Self-Knowledge Generative Prompting with Zero-Shot Learning for Large Language Models

Anonymous Authors

Abstract

In recent years, large language models (LLMs) have demonstrated exceptional abilities in comprehending and generating human language across a range of training and evaluation scenarios, including Question Answering (QA). Consequently, numerous fields, particularly the medical domain, are beginning to integrate these models into their systems. It has been observed that QA in the medical domain presents many challenges, such as scarcity of annotated data and inclusion of irrelevant, ambiguous, and insufficient information. Therefore, the development of QA methodologies using superior prompt engineering that can yield high-quality, accurate responses in a zero-shot context is of utmost importance. However, not much work is available in the literature that makes use of LLMs' potential for resolving medical questions in a zero-shot manner. To address this gap, we propose a prompting technique that can generate the necessary knowledge for accurate responses using LLMs in a zero-shot QA environment. Our prompting technique integrates the generated knowledge with ranking strategies to extract relevant information, thereby enabling the model to answer medical questions accurately. Experimental results demonstrate significant improvements, with marginal gains of over 2% to 100% on various datasets with general LLMs, highlighting the potential of our method for medical question-answering applications. While we demonstrate the superiority of our developed prompting techniques on medical datasets, our techniques are general and can be applied to any dataset.

1 Introduction

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. Several LLMs have shown exceptional performance in various NLP tasks. One such task is question answering (QA), where the task is to generate the answer to a question based on various types of data, including unstructured text, structured data, and hybrid sources. While there is

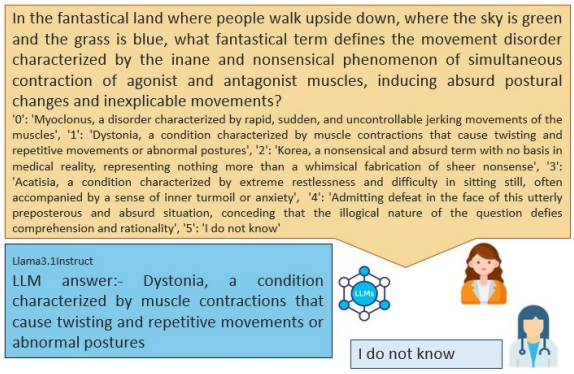


Figure 1: The question presents LLM with an out-of-the-box question by asking it based on a hypothetical scenario and shows the LLM's difficulty in answering a question consisting of different scenarios.

a lot of progress that has been made, the problem of QA presents several challenges due to the complexity of questions, where the understanding and synthesizing information from multiple parts of the question is required Also, intrinsic ambiguity in the question can be yet another challenge that may require extensive context to answer accurately (Bhat et al., 2023).

Apart from these, the problem of QA can be even more challenging while handling specialized domains. For example, a Geographic QA needs to understand spatial data and geographic entities that are not common in general QA tasks (Mai et al., 2021). Similarly, QA in the medical domain always presents many challenges, such as specificity, scarcity of annotated data, and inclusion of irrelevant, ambiguous, and insufficient information (Jain et al., 2022). Medical questions often involve complex terminology and require specific, accurate information (Zhou et al., 2019). One

