers without directly quoting from the documents.

TABLE VI

CHERRY-PICKED CASES IN THE FID EXTRACTIVE GENERATION TASK. WE USE YELLOW TO MARK OUT FACTUAL KNOWLEDGE SNIPPETS OF THE FID'S ABILITY TO SYNTHESIZE KNOWLEDGE, READ, AND UNDERSTAND THE CONTEXT TO REWRITE ANSWERS.

Question	How do I get checked for STDs without my parents knowing?
Answer	Every state allows teens to be tested and treated for STIs without their partner knowing. Talk with your doctor about your concerns and how to receive this type of care. If you're too embarrassed or worried about how your parents will react, it's important that you talk with an adult who can give you trusted advice, like your doctor.
Support docs	If you're 18 or older, your medical records are private according to the law. Any information that health staff receive or discuss with you at a clinic is completely confidential and cannot be shared with your parents or people you know without your permission
	While it's best to talk with your parents (or guardians), they might not be your first choice. If you are too embarrassed or worry about how your parents will react, it's important that you talk with an adult who can give you trusted advice, like your doctor
	You can and should talk with your doctor (or the office nurse) about ANYTHING and EVERYTHING. Sometimes your doctor will ask questions about school, your friends, and family members. Sometimes your doctor will ask you personal things like how you're feeling or what you like to do in your free time. The more your doctor knows about you, the better he or she is able to answer your questions or concerns.
	Every state allows for teens to be tested and treated for STIs without their parents knowing. Talk with your doctor about your concerns and how to receive this type of care.
	Your doctor will keep the details of what you talk about private or confidential. The only times when your doctor cannot honor your privacy is when someone is hurting you or you are going to hurt yourself or someone else. There are state laws that require doctors to share information when there is a concern about someone possibly getting hurt. If this happens, you and your doctor will talk about how to share any information necessary to keep everybody safe.
Question	What are the signs of a healthy relationship?
Answer	Healthy relationships involve honesty, trust, respect, and open communication between partners and they take effort and compromise from both people. There is no imbalance of power. Partners respect each other's independence, can make their own decision without fear of retribution or retaliation, and share decisions. If or when a relationship ends, there is no stalking or refusal to let the other partner go.
Support docs	Healthy relationships involve honesty, trust, respect, and open communication between partners and they take effort and compromise from both people. There is no imbalance of power. Partners respect each other's independence, can make their own decision without fear of retribution or retaliation, and share decisions. If or when a relationship ends, there is no stalking or refusal to let the other partner go.
	Healthy relationships are built on trust. And trust applies to fidelity, finances, parenting, and more. Trustworthy partners are predictable, faithful, and reliable.
	The best way to know that you're in a healthy relationship is to consider how your relationship makes you feel. Infatuation can feel exciting and drama can masquerade as passion, but true love feels safe, calm, and empowering. If your partner regularly shows that they care for you and support you, that's a good sign of a healthy relationship.
	Your friends and family support your relationship. What you think of your relationship is the most important opinion to consider, of course. But if the majority of people who love you and want the best for you think you're better off single, it's a good idea to take an impartial look at your relationship to determine why other people don't see it as a healthy one.
	In a healthy relationship, partners collaborate on decisions. Whether you're deciding something as mundane as what to have for dinner or something as momentous as where to live, you and your significant other should listen to each other and come to a mutually agreeable decision.

TABLE VII

CHERRY-PICKED CASES IN THE FID ABSTRACTIVE GENERATION TASK. WE USE YELLOW TO MARK OUT FACTUAL KNOWLEDGE SNIPPETS OF FID'S USE THE KNOWLEDGE EXTRACTED FROM THE SUPPORT DOCUMENTS AS KEYWORDS COMBINED WITH THE KNOWLEDGE STORED IN THE PARAMETERS IN THE TRAINING STATE THAT GENERATES ANSWERS COMPLETELY WITHOUT DIRECTLY QUOTING FROM THE DOCUMENTS.

Question	How do I convince my parents to stop telling me to act more girly and push gender roles onto me?
Answer	I don't know about you, but I'm pretty sure it's a lot easier to convince your parents to stop telling you to act more girly and push gender roles onto you. If you're a girl, you'll be able to build a brighter future for new generations by challenging gender roles and stereotypes.
Support docs	It is clear many young people are living in gendered environments, and are affected by traditional male or female stereotypes day in and day out. These perceived gender norms impact young people's well-being. Children who chose 'being tough' as the most important trait for boys, or 'having good clothes' as the most important trait for girls, are shown to have the lowest well-being across the group. On the other hand, children who chose 'working hard at school' as the most important quality scored highest for well-being.
	We all have the responsibility to build a more inclusive and accepting society. By challenging gender roles and stereotypes, we can build a brighter future for new generations.
	Girls are vulnerable, and boys are strong. How son's gender identity changed. The study calls it the hegemonic myth: the perception that men are the dominant sex, strong and independent, while women need to be protected. This idea starts in early childhood, reinforced by schools, parents and media.
	Once puberty hits, it's all about (preventing) sex. Puberty deepens the divide, especially when it comes to sexuality, turning boys into predators and girls into potential targets, the study found. Messages such as "don't sit like that," "don't wear that" and "boys will ruin your future" reinforce the gender division of power and promote sex segregation with the aim of preserving a girl's sexuality, the study says.
	Gender-nonconforming has consequences. A recurring theme in conversations with adolescents and adults is those who defied gender expectations, resulting in "significant sanctions and pressures to conform," the study said. The study focuses on three common manifestations of defiance: boys wearing nail polish, girls playing football or soccer, and gender-"inappropriate" clothing. For young children, it may start as acceptable to cross gender boundaries.
Question	How to convince your parents when you have a lover?
Answer	If everyone in the family has comments and misjudges about your lover, you should create a meeting at home or at a restaurant that is full of both parties on a beautiful evening. People talk together, maybe the misunderstanding will be resolved when the two sides understand each other better. Remember not to let the conversation last too long, and do not forget to suggest to your lover to show his humor and charming talk.
Support docs	You can share and make sure to show your parents that your relationship will not affect your studies/work and have a positive impact on your quality of life. Remember not to let the conversation last too long, no matter how your parents react, stay calm and patiently prove to them that you and your partner have enough knowledge in love and sex-related issues for your parents' peace of mind about both.
	Parents always have standards for their children and are more and more strict with the partner\" you choose, the source of this insecurity comes from love. So to share this issue, it is most necessary to trust and you have to prove that both of you have enough understanding to have a healthy relationship.
	Convincing parents does not stop at words but also requires specific actions, which is a process of building trust. Make sure you know enough about love and sex for a relationship and become a connection point so that parents and your lover can understand and create sympathy for each other.
	Parents' intervention in their children's love comes from the mentality of always wanting the best for their children. In order not to accidentally become a cause of distance between parents and children, let's sit down together to share the views of both sides. That's the best way to show respect for each other.
	Instead of imposing a restriction with the thought "for the good of the child\", sit down and talk and listen to your child's wishes. At that time, you will understand your child's mind and why the child believes in their love. From there, you can give sincere advice to positively influence your child's love.