ZK-GenMed: A Zero-shot Knowledge Generative Medical Large Language Model

Abstract

Advancements in Natural Language Processing (NLP) have led to the development of Large Language Models (LLMs), which have demonstrated remarkable capabilities in various tasks, domains, and settings. These models have demonstrated efficacy in various training and evaluation scenarios, including zero-shot learning and instruction settings. They have been effectively applied to tasks including reasoning, summarizing, and answering questions. Moreover, LLMs have been used in a variety of industries, including the medical profession, where they have been used to jobs requiring accuracy, such answering questions. However, much research hasn't been done on LLMs' potential for resolving medical questions in a zero-shot manner. In order to close this knowledge gap, we provide a novel framework called ZK-GenMed, which uses LLMs' advantages to produce the information needed for medical question answering in a zero-shot scenario. This framework combines the generated knowledge with ranking strategies to extract relevant information, enabling the model to answer medical questions meaningfully. Experimental results demonstrate significant improvements, with marginal gains of over 10% on various datasets, highlighting the potential of ZK-GenMed for medical question-answering applications.

1 Introduction

The emergence of extensive pre-trained language models (LLMs) has transformed the domain of natural language processing, allowing for zero-shot generalisation across a range of tasks, such as text classification, question answering, and text summarization (Sanh et al., 2022) (Brown et al., 2020). These models have performed remarkably in several areas, such as question-answering (QA) systems. Zero-shot approaches have effectively been utilised in QA systems to improve their performance in areas with a scarcity of labelled data. Recent research have introduced zero-shot open-book QA systems that employ retrievers to search for pertinent documents and extractors to identify answers within those documents, all without the need for domain-specific training data (Gholami and Noori, 2021). This technique has demonstrated potential when there is limited annotated data, such as in medical fields where data availability is frequently difficult.

Nevertheless, although LLMs possess remarkable powers, they are susceptible to hallucinations. Hallucinations are instances where wrong or illogical information is generated, which may seem believable but is actually factually incorrect or completely made up. The occurrence of hallucinations in present-day LLMs is extensively established, and several research has emphasised the frequency and significance of these mistakes. To overcome this constraint, scholars have suggested frameworks that combine knowledge graphs and employ hypothesis knowledge graphs to offer contextually pertinent information and enhance the dependability of medical question-answering solutions. The HyKGE architecture incorporates knowledge graphs to offer better-organised and precise answers, especially in intricate diagnostic situations. The authors of (Huang et al., 2023) highlight that relying solely on fine-tuning and post-processing methods is inadequate for effectively addressing hallucinations, especially in situations requiring high factual accuracy, such as medical diagnosis and treatment suggestions. Nevertheless, their practicality is constrained by data availability for pertinent situations.

This work introduces a new and innovative multi-step hybrid prompting approach. The technique is specifically developed to address different question-answering situations in LLMs while minimising the occurrence of hallucinations. Importantly, this technique does not rely on any external data. Our approach utilises multi-step deep dive prompting

techniques. This involves constructing a knowledge base by presenting multiple questions to LLM. The results of previous queries are combined using a ranking strategy to create deep prompts. These prompts are combined with an instruction-based approach to generate the final output. This strategy effectively tackles the limitations of current methods and offers a scalable way to improve the reliability of medical AI systems.

First, we explore the related work on question-answering in the medical domain, followed by a discussion of the proposed framework. Later, we will present our experimental setup and results of comprehensive testing and evaluation to demonstrate the effectiveness of our approach in significantly reducing hallucinations. Finally, we will discuss the improved safety and utility of LLMs in medical applications and outline directions for future work.

2 Related Work

Question answering (QA) has witnessed significant advancements over the years, driven by the availability of large datasets and the development of sophisticated natural language processing techniques. This section provides an overview of the key studies and methods that have contributed to the field, focusing on approaches that enhance accuracy, leverage zero-shot learning, and integrate external knowledge sources.