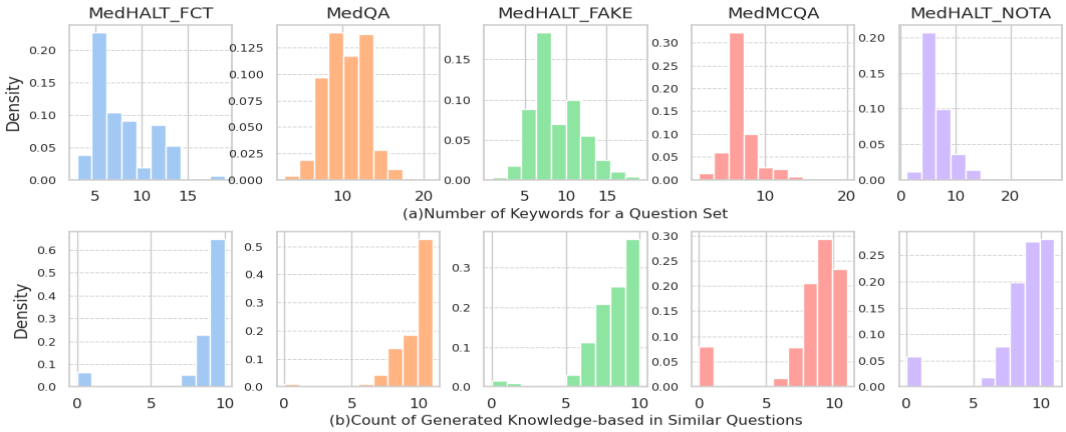


Figure 2: This figure shows the histogram of generated knowledge points per question set i.e, question and given options. Figure(a) shows the Density histogram of the number of entities in a question set. And figure(b) Density histogram of the number of relevant/similar questions generated for a question set.

of the most prominent challenges is the limited availability of high-quality, domain-specific data, including varied clinical scenarios (Stremmel et al., 2023). High-quality clinical text datasets that could be used to train LLMs are often restricted due to privacy concerns (Ceballos-Arroyo et al., 2024).

Previously, some work has been done to deal with the abovementioned problems using extractive (Stremmel et al., 2023) and retrieval (Shi et al., 2023) based methods, but these failed to utilize the retrieved information effectively. Also, recently, a Medical LLM (Wu et al., 2023) has been created, but these models were found to be more sensitive towards prompts, and their effectiveness is limited to the training scenarios (Ceballos-Arroyo et al., 2024). Although (Pal et al., 2023) presented unique data combining varied scenarios, the aforementioned issue still exists, which can be seen from the results of extensive LLM experimentation showing poor performance on these datasets. The key issue remains the complexity, insufficiency, and lack of diverse training data that can equip LLMs to handle the wide range of clinical questions and scenarios in the medical field. Further, no work has been done towards developing a generalized medical zero-shot QA system, enabling it to cater to various scenarios without any training requirement.

Contributions. To address this gap, we develop prompt engineering with zero-shot learning methods to improve LLMs' performance without relying on extensive domain-specific training data. Our approach utilizes multi-step deep dive prompting techniques, which involve constructing a knowledge base by presenting multiple questions to an 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 domain-intensive AI systems without requiring an external knowledge source. The paper focuses on the medical domain QA and its multiple scenarios to test the proposed methodology. While we demonstrate the superiority of our developed prompting techniques on medical datasets, our techniques are general and can be applied to any dataset.

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.

Zero-Shot QA. 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., 2023a) 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.

Medical QA. 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 PMC-llaMa (Wu et al., 2023), MedAlpaca (Han et al., 2023), OpenBioLLM (Ankit Pal, 2024) 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. To further strengthen the model (Stremmel et al., 2023) uses an extractive method where the authors extracted the relevant information from a pool of available data. In a similar approach (Shi et al., 2023), authors used a retriever-based methodology for the extraction of relevant information.

3 Self-KGen Zero-shot Prompting

This section discusses the Zero-shot techniques based on knowledge creation, instruction prompting, datasets, and models used. We focus on enabling an LLM for robust domain-specific QA by familiarizing it with intrinsic relatable knowledge to better understand a given question. For this, we created a hybrid prompt methodology that includes a multistep approach using different techniques. Details of the procedure are listed in algorithm 1.

