Retrieval-Augmented Medical Large Language Models

Hala Youssef *, Jacob Turner , Evan Sanders

Posted Date: 22 October 2024

doi: 10.20944/preprints202410.1658.v1

Keywords: Biomedical Language Models; Self-Reflection Mechanism; Medical Question-Answering;Retrieval-Augmented Models

 doi:10.20944/preprints202410.1658.v1

Article

Retrieval-Augmented Medical Large Language Models

Hala Youssef *, Jacob Turner and Evan Sanders

Minia University
* Correspondence: s30311201800374@nur.dmu.edu.eg

Abstract: Biomedical large language models (LLMs) have made significant strides, but their reliance on external retrieval mechanisms presents challenges in accuracy and computational efficiency. To address these issues, we propose MedRAG-Refine, a generative LLM designed specifically for the biomedical domain. Our model integrates a two-stage fine-tuning process, incorporating a self-reflection mechanism to improve reasoning quality. We evaluate our model on MedQA, MedMCQA, and MMLU datasets, demonstrating superior performance over state-of-the-art methods. Additionally, human evaluations confirm the enhanced accuracy and reasoning quality of our model in real-world medical tasks.

Keywords: biomedical language models; self-reflection mechanism; medical question-answering; retrieval-augmented models

1. Introduction

In recent years, large language models (LLMs) have shown remarkable potential in a variety of domains, including healthcare and biomedical applications. However, biomedical knowledge is vast, rapidly evolving, and specialized, which presents significant challenges for LLMs that are trained primarily on general corpora. To address this gap, retrieval-augmented large language models (RAGs) have been developed, enabling models to fetch and integrate external biomedical resources during the inference process. These models have been particularly valuable in medical reasoning, providing clinicians and researchers with advanced tools for answering complex biomedical questions by retrieving relevant documents from specialized databases such as PubMed and PMC [1,2].

Despite the effectiveness of RAGs in improving biomedical language processing, they face several challenges. First, relying on external retrieval mechanisms can introduce inefficiencies, as the quality and relevance of retrieved documents are not always guaranteed, leading to potential misinterpretations or hallucinations. Additionally, the integration of external retrieval systems requires significant computational resources, making the real-time application in clinical settings less feasible. Finally, the rapid evolution of biomedical literature creates a persistent risk that the retrieved documents may become outdated, which is problematic in a field that demands up-to-date information [3–6]. These challenges highlight the need for models that can internalize domain-specific knowledge while still being flexible enough to adapt to new information.

To address these limitations, we propose a novel approach to training large language models tailored specifically for the biomedical domain, without over-relying on external retrieval mechanisms. Our approach consists of a two-stage fine-tuning process: first, the model undergoes domain-specific pre-training on large-scale biomedical corpora, including research papers, clinical guidelines, and textbooks. This enables the model to acquire deep foundational knowledge within the domain. Following this, task-specific fine-tuning is applied using medical question-answering datasets like MedQA, MedMCQA, and MMLU, during which we introduce a self-reflection loss mechanism. This loss function encourages the model to self-assess its generated answers, promoting higher accuracy and relevance in the absence of external retrieval [2,5]. By embedding domain-specific reasoning directly into the model parameters, we aim to reduce the dependency on external retrieval, leading to more efficient, real-time medical decision support.

Our experimental evaluations are conducted using a range of established biomedical datasets, such as MedQA, MedMCQA, and the medical subset of the MMLU benchmark. The model's performance is assessed using standard evaluation metrics, including accuracy and the quality of reasoning

 doi:10.20944/preprints202410.1658.v1

behind its answers. Comparative results with state-of-the-art models such as Self-BioRAG [2], RAG [1], and LLaMA2 [4] demonstrate that our approach achieves superior performance, particularly in complex medical reasoning tasks, while also improving model efficiency by reducing reliance on real-time retrieval.

1. We propose a novel two-stage fine-tuning process for biomedical large language models that reduces reliance on retrieval mechanisms by embedding domain-specific knowledge directly into the model.
2. We introduce a self-reflection loss mechanism during task-specific fine-tuning, which encourages the model to self-assess its answers, improving accuracy and reasoning quality.
3. Extensive experiments on medical question-answering datasets demonstrate that our method outperforms existing state-of-the-art models in both accuracy and efficiency, making it more suitable for real-time clinical applications.