Several studies have explored the impact of multi-task training on zero-shot generalization in natural language processing tasks(Zhou et al., 2022). For instance, (Ma et al., 2021) demonstrates that training on a few key tasks can significantly boost performance, with QA tasks playing a crucial role. Other studies have investigated the role of knowledge sources, question-generation strategies, language models, and training regimes in medical QA (Zhao et al., 2022). Furthermore, researchers have developed techniques to generate verbalizers using embeddings automatically (Wang et al., 2023) and convert zero-shot learning into multiple-choice tasks (Gramopadhye et al., 2024). Similarly, (Teney and van den Hengel, 2016) proposed methods for zero-shot visual question answering (VQA), which highlight the importance of generalizing beyond the training examples to answer new questions about unseen images. These strategies underscore the potential of zero-shot techniques in extending the capabilities of QA systems across various domains. In the medical domain, few studies have explored zero-shot learning, including (Ji et al., 2023), which presents a COT-based verifier model with few-shot learning for response selection and highlights the importance of tailored prompting methods for accurate medical question-answering. Over the years medical QA has benefited from developing specialized datasets that facilitate the training and evaluation of QA models. Notable datasets include MedQA (USMLE)(Jin et al., 2020), MedMCQA(Pal et al., 2022), MMLU-Medicine(Hendrycks et al., 2021), and PubMedQA(Jin et al., 2019), which cover a wide range of medical topics and provide a benchmark for evaluating the performance of QA systems. One notable approach to improving medical QA involves the use of large pre-trained language models (LLMs) that are fine-tuned on medical datasets. The development of medical-specific LLMs, such as MedAlpaca (Han et al., 2023), OpenBioLLM (Ankit Pal, 2024), Mistral and Meditron (Chen et al., 2023),, has shown significant improvements in tasks such as question answering and treatment recommendation. These models leverage the vast amounts of medical literature and patient records to provide accurate and contextually relevant responses.

Despite the advancements in medical QA, challenges remain to be addressed. For instance, (Ji et al., 2023) identifies self-contradiction as a significant factor contributing to hallucinations in medical generative QA systems and proposes structured comparative reasoning and self-reflection methodologies to reduce hallucinations. Additionally, (Gao et al., 2023) introduces Retrofit Attribution using Research and Revision (RARR), a post-generation refinement technique that automatically aligns generated content with retrieved evidence, highlighting the need for real-time verification and validation processes to ensure the accuracy of AI-generated medical information. The comprehensive survey by (Tonmoy et al., 2024) outlines various feedback mechanisms and iterative refinement processes, such as the Decompose-and-Query framework and the EVER (Real-time Verification and Rectification) strategy, which attempt to detect and correct hallucinations during the generation phase. However, these limitations necessitate the development of more robust techniques to enhance the reliability of AI-generated medical information.

3 Methodology

This section discusses the Zero-shot techniques based on knowledge creation and instruction prompting, dataset, and models used. We begin with our proposed prompt techniques:

Algorithm 1 ZK-GenMed

function CALCULATEBARTSCORE(Q_o, Q_g)
    Calculate BARTScore for Q_o and each Q_g
    return $BARTScores$

function SELECTTOPQUESTIONS(gen_Q, BARTScores)
    Select top K questions with highest $BARTScores$
    return $topQuestions$

function GENERATEANSWERS(Q,option,KB)
    Prompt : "Consider the following Knowledge Base: {$answers$}Answer the Question {Q} and given {$options$}. What is the most appropriate option for the given Question?"
    answers = LLM(prompt)
    return answers

procedure (Answer generation)(Q, ops)
    $keyword_q \leftarrow$instructLLM(Q)
    $keyword_o \leftarrow$instructLLM(ops)
    $KB1 \leftarrow$instructLLM($keyword_q, keyword_o$)
    $gen_Q \leftarrow$instructLLM(question)
    $BARTScores \leftarrow CalculateBARTScore(Q, gen\_Q)$
    $topQuestions \leftarrow SelectTopQuestions(gen\_Q, BARTScores)$
    $KB \leftarrow topQuestions + KB1$
    $Answer \leftarrow GenerateAnswers(Q, ops, KB)$
    return Answer

This method 1 employs a hybrid approach. Initially, we identify the entity present in the original question and generate 10 related questions based on this entity. We then select the top 3 questions that are most similar to the original question and generate answers for them. The generated answers serve as the knowledge base for our final query to the LLM. The final query includes the original question, multiple-choice options, and the generated knowledge as context.