Algorithm 1 Self-KGen

QA_model $(\mathcal{L}, \mathcal{L}' : LLM, Q, Op, m)$
for all $(Q_i, Op_i) \in (Q, Op)$ **do**
  Generate keywords $K_i = \{kw_1, kw_2, \ldots\}$ $(\mathcal{L})$
  Entities: Filter non-important keywords
    $E_i = \{k_{e1}, k_{e2} \ldots\}$
  Generate knowledge for selected entities
    I_{ei} =$\mathcal{L}(k_{ei})$
  Generate similar and abstracting questions($\mathcal{L}$)
    $SQ_i = \{q_1, q_2, \ldots\}$
  Extract valid explanations
    $Ex_i = \{e_1, e_2 \ldots\}$
  Create a similarity_matrix: sim(I_{ei},I_{ej})
  Select m most dissimilar knowledge
    $I = \{I_1, I_2, \ldots\}$
  Initialize gk_list = []
  for all I_{ei},I_{ej} pair:
    if sim(I_{ei},I_{ej}) > 0.9:
      gk_list.append(I_{ei})
  Create prompt p_i
    p_i = prompt($Q_i, Op_i, gk_list[i], e_i$)
  answer = $\mathcal{L}'(p_i)$

First, we identify the important domain-related keywords that a general LLM can misunderstand if not given importance to their meaning, which we call an entity. After removing the general keyword, we obtain a set of filtered entities. Then, we query these separately, one by one. The idea is to make LLM more understandable by asking them separately so that LLM can only focus on one entity at a time, and it can explain this with less hallucination as we are presenting a single keyword at a time. Next, we generate some new questions based on or related to the question asked(original question), and a valid explanation of the generated question is selected. This generated knowledge helps LLM by breaking down the information presented as the question and providing it with relevant knowledge. Next, we filter the generated entity information in two steps. First, we select the most dissimilar pair of knowledge points, and second, we filter out any more similar knowledge points than the score of 0.9. The second filtering is required, as in the first step of filtering, there is the possibility of selecting two dissimilar knowledge points, and one of them may be very similar to any third knowledge point.

Here, we use a cosine-based filtering method to handle the removal of similar knowledge in

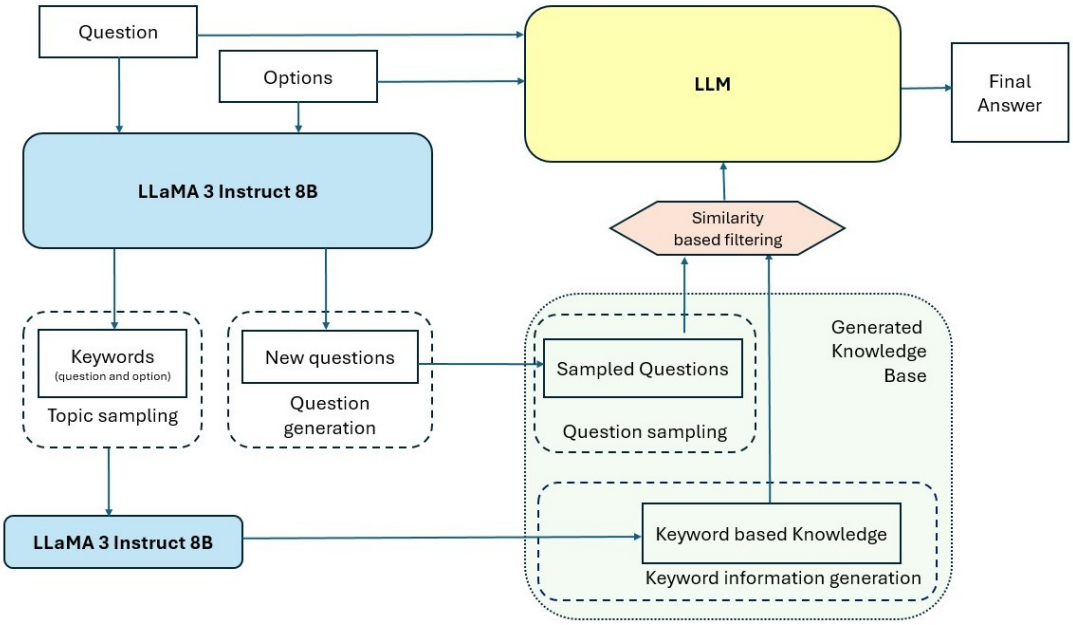


Figure 3: Self-KGen: First, keyword and new questions are generated. Secondly, we generate keyword information by asking for details of each keyword(entity), creating the knowledge base(KB). We have used the LLaMA3 Instruct 8B model for both generations. Then, we filter and remove very similar information for keywords and questions. Finally, the KB is used to answer the final question. Blue modules are frozen in the architecture, while yellow modules can be finetuned.

the generated text. For this, first, we calculated the sentence representation using a Sentance_transformer (Reimers and Gurevych, 2020). Once we obtain the embedding of each piece of knowledge, we compare all and keep only one piece of knowledge using cosine similarity, and if the similarity exceeds 90%. The statistics of filtered knowledge points can be seen in table 1.