2. Related Work

2.1. Large Language Models

Large Language Models (LLMs) have emerged as a cornerstone of modern natural language processing (NLP) by demonstrating exceptional performance across a wide range of tasks, including text generation, question-answering, and reasoning. These models, typically built upon transformer architectures, have evolved significantly in recent years, driven by advances in training techniques, scaling strategies, and more efficient attention mechanisms. One of the early breakthroughs in LLMs was the development of models such as BERT and GPT, which introduced the encoder and decoder transformer architectures, respectively. BERT focused on bidirectional context representation, while GPT employed autoregressive decoding for generating coherent text. With models like GPT-3 and BERT-variants, LLMs scaled to hundreds of billions of parameters, vastly improving their ability to handle a diverse set of tasks through unsupervised pre-training on massive corpora followed by task-specific fine-tuning [7–10]. More recently, several surveys have explored the capabilities and limitations of LLMs. These works outline the critical role of pre-training on extensive datasets and highlight the importance of prompt engineering in guiding LLMs to perform specific tasks [11]. For instance, prompt-based learning techniques have enabled models like GPT-3 to perform well on zero-shot and few-shot tasks by utilizing simple textual cues [8]. Additionally, the use of chain-of-thought prompting and in-context learning has enhanced the reasoning capabilities of LLMs, making them applicable to complex tasks such as arithmetic reasoning and commonsense logic [12–14]. Several models have also been developed with specific goals in mind, such as improving multilingual capabilities. Research on multilingual LLMs like PaLM, BLOOM, and Mistral has demonstrated the importance of scaling and training on diverse language corpora to enhance performance across languages [15–17]. Furthermore, there has been growing interest in using LLMs for dense retrieval tasks, where models like BERT and T5 are used to retrieve and rank relevant documents from large datasets, highlighting the versatility of LLMs beyond traditional text generation tasks [18–21]. As the field progresses, newer models such as LLaMA and PaLM 2 aim to tackle issues of efficiency and scalability by adopting optimized architectures and better training techniques. These models have been applied to a variety of domains, including code generation, reasoning, and medical applications, demonstrating the broad applicability of LLMs [22–26]. Overall, the rapid development of LLMs has raised both excitement and concern within the NLP community. While the potential of these models is vast, challenges such as hallucination, energy consumption, and ethical considerations remain areas of active research [27]. Continued innovation in LLM training and evaluation will be crucial to addressing these challenges while pushing the boundaries of what LLMs can achieve.

 doi:10.20944/preprints202410.1658.v1

2.2. Retrieval-Augmented Large Language Models

Retrieval-Augmented Large Language Models (RAGs) have emerged as a powerful approach to addressing the limitations of traditional LLMs, particularly in overcoming hallucinations and keeping responses grounded in factual knowledge. Unlike standard LLMs, which rely solely on the data embedded in their pre-trained weights, RAGs are augmented with a retrieval mechanism that allows them to fetch relevant external information during inference, leading to more accurate and contextually appropriate outputs. Early work in this area focused on enhancing the performance of LLMs by integrating information retrieval techniques into the model architecture. For instance, Izacard et al. introduced retrieval-augmented models that leverage external documents to improve the model's ability to answer fact-based questions and perform few-shot learning tasks, showing significant improvements in accuracy and reliability over models like GPT-3 [28–30]. These approaches integrate document retrieval into the generation process, enabling the model to cross-reference and synthesize information from multiple sources. Recent surveys have further explored the effectiveness of RAGs across a variety of applications, highlighting the role of external data in resolving issues such as outdated knowledge and misinformation in LLMs. These surveys emphasize the flexibility of RAGs, particularly when used in combination with long-context models, showing that they can outperform purely long-context LLMs in scenarios involving large external data sources [31,32]. Additionally, advanced techniques like active retrieval-augmented generation allow models to iteratively refine their outputs by continuously retrieving more relevant documents, improving both accuracy and computational efficiency [33]. Another important development in this field is the introduction of toolkits such as RETA-LLM, which facilitate the development of retrieval-augmented models by providing modular components for document retrieval, passage extraction, and answer generation [34]. These toolkits make it easier for researchers to build RAGs tailored to specific domains, including medical and legal applications. Recent benchmarking efforts have led to the creation of the Retrieval-Augmented Generation Benchmark (RGB), which provides a systematic way to evaluate RAG models on tasks requiring external knowledge integration. This benchmark assesses models based on several dimensions, including noise robustness and information integration, ensuring that they can handle noisy or incomplete data effectively [35]. In summary, RAGs represent a promising direction for improving the factuality and reliability of LLMs, particularly in knowledge-intensive domains. By combining the strengths of retrieval systems with the generative capabilities of LLMs, these models can address some of the key limitations that have historically affected purely generative models.