To identify the top 3 similar questions to the original question, we employ similarity measures including BARTScore. BARTScore is a metric based on the BART (Bidirectional and Auto-Regressive Transformers) model, which is pre-trained for sequence-to-sequence tasks such as summarization and translation. Given a sequence-to-sequence pre-trained model like BART, the BARTScore is calculated using the log probability of the target text given the revised text as the source text.

Let Q_o represent the sequence of tokens in the original question and Q_g represent the sequence of tokens in the generated related questions. The BARTScore is computed as follows:

$$BARTScore = \sum_{t=1}^{k} \log p(Q_{gt}|Q_{g(<t)}, Q_o, \theta)$$

where θ represents the parameters of the BART model. This score measures the semantic coverage between the original question and the generated questions.

Additionally, we use cosine similarity to further quantify the relationship between the original question and each generated question. Let Q represent the vectorized form of the original question and R_i represent the vector for the i-th generated question. The cosine similarity between the original question and each generated question is calculated as follows:

$$similarity(Q, R_i) = \frac{Q \cdot R_i}{\|Q\|\|R_i\|}$$

This similarity score helps in understanding how closely each generated question aligns with the core elements of the original question.

By combining BARTScore and cosine similarity, we rank the generated questions and select the top 3 most similar ones to the original question. These selected questions and their generated answers serve as the knowledge base for our final query to the LLM. This approach enhances the accuracy of the final responses and reduces the likelihood of hallucinations by ensuring the contextual relevance of the information considered.

4 Experimental Setup

In this study, we aimed to evaluate the effectiveness of our multi-step hybrid prompting technique in reducing hallucinations and enhancing the reliability of AI-generated medical information. Our experimental setup consisted of three main components: datasets, language models (LLMs), and prompt techniques.