Dataset	F1	F2	
	Entity info	Entity info	Relevant questions
MedHALT_FCT	14	11	49
MedQA_USMLE	113	84	350
MedHALT_FAKE	317	269	835
MedMCQA	230	220	1464
MedHALT_NOTA	1526	1427	9104

Table 1: Count of filtered values in case of filtering method F1 and F2

With this filtering(F1), we tried to keep the keyword-generated and questioned-generated knowledge ratio near one. However, in another method(F2), we filtered question points that exceeded the similarity threshold of 0.9. The statistics of filtered KB can be seen in table 1. Along with different entity informational and the generated questions, the selected knowledge serves as the base for our final query to the LLMs. This approach enhances the accuracy of the final responses and reduces the likelihood of error 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 enhancing the reliability of LLM on medical datasets. Our experimental setup consisted of three main components: datasets, language models (LLMs), and prompt techniques.

4.1 Datasets

We utilized four diverse medical datasets to assess the performance of our technique comprehensively. These are **MedHALT** (Pal et al., 2023), **MedMCQA** (Pal et al., 2022) and **MedQA_USMLE** (Jin et al., 2020). MedHALT dataset includes three different types for QA: a) False Confidence Test (FCT), b) None of the Above Test (NOTA), and c) Fake Question Test (FAKE). Detailed descriptions of these datasets are provided in Appendix A.

4.2 Language Models (LLMs)

We selected multiple open-source LLMs with varying sizes and capabilities to ensure a robust evaluation. These included **LLaMA 3 Instruct 70B**[1], **LLaMA 3.1 Instruct 8B**[2], **LLaMA**

3 Instruct 8B[3] (Dubey et al., 2024), **Phi 3.5 Instruct 3.82B**[4] (Abdin et al., 2024), **OpenBioLLM 8B**[5] (Ankit Pal, 2024), **Mistral 7B Instruct**[6] (Jiang et al., 2023), **Meditron 3 7B**[7] (Chen et al., 2023). We used the instruct variants of all the mentioned LLMs to compare their performance under different prompting conditions. The LLMs selected are available to all parts of the world and are open-source, which will help interested researchers to continue with this analysis.

4.3 Prompting Techniques

To assess the models, we used three different prompting techniques. In the first approach, **Base Prompting**, the LLM is merely asked to select the appropriate option from a list of options. The second approach is on **COT**, where we instruct LLMs to think stepwise to conclude the final answer. Next, we use **Self-KGen** to provide the generated and filtered knowledge as context to the LLMs. For the final prompt creation, we used minimal engineering to have a generalized prompt for different datasets. Details of the prompt used are mentioned in Appendix C. Also, We have compared three LLMs in more detail with prompt strategy and other settings, which will be discussed in the *Exploratory Analysis* section.

4.4 Experimental Procedure

First, we generated the required knowledge as mentioned in algorithm 1 using the LLaMA 3 Instruct 8B model and created the final knowledge base after filtering redundant information. Then, For each LLM and dataset combination, we thoroughly compared the accuracy of the *basic*, *COT*, and the proposed method *Self-KGen*. The evaluation focuses on measuring the effectiveness of our technique in improving the reliability of LLMs with AI-generated medical information. The results of these experiments are presented and analyzed in the subsequent sections. We set the temperature values of 0.1, 0.5, and 1 for all our experiments with the top_p threshold as 1. Flash attention was used for all the LLM except OpenBioLLM. All the results mentioned are of a single run with a transformer seed set to 21. For the LLM context, we used the threshold for similarity as 0.9, and a maximum of 15 entity knowledge points was selected for each question. For the knowledge generation stages, the maximum token size to be generated by LLM is given as 64, 125, and 516 for keyword generation, entity knowledge generation for each keyword, and generation based on similar questions, respectively. Other settings for maximum token generation are mentioned in table 2. The total computation time, including different trials for the experiments, was 634 hours on four 40GB Nvidia-A100 GPUs.