2.3. Medical Large Language Models

The emergence of Medical Large Language Models (Med-LLMs) has revolutionized healthcare by enabling advanced language processing for a variety of medical applications, including clinical decision support, patient communication, and medical education. These models, typically fine-tuned versions of general-purpose LLMs, have shown significant potential in interpreting complex medical texts, generating accurate diagnoses, and answering medical queries. However, they also face unique challenges such as ensuring data privacy, maintaining factual accuracy, and mitigating risks related to hallucinations and ethical concerns. Several surveys have highlighted the broad range of applications and challenges for Med-LLMs. For instance, a comprehensive review by [36] focuses on the principles, applications, and challenges of Med-LLMs, including their use in diagnosis, clinical report generation, and decision-making support. This work also emphasizes the importance of trustworthiness and interpretability, especially when Med-LLMs are integrated into clinical workflows. Similarly, [37] provides an overview of the technological aspects of Med-LLMs, exploring their application in tasks such as medical language translation and medical robotics, while also addressing the ethical and legal implications of using LLMs in healthcare. Another major development in this field is the fine-tuning of existing LLMs, such as BLOOM and PaLM, on medical datasets to create domain-specific models. For example, [38] introduced ClinicalGPT, a fine-tuned model designed for medical conversations and diagnostic tasks, showing improvements over general LLMs in medical dialogue systems. Similarly,

 doi:10.20944/preprints202410.1658.v1

[39] evaluated Med-PaLM 2 on medical question-answering benchmarks, demonstrating that fine-tuning general LLMs on specialized medical datasets significantly enhances their performance in professional medical tasks. Furthermore, the integration of external knowledge bases through retrieval-augmented techniques has been explored to improve the accuracy of Med-LLMs. [40] proposed Medical Graph RAG, a framework that uses graph-based retrieval to provide safe and reliable medical recommendations, addressing the issue of outdated or inaccurate information in LLM responses. Despite these advancements, there remain substantial challenges in scaling Med-LLMs for real-world use. Issues such as hallucinations, biased decision-making, and privacy concerns need to be addressed before widespread clinical adoption can be achieved [37,38]. Moreover, ethical considerations such as accountability and fairness, particularly in patient outcomes, are critical in ensuring that these models are trustworthy and beneficial in clinical practice.

3. Method

3.1. Model Architecture

Let $x \in \mathbb{R}^n$ represent an input biomedical query, and let the LLM generate a corresponding answer $\hat{y}$. The generative process of the model can be formulated as maximizing the likelihood of the correct output y given the input x, which is mathematically expressed as:

$$\max_{\theta} \log p_\theta(y|x) \tag{1}$$

where θ represents the parameters of the model. During training, the model is fine-tuned using a mixture of biomedical documents and medical question-answer datasets, with the aim of adapting the pre-trained LLM's parameters to domain-specific tasks.

3.2. Two-Stage Fine-Tuning

The training of the proposed model consists of two distinct stages:

1. Domain-Specific Pre-Training: In this stage, we pre-train the LLM on a large corpus $\mathcal{D}$ of biomedical texts, including textbooks, research papers, and clinical guidelines. The objective is to maximize the likelihood of text generation given the domain-specific inputs, which is formulated as:

$$\mathcal{L}_{\text{pre-train}} = -\sum_{(x,y)\in\mathcal{D}} \log p_\theta(y|x) \tag{2}$$

This stage enables the model to internalize critical biomedical knowledge, ensuring that it has a foundational understanding of domain-specific concepts before being fine-tuned on specific tasks.

2. Task-Specific Fine-Tuning with Self-Reflection Loss: Once pre-training is complete, we fine-tune the model on biomedical question-answering datasets such as MedQA, MedMCQA, and MMLU. The training loss during this stage is modified to introduce a self-reflection loss, which encourages the model to evaluate its own outputs. Let $\hat{y}$ be the model's predicted answer, and y be the ground truth. The task-specific loss consists of two components:

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{QA}} + \lambda \mathcal{L}_{\text{self-reflect}} \tag{3}$$

The QA loss $\mathcal{L}_{\text{QA}}$ is a standard cross-entropy loss that compares the predicted answer $\hat{y}$ with the ground truth y:

$$\mathcal{L}_{\text{QA}} = -\sum_{(x,y)} \log p_\theta(\hat{y}|x) \tag{4}$$

 doi:10.20944/preprints202410.1658.v1

The self-reflection loss $\mathcal{L}_{\text{self-reflect}}$ penalizes the model if its output deviates from expected reasoning standards. Specifically, the model generates a self-assessment score $\hat{s}$ for each predicted answer, which is compared to a target score s (evaluating how close the answer is to the ground truth):

$$\mathcal{L}_{\text{self-reflect}} = \sum_{(x,y)} (\hat{s} - s)^2 \tag{5}$$

where λ is a hyperparameter that balances the importance of the QA loss and the self-reflection loss.