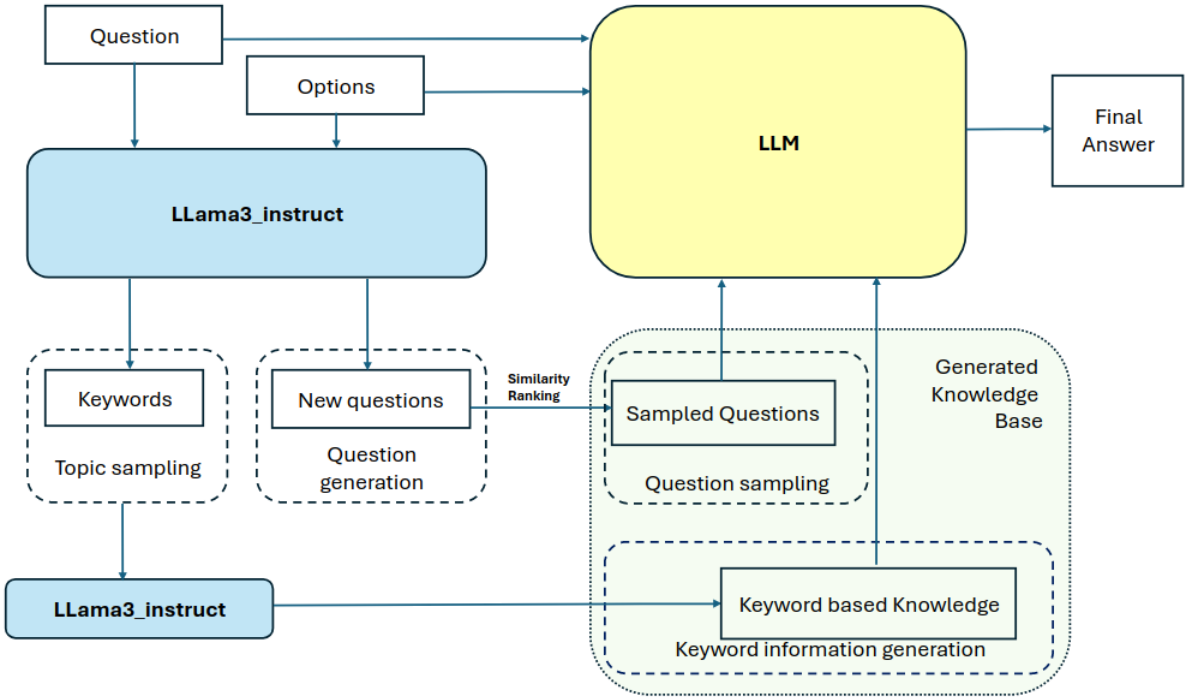

Figure 1: ZK-GenMed Framework: In the first step, The LLM generates the important keywords and similar questions. In step two, the KB is compiled, and later, it's used for the final question answering. In the architecture, Blue modules are frozen, while yellow modules can be finetuned.

4.1 Datasets

We utilized four diverse medical datasets to assess the performance of our technique comprehensively: **MedHALT**(Pal et al., 2023): This dataset includes three distinct tests to evaluate different aspects of model performance: The *False Confidence Test (FCT)* presents multiple-choice medical questions with a randomly suggested correct answer. The model evaluates the validity of the proposed answer and provides detailed explanations. It contains 95 questions. The *None of the Above Test (NOTA)* involves multiple-choice questions where the correct answer is replaced by 'None of the above'. The model must identify this and justify its selection. This test includes 18,865 questions. The *Fake Question Test (FAKE)* presents fake or nonsensical medical questions to determine if the model can correctly identify and handle such queries. This test contains 1,857 questions.
MedMCQA(Pal et al., 2022): This dataset consists of over 194k high-quality AIIMS and NEET PG entrance exam multiple-choice questions covering 2.4k healthcare topics and 21 medical subjects.
MedQA_USMLE(Jin et al., 2020): This dataset includes 12,723 4-way multiple-choice questions from practice tests for the United States Medical License Exams (USMLE), requiring biomedical and clinical knowledge.

4.2 Language Models (LLMs)

We selected multiple LLMs with varying sizes and capabilities to ensure a robust evaluation. These included **LLama3 Instruct - 8B**, **LLama3 - 8B**, **LLama2 chat - 7B** (Touvron et al., 2023), **LLama2 - 7B** (Touvron et al., 2023), **OpenBioLLM-8B** (Ankit Pal, 2024), **Mistral-7B**, **Mistral-7B-instruct**, **Meditron-7B** (Chen et al., 2023) and **Phi-3-mini-4k-instruct - 3.82B** (**?**). We used both the normal and instruct variants of LLama3 and the base and chat versions of LLama2 models to compare their performance under different prompting conditions.

4.3 Prompting Techniques

To assess the models, we used three different prompting techniques. In the first approach, **Normal Prompting**, the LLM is merely asked to select the appropriate option from a list of options. **Multi-Step Hybrid-1**, on the other hand, resembles the suggested method 1 but is marginally different in that it utilises the pertinent key entities as context rather than passing the same question. The preceding section's **Multi-step Hybrid Prompting**

method evaluates each alternative's applicability to the query first. After then, it makes the ultimate decision by combining all of the information at its disposal with additional guidelines.

4.4 Experimental Procedure

For each LLM and dataset combination, we conducted a thorough comparison of accuracy between the normal prompting and our advanced prompting technique. The evaluation focused on measuring the effectiveness of our technique in reducing hallucinations and improving the reliability of AI-generated medical information. The results of these experiments are presented and analyzed in the subsequent sections.