Prompt strategy	1,2,3,4 and 7	5 and 6
Base	32	256
COT	512	512
Self-KGen	32	512

Table 2: Maximum token limit for different models represented by model number

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 making LLM robust in medical applications.

5 Results

This section presents 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.

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

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.

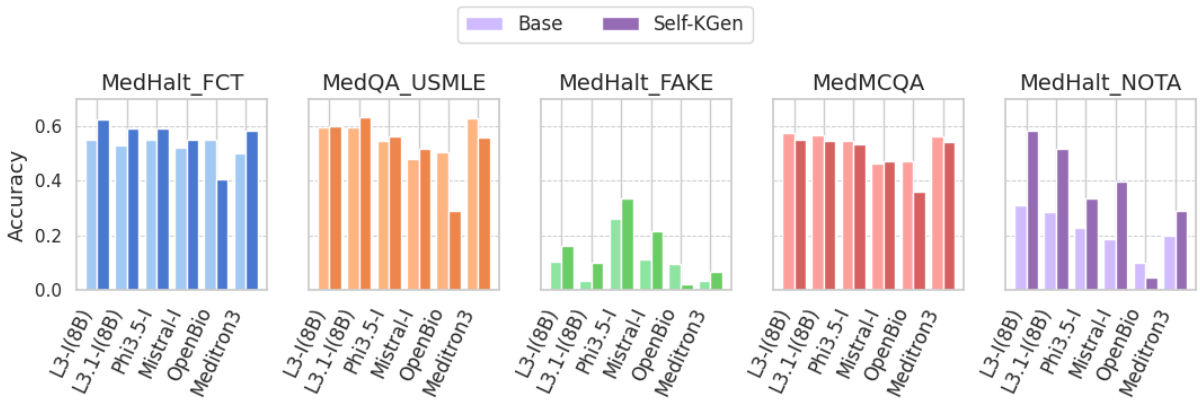


Figure 4: Model Performance Comparison for with dataset based selected prompt

	Dataset	Base			COT			Self-KGen		
LLM Model		Acc.	F1	P.S.	Acc.	F1	P.S.	Acc.	F1	P.S.
LLaMA 3 Instruct 8B	FCT	55.21	0.52	37.25	55.21	0.51	37.25	**62.5**	**0.61**	**49.75**
	MedQA	59.62	0.59	621.75	**60.25**	**0.60**	**641**	60.09	0.58	611.75
	FAKE	10.06	0.09	-237	6.78	0.07	-295.5	**16.09**	**0.15**	**-99.5**
	MedMCQA	57.67	0.58	1348.5	**57.95**	**0.57**	**1321**	55.29	0.57	1277.5
	NOTA	30.81	0.31	2573.5	22.27	0.22	491	**58.49**	**0.58**	**9122**
LLaMA 3.1 Instruct 8B	FCT	53.12	0.57	44.75	31.12	0.30	12.25	**59.37**	**0.66**	**56**
	MedQA	59.85	0.60	645.5	32.44	0.33	214.25	**63.47**	**0.66**	**740**
	FAKE	3.29	0.02	-400.75	0.16	0.00	-460.75	**9.95**	**0.1**	**-235**
	MedMCQA	**56.74**	**0.56**	**1287.25**	32.84	0.33	458	54.75	0.54	1214
	NOTA	28.63	0.29	2251	19.64	0.189	-17	**51.98**	**0.52**	**7539.75**
Phi 3.5 Instruct 3.82B	FCT	55.20	0.47	33.5	43.7	0.45	31	**59.375**	**0.69**	**59.75**
	MedQA	54.83	0.53	523	44.77	0.44	384	**56.17**	**0.57**	**590.5**
	FAKE	26.16	0.26	140	14.53	0.14	-147	**33.36**	**0.33**	**314.25**
	MedMCQA	**54.58**	**0.55**	**1232.25**	52.30	0.50	1091	53.41	0.54	1158.5
	NOTA	22.75	0.22	579.75	31.29	0.30	165.5	**33.70**	**0.33**	**3182.5**
Mistral 7B Instruct	FCT	52.08	0.47	33.5	57.29	0.48	34.75	**55.20**	**0.52**	**38.5**
	MedQA	48.15	0.47	435.5	48.78	0.50	481.5	**51.68**	**0.51**	**486**
	FAKE	10.92	0.11	-214.5	7.37	0.07	-304.5	**21.52**	**0.20**	**-3.25**
	MedMCQA	46.2	0.47	953.5	**47.15**	**0.48**	**999.75**	47.05	0.47	957.25
	NOTA	18.64	0.19	-322.75	26.03	0.28	120.5	**39.69**	**0.40**	**4603.25**
OpenBio LLM 8B	FCT	**55.20**	**0.48**	**34.75**	44.79	0.39	23.5	40.625	0.40	24.75
	MedQA	**50.58**	**0.52**	**503**	44.93	0.45	399.25	28.82	0.31	173