3.3. Self-Reflection Mechanism

The self-reflection mechanism plays a key role in improving the model's reasoning ability. After generating an initial output $\hat{y}$, the model generates a self-assessment score $\hat{s}$ by evaluating its own answer through a secondary pass over the generated output. The self-assessment process can be described as:

$$\hat{s} = f_{\text{reflect}}(\hat{y}, x) \tag{6}$$

where f_{reflect} represents a reflection function implemented as a sub-module of the LLM that reviews both the generated output and the original input query. This process encourages the model to produce more accurate and well-reasoned answers by explicitly scoring the relevance and correctness of its output.

3.4. Optimization Objective

The final objective of the model is to minimize the combined loss function, incorporating both the QA loss and the self-reflection loss. The complete optimization problem can be summarized as:

$$\min_{\theta} \mathcal{L}_{\text{task}} = \min_{\theta} (\mathcal{L}_{\text{QA}} + \lambda \mathcal{L}_{\text{self-reflect}}) \tag{7}$$

By incorporating the self-reflection mechanism, we ensure that the model not only generates accurate answers but also learns to evaluate the quality of its own responses, reducing the likelihood of generating erroneous or incomplete answers.

3.5. Training Strategy

We adopt a standard Adam optimizer with learning rate scheduling during the training process. The domain-specific pre-training is conducted over large biomedical corpora, while the task-specific fine-tuning stage is performed on the question-answer datasets. The self-reflection mechanism is applied iteratively, with the model performing a second pass over its outputs to compute self-assessment scores. This iterative approach allows the model to refine its output during training, ensuring more accurate predictions in biomedical question-answering tasks.

4. Experiments

To validate the effectiveness of our proposed approach, we conducted a series of experiments comparing our method, MedRAG-Refine, with multiple state-of-the-art models, including Self-BioRAG, RAG, and LLaMA2. These models were evaluated on various biomedical question-answering datasets, such as MedQA, MedMCQA, and the medical subset of MMLU. Our experiments aim to showcase the performance improvements achieved by our method, particularly in complex medical reasoning tasks.

4.1. Dataset and Experimental Setup

We evaluated all models on three datasets:

1. MedQA: A dataset focusing on medical licensing examination questions.
2. MedMCQA: A multiple-choice medical QA dataset covering a wide range of medical subjects.

 doi:10.20944/preprints202410.1658.v1

3. MMLU (Medical subset): A dataset that includes medical questions for evaluating general reasoning skills in the biomedical domain.

All models were trained and evaluated on the same sets of data to ensure a fair comparison. The metrics used for evaluation include accuracy and reasoning quality, which capture how well each model understands and answers biomedical questions.

4.2. Comparative Results

The experimental results indicate that our method outperforms the baseline models in all three datasets. Below is a summary of the accuracy results in table format:

Table 1. Accuracy comparison between our method and other models on MedQA, MedMCQA, and MMLU (Medical subset).

Model	Params	MedQA	MedMCQA	MMLU (Med)
MedRAG-Refine (7B)	7B	**44.8%**	**43.5%**	**55.1%**
MedRAG-Refine (13B)	13B	**50.1%**	**45.6%**	**59.3%**
Self-BioRAG (7B)	7B	43.6%	42.1%	53.9%
RAG (7B)	7B	36.2%	38.3%	47.7%
LLaMA2 (7B)	7B	35.2%	36.3%	46.3%

The results demonstrate that MedRAG-Refine consistently outperforms other models across all datasets. Notably, the accuracy improvement is more pronounced in the MMLU medical subset, highlighting the model's ability to handle general biomedical reasoning tasks.

4.3. Ablation Study

To further validate the effectiveness of our proposed self-reflection loss, we conducted an ablation study where the self-reflection component was removed. The comparison is shown below:

Table 2. Ablation study comparing the full model and the version without the self-reflection mechanism.

Model	MedQA	MedMCQA	MMLU (Med)
MedRAG-Refine (Full)	**44.8%**	**43.5%**	**55.1%**
MedRAG-Refine (No Self-Reflect)	41.5%	40.3%	50.6%

As seen from the table, removing the self-reflection mechanism results in a noticeable drop in accuracy across all datasets, confirming the importance of this component in improving the model's reasoning capabilities.

4.4. Human Evaluation

In addition to automatic metrics, we also performed a human evaluation to assess the quality of reasoning and correctness of answers generated by the models. A team of medical professionals was asked to evaluate a random subset of answers generated by each model based on accuracy, relevance, and reasoning quality. The results are presented in the following table:

Table 3. Human evaluation of generated answers across different models based on accuracy, relevance, and reasoning quality.