By leveraging a diverse set of datasets, advanced language models, and a rigorous evaluation framework, our study provides a comprehensive assessment of the proposed prompting technique's impact on mitigating hallucinations in medical applications.

Below is the Hyperparameter setting used for the experiments

- Number of Question selected after BARTScore : 5
- LLM setting
  - Number of Parameters : Parameters varies from 3.8B to 8B.
  - P value : 0.2
  - Top p : 0.9

Total computation time for the experiments was 500 hours on 40GB Nvidia-A100 GPUs.

5 Results

In this section, we present the evaluation results of different prompting techniques on various datasets using multiple LLMs. The objective is to assess the performance of each approach and provide insights into their effectiveness in different scenarios.

5.1 Evaluation Metrics

We used the following metrics to evaluate performance: **Accuracy, F1 Score, and Pointwise Score**.

The *Pointwise Score* is a detailed evaluation metric that accounts for both correct and incorrect predictions. Each correct prediction is awarded +1 point, while each incorrect prediction incurs a penalty of -0.25 points. The final Pointwise Score is an average of these individual scores. The formula for this is shown below.

$$S = \frac{1}{N} \sum_{i=1}^{N} (I(y_i = \hat{y}_i) \cdot P_c + I(y_i \neq \hat{y}_i) \cdot P_w) \tag{1}$$

where S is the final score, N is the total number of samples, y_i is the true label of the i-th sample, $\hat{y}_i$ is the predicted label of the i-th sample, $I(condition)$ is the indicator function that returns 1 if the condition is true and 0 otherwise, P_c is the points awarded for a correct prediction, and P_w is the points deducted for an incorrect prediction.

5.2 Results by Dataset

The performance of different Medical LLMs on various datasets is summarized in Table 2 for baseline results and Table 1 for the results using the proposed prompting techniques. The analysis focuses on the accuracy improvements observed with the multi-step hybrid prompting methods.

The performance of different Medical LLMs on various datasets shows significant improvements in accuracy with the application of the proposed Multi-step Hybrid prompting techniques.

For the **MedHALT-FCT** dataset, the baseline results show that OpenBioLLM-8B achieved an accuracy of 32.81%. With the Multi-step Hybrid technique 1 approach, Phi-3-mini-4k-instruct improved to 46.88%. LLaMA-3-instruct-8B also demonstrated substantial improvements, achieving 51.74% with the Hybrid technique 1 method.

In the **MedHALT-FAKE** dataset, baseline performance was low, with the Medalpaca-7B model achieving only 4.63% accuracy. However, Phi-3-mini-4k-instruct increased to 9.47% with Hybrid technique 1, and LLaMA-2-7B achieved 22.55%.

For the **MedHALT-NOTA** dataset, Mistral-7B-instruct had a baseline accuracy of 24.12%. With the advanced prompting methods, Phi-3-mini-4k-instruct achieved 21.50% with Hybrid technique 1.

In the **MedQA_USMLE** dataset, OpenBioLLM-8B had a baseline accuracy of 20.24%. Phi-3-mini-4k-instruct improved accuracy to 52.50% using Hybrid technique 1, while LLaMA-3-instruct-8B reached 54.88% with Hybrid technique 2.

For the **MedMCQA** dataset, OpenBioLLM-8B's baseline accuracy was 34.39%. Phi-3-mini-4k-instruct achieved 47.61% with Hybrid technique 1. LLaMA-3-instruct-8B also showed improvements,