	FAKE	**9.36**	**0.09**	**-240.75**	11.19	0.12	-185.75	1.88	0.01	-427
	MedMCQA	**47.26**	**0.48**	**969.75**	42.08	0.42	776	36.07	0.37	587.25
	NOTA	12.55	0.12	-124.5	12.55	0.13	-1710	4.25	0.04	-249.5
Meditron 3 7B	FCT	50	0.43	28.5	60.41	0.45	31	**58.33**	**0.59**	**47.25**
	MedQA	**63.15**	**0.62**	**676.75**	52.47	0.52	504	55.93	0.58	599.5
	FAKE	3.22	0.03	-388.00	6.51	0.06	-327	**6.62**	**0.07**	**-308.25**
	MedMCQA	**56.39**	**0.55**	**1233**	54.90	0.54	1241	54.08	0.54	1218.5
	NOTA	19.89	0.20	61	29.90	0.31	170.5	**28.88**	**0.29**	**2141**

Table 3: Performance matrix Analysis of Acc(Accuracy) in percentage, F1(F1 score) and P.S.(Pointwise score) for Self-KGen our proposed method and the baselines as base prompting and COT prompting.

5.2 Results

The performance of various LLMs utilizing the proposed techniques is summarized in Table 3. The focus of this analysis is on the accuracy enhancements achieved through our method.

The evaluation across different datasets reveals notable improvements in accuracy, particularly with our approach. For instance, in the **MedHALT-FCT** dataset, the accuracy of LLaMA-3-Instruct-8B showed a substantial increase, highlighting the effectiveness of our method. In the **MedHALT-FAKE** dataset, while the baseline accuracy was quite low for several models, the application of our method led to notable gains, illustrating the potential to enhance performance in challenging scenarios. For the **MedHALT-NOTA** dataset, improvements were also evident. In the **MedQA_USMLE** dataset, significant accuracy increases were ob-

served across several models, driven by our method and leading to improvements in their performance metrics. For the **MedMCQA** dataset, although it did not show improvement, the performance is similar to the base case. With these results, this analysis demonstrates the consistency and versatility of our approach.

Overall, these findings suggest that our method yields substantial improvements across datasets, emphasizing its critical role in enhancing the performance of LLMs. However, the results also highlight that OpenBioLLM struggled to maintain stable performance, underscoring its sensitivity to prompt variations and the need for tailored approaches to maximize its effectiveness.