	Datset	Normal			ZK GenMed-1			ZK GenMed-2		
LLM Model		Acc.	F1	P.S.	Acc.	F1	P.S.	Acc.	F1	P.S.
Phi-3-mini-4k-instruct	FCT	43.75	0.32	28.50	46.88	0.46	32.25	32.81	0.32	15.38
	NOTA	21.50	0.16	18.75	32.50	0.26	156.25	16.40	0.15	-45.0
	MedQA	33.58	0.24	216.12	52.50	0.49	517.17	35.66	0.34	249.25
	MedMCQA	38.23	0.28	641.62	47.61	0.38	971.83	33.77	0.32	484.75
	FAKE	1.13	0.00	-438.25	9.47	0.02	-244.50	3.07	0.01	-393.25
LLaMA 2-7B	FCT	22.92	3.50	7.25	19.79	0.17	-0.25	18.75	0.17	-1.50
	NOTA	9.65	0.09	-129.38	12.70	0.10	-91.25	11.00	0.08	-112.50
	MedQA	24.98	0.22	79.25	24.98	0.22	79.25	16.42	0.15	-57.00
	MedMCQA	26.03	0.25	212.25	20.61	0.17	21.62	29.94	0.26	349.75
	FAKE	0.93	0.00	-442.83	22.55	0.04	59.25	-	-	-
LLaMA 3-8B	FCT	38.54	0.38	22.25	20.31	0.21	0.38	38.54	0.34	22.25
	NOTA	7.10	0.07	-161.25	7.10	0.07	-161.25	9.60	0.09	-130.0
	MedQA	41.87	0.42	348.0	45.40	0.45	404.25	24.46	0.26	71.00
	MedMCQA	44.46	0.44	861.0	44.46	0.44	861	27.92	0.27	278.92
	FAKE	0.94	0.00	-442.62	2.66	0.01	-402	6.62	0.02	-310.75
LLaMA 3-instruct -8B	FCT	-	-	-	51.74	0.48	38.08	52.26	0.44	38.71
	NOTA	33.10	0.26	163.75	32.0	0.26	150.0	32.0	0.26	150.0
	MedQA	45.40	0.45	404.25	53.69	0.43	536.12	54.88	0.44	555.08
	MedMCQA	-	-	-	50.75	0.40	1082.25	50.37	0.41	1069.12
	FAKE	10.01	0.03	-232.00	12.33	0.03	-178.25	-	-	-

Table 1: Performance matrix Analysis of Acc(Accuracy), F1(F1 score) and P.S.(Pointwise score.)

Dataset	**Metric**	**Medalpaca-7B**	**Mistral-7B-instruct**	**OpenBioLLM-8B**	**Meditron-7B**
FAKE	Acc.	4.63	4.41	25.87	0.0
	F1	0.01	0.01	0.06	0.0
	P.S.	-357.00	-362.00	136.33	-464.5
FCT	Acc.	37.50	50.0	32.81	10.42
	F1	0.23	0.40	0.28	0.06
	P.S.	21.0	36.0	15.38	-11.50
MedQA	Acc.	34.80	39.83	20.24	12.33
	F1	0.25	0.34	0.23	0.07
	P.S.	235.50	315.50	3.83	-122.00
MedMCQA	Acc.	37.14	40.98	34.39	10.03
	F1	0.29	0.32	0.31	0.06
	P.S.	603.50	150	506.42	-11.50

Table 2: Comparison of Baseline for Medical LLMs.

achieving 50.75% with Hybrid technique 1 and 50.37% with Hybrid technique 2.

Additionally, Phi-3-mini-4k-instruct showed improvements across multiple datasets, notably achieving 33.58% accuracy on **MedQA_USMLE** with the Normal approach, and 52.50% with Hybrid technique 1. Similarly, on **MedMCQA**, the accuracy improved from 38.23% to 47.61% with Hybrid technique-1.

LLaMA-2-7B improved its performance on the **MedQA_USMLE** dataset from a baseline of 24.98% to 24.98% with both Normal and Hybrid technique 1 approaches. On the **MedMCQA** dataset, its accuracy increased from 26.03% to 29.94% with Hybrid technique 2.

LLaMA-3-8B demonstrated significant gains on the **MedQA_USMLE** dataset, improving from 41.87% to 45.40% with Hybrid technique 1 and reaching 53.69% with Hybrid technique 2. On the **MedMCQA** dataset, its performance increased

from 44.46% to 50.75% with Hybrid technique 1. Overall, the results indicate that the advanced prompting techniques yielded substantial improvements across all datasets, highlighting their effectiveness in enhancing the performance of Medical LLMs.