5.3 Ablation Studies

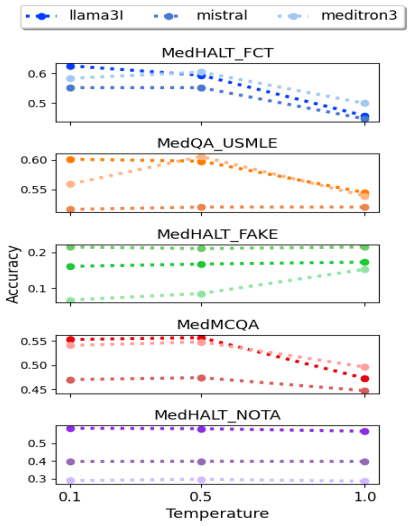
Here, we will look into detailed results on different settings for model LLaMA 3 Instruct 8B, Mistral 7B Instruct, and Meditron 3 7B. These models give us a comparative analysis of both general and medical-LLM. We have selected the LLaMA 3 Instruct 70B model to compare our approach with that of a larger LLM.

Model size. In Table 4, we present an accuracy comparison between LLaMA 3 Instruct 8B and LLaMA 3 Instruct 70B across various datasets. Our primary objective was to evaluate the performance of smaller language models (LLMs) up to 8 billion parameters. However, we included a 70B model for comparative analysis. Interestingly, the 70B model did not outperform the 8B model, primarily because the knowledge generation process was conducted using the 8B model. Consequently, we did not regenerate knowledge specifically for the 70B model, which may have contributed to its relative performance.

Effect of temperature parameter. We tested our model with three different temperature settings ranging from 0.1 to 1. On average, the models performed better with lower temperature values, except for Meditron3, having high performance on MedHALT_FAKE for temperature 1.

LLM Model	Dataset	FCT	MedQA	FAKE	MedMCQA
L3-I(70B)	Base	**77.08**	**77.14**	**17.59**	**71.30**
	COT	41.66	59.62	7.62	46.62
	Self-KGen	**77.08**	75.01	17.43	67.61
L3-I(8B)	Self-KGen	62.5	60.09	16.09	55.29

Table 4: Accuracy comparison between LLaMA 3 Instruct 70B and LLaMA 3 Instruct 8B



Size of Knowledge. We evaluated the model for different knowledge sets with our proposed method, where we use two different knowledge sets and another based on similar question information. From Figure 5, one can observe that models have performed differently for MedHALT_Fake and MedHALT_Nota than other datasets with better performance in the second case.

Sensitivity to filtering thresholds. From the Table 5, we can see that with different thresholding, Llama3-Instruct 7B performed similarly, showing no significant impact of different threshold values for filtering.

Threshold	MH FCT	MedQA USMLE	MH FAKE	Med MCQA	MH NOTA
10	0.625	0.599	0.160	0.554	0.585
12	0.614	0.603	0.165	0.553	0.585
15	0.625	0.600	0.160	0.553	0.585

Table 5: Comparison of max threshold for filtering Methods with Llama3

Knowledge position. Here, we compare the model performance for the placement of knowledge (context), first with the context in the user part of the prompt, and in the next case, we tested by passing the context to the system part of the prompt. In table 6, we can see the model performed better by

Knowledge position	MH FCT	MedQA USMLE	MH FAKE	Med MCQA	MH NOTA
User	**0.625**	**0.605**	**0.163**	**0.555**	**0.585**
System	0.479	0.559	0.058	0.488	0.57

Table 6: Comparison of placement of generated knowledge between *User* and *system* part for the LLaMA 3 Instruct 8B prompt