6 Conclusion

In this study, we proposed an effective framework *ZK-GenMed*, a multi-step text-generative hybrid prompting technique for zero-shot QA setting, along with the analysis of its performance of various Medical Large Language Models (LLMs) across several medical datasets, including MedHALT-FCT, MedHALT-FAKE, MedHALT-NOTA, MedQA_USMLE, and MedMCQA. Our proposed methods demonstrated significant improvements over multiple models on accuracy, F1 score, and pointwise score across all datasets compared to baseline results.

The experimental findings demonstrated the possibility of sophisticated prompting techniques to improve Medical LLM performance, particularly in difficult datasets with low initial baseline accuracies. The LLaMA-3-instruct-8B models and Phi-3-mini-4k-instruct, for example, demonstrated considerable accuracy gains, demonstrating the versatility and resilience of the suggested techniques. Furthermore, the uniform performance improvements on several datasets highlight how broadly applicable our method is.

In conclusion, this study opens the door for more precise and dependable medical NLP applications by demonstrating the effectiveness of multi-step generative hybrid prompting strategies in improving the performance of medical language models. The developments discussed here can help to improve diagnostic tools, medical record analysis, and overall healthcare management through enhanced language understanding and creation skills.

7 Future Work

This work can be extended by exploring its effectiveness in general domain question answering and by adding more domain-specific knowledge and examining how these prompting strategies affect other NLP tasks like named entity identification and relation extraction; future research may improve these prompting strategies even further. A deeper understanding of these methods' scalability and efficacy may also be obtained by merging them with bigger and more varied datasets. Additionally, adding verified facts to Knowledge Graphs could improve the factual accuracy and dependability of the LLMs' outputs, guaranteeing that the information produced is accurate and reliable.

8 Limitations

In the ZK-GenMed framework, we use the generated text from an LLM to create a knowledge base, which is later used to direct the development of responses. Also, our technique needs more processing power to conduct concentrated and multi-step thinking before response creation. Also, the prompts in our system are human-written; they are not universally applicable. Along with this, the generated text may not be free from the issue of LLM hallucination and may contain incorrect information. Since the the generation of relevant text depends on the reasoning abilities of LLMs, and the manual prompts asked by the user may impact, it may produce incorrect phrases during the pondering or extraction or in the inference phase. The technical method of creating these prompts requires more work. Our goal is for future research to build on our approach, which is more error-resilient and by augmenting current implementation with real-world correct data and more resilient in variances of automatic prompt engineering. Hence, it can assist the existing framework in generating high-quality knowledge used in the later stages.

References

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen,

Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Sia Gholami and Mehdi Noori. 2021. Zero-shot open-book question answering. *Preprint*, arXiv:2111.11520.

Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv preprint arXiv:2403.04890*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023. Rho (ρ): Reducing hallucination in open-domain dialogues with knowledge grounding. *Preprint*, arXiv:2212.01588.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13507–13515.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *Preprint*, arXiv:2307.15343.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Damien Teney and Anton van den Hengel. 2016. Zero-shot visual question answering. *Preprint*, arXiv:1611.05546.

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,

Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Junjie Wang, Ping Yang, Ruyi Gan, Yuxiang Zhang, Jiaxing Zhang, and Tetsuya Sakai. 2023. Zero-shot learners for natural language understanding via a unified multiple-choice perspective. *IEEE Access*.

Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu, and Lei Li. 2022. Pre-trained language models can be fully zero-shot learners. *arXiv preprint arXiv:2212.06950*.

Jing Zhou, Zongyu Lin, Yanan Zheng, Jian Li, and Zhilin Yang. 2022. Not all tasks are born equal: Understanding zero-shot generalization. In *The Eleventh International Conference on Learning Representations*.