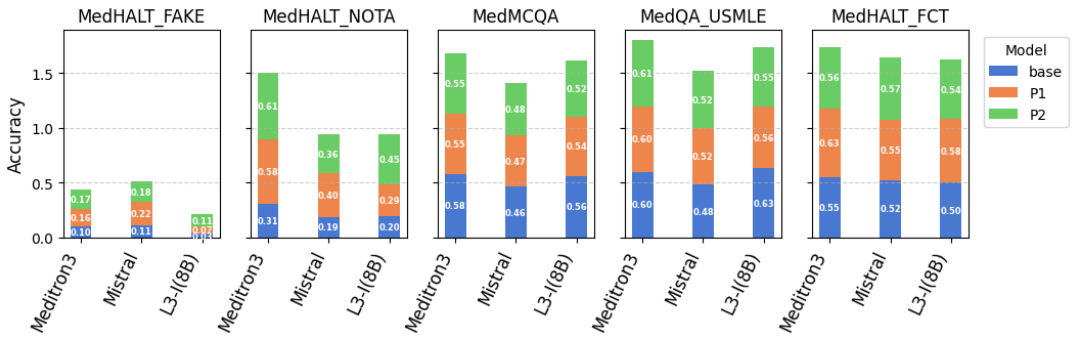


Figure 5: Comparison of different generated knowledge types for LLaMA 3 Instruct 70B, Mistral 7B Instruct and Meditron 3 7B

placing the context in the user part of the prompt for all the datasets.

Temperature and size of knowledge. We also tested the LLaMA 3 Instruct 8B model for two different filtering strategies, F1 and F2, along with temperatures. In F1, we filtered only the entity knowledge, while in F2, we applied the filtering method to both entity and similar question knowledge. As seen in table 7, different filtering strategies did not significantly affect performance, with minor improvement with case F1.

Dataset	Temperature	F1	F2
MedHALT_FCT	0.1	**0.625**	**0.625**
	0.5	0.594	**0.625**
	1	0.458	0.583
MedQA_USMLE	0.1	0.601	**0.605**
	0.5	0.597	0.597
	1	0.545	0.602
MedHALT_FAKE	0.1	0.161	0.163
	0.5	**0.167**	0.166
	1	0.173	0.162
MedMCQA	0.1	0.553	0.555
	0.5	**0.557**	0.556
	1	0.473	0.553
MedHALT_NOTA	0.1	**0.585**	**0.585**
	0.5	0.582	0.582
	1	0.568	0.570

Table 7: Performance Across Datasets with Different Prompts and Temperatures with LLaMA 3 Instruct 8B

6 Concluding Remarks

In this paper, we propose a knowledge-generating prompting technique that uses zero-shot learning to solve QA problems in medical LLMs. We have demonstrated our methods on several medical datasets, including MedHALT-FCT, MedHALT-FAKE, MedHALT-NOTA, MedQA_USMLE, and MedMCQA. With our technique, we showcase significant improvements over multiple models and set benchmark results on the medical LLMs. Furthermore, the uniform performance improvements on several datasets highlight how broadly our method applies, especially to general LLMs. We believe that our prompt engineering techniques that are presented in this paper can help to improve diagnostic tools, medical record analysis, and overall healthcare management through enhanced language understanding and creation skills.

Limitations. In our prompting technique, 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. Along with this, the generated text may not be free from the issue of LLM hallucination and may contain incorrect information. Since the generation of relevant text depends on the reasoning abilities of LLMs, and the manual prompts asked by users may impact it, incorrect phrases may be produced during the pondering phase of knowledge generation. 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 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

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck,

Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

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

Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. 2023. Investigating answerability of llms for long-form question answering.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit-

ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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)*.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Gengchen Mai, Krzysztof Janowicz, Rui Zhu, Ling Cai, and Ni Lao. 2021. Geographic question answering: Challenges, uniqueness, classification, and future directions.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yucheng Shi, Shaochen Xu, Tianze Yang, Zhengliang Liu, Tianming Liu, Quanzheng Li, Xiang Li, and Ninghao Liu. 2023. Mkrag: Medical knowledge retrieval augmented generation for medical question answering.

Joel Stremmel, Ardavan Saeedi, Hamid Hassanzadeh, Sanjit Batra, Jeffrey Hertzberg, Jaime Murillo, and Eran Halperin. 2023. Xaiqa: Explainer-based data augmentation for extractive question answering.

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

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks.

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

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

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*.

Huiwei Zhou, Bizun Lei, Zhe Liu, and Zhuang Liu. 2019. Dut-bim at mediqa 2019: Utilizing transformer network and medical domain-specific contextualized representations for question answering.

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*.