Model	Accuracy	Relevance	Reasoning Quality
MedRAG-Refine (7B)	**92%**	**88%**	**90%**
Self-BioRAG (7B)	85%	82%	84%
RAG (7B)	78%	75%	76%
LLaMA2 (7B)	73%	70%	72%

 doi:10.20944/preprints202410.1658.v1

The human evaluation results align with the automatic metrics, showing that MedRAG-Refine generates more accurate, relevant, and higher-quality answers compared to other models. The evaluation underscores the efficacy of our method in real-world medical question-answering scenarios, where reasoning quality is critical.

4.5. Discussion

Our experiments demonstrate that the introduction of the self-reflection loss in the training process significantly improves both the automatic and human-evaluated performance of the model. By embedding domain-specific knowledge and reasoning mechanisms into the model, MedRAG-Refine achieves better results across various medical question-answering datasets. The consistent improvement across both automatic and human evaluation metrics suggests that our approach is more suitable for practical biomedical applications, particularly in settings where real-time, accurate responses are crucial.

5. Conclusion

In this work, we introduced MedRAG-Refine, a retrieval-independent large language model optimized for biomedical applications. By embedding domain-specific knowledge through a two-stage fine-tuning process and enhancing reasoning quality with a self-reflection mechanism, our model significantly improves both accuracy and relevance in medical question-answering tasks. Comparative experiments with state-of-the-art models confirmed the superiority of MedRAG-Refine, both in automated evaluations and human assessments. These results indicate that our approach not only advances medical reasoning but also demonstrates strong potential for real-world, real-time clinical applications.

References

1. Li, M.; Kilicoglu, H.; Xu, H.; Zhang, R. BiomedRAG: A Retrieval Augmented Large Language Model for Biomedicine. *CoRR* **2024**, *abs/2405.00465*, [2405.00465]. doi:10.48550/ARXIV.2405.00465.
2. Jeong, M.; Sohn, J.; Sung, M.; Kang, J. Improving Medical Reasoning through Retrieval and Self-Reflection with Retrieval-Augmented Large Language Models. *CoRR* **2024**, *abs/2401.15269*, [2401.15269]. https://doi.org/10.48550/ARXIV.2401.15269.
3. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Shen, J.; Long, G.; Xu, C.; Jiang, D. Fine-grained distillation for long document retrieval. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 19732–19740.
4. Li, M.; Zhan, Z.; Yang, H.; Xiao, Y.; Huang, J.; Zhang, R. Benchmarking Retrieval-Augmented Large Language Models in Biomedical NLP: Application, Robustness, and Self-Awareness. *CoRR* **2024**, *abs/2405.08151*, [2405.08151]. doi:10.48550/ARXIV.2405.08151.
5. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024; Ku, L.; Martins, A.; Srikumar, V., Eds. Association for Computational Linguistics, 2024, pp. 6233–6251. doi:10.18653/V1/2024.FINDINGS-ACL.372.
6. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; Jiang, D. Towards Robust Ranker for Text Retrieval. Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5387–5401.
7. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers); Burstein, J.; Doran, C.; Solorio, T., Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.
8. Wang, Z.; Li, M.; Xu, R.; Zhou, L.; Lei, J.; Lin, X.; Wang, S.; Yang, Z.; Zhu, C.; Hoiem, D.; Chang, S.; Bansal, M.; Ji, H. Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners. Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems

2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.

9. Zhou, Y.; Shen, T.; Geng, X.; Long, G.; Jiang, D. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2559–2575.
10. Zhou, Y.; Geng, X.; Shen, T.; Long, G.; Jiang, D. Eventbert: A pre-trained model for event correlation reasoning. Proceedings of the ACM Web Conference 2022, 2022, pp. 850–859.
11. Sanh, V.; Webson, A.; Raffel, C.; Bach, S.H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M.S.; Xu, C.; Thakker, U.; Sharma, S.S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N.V.; Datta, D.; Chang, J.; Jiang, M.T.; Wang, H.; Manica, M.; Shen, S.; Yong, Z.X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Févry, T.; Fries, J.A.; Teehan, R.; Scao, T.L.; Biderman, S.; Gao, L.; Wolf, T.; Rush, A.M. Multitask Prompted Training Enables Zero-Shot Task Generalization. The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
12. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
13. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.H.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
14. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022; Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; Oh, A., Eds., 2022.
15. Zhou, Y.; Geng, X.; Shen, T.; Zhang, W.; Jiang, D. Improving zero-shot cross-lingual transfer for multilingual question answering over knowledge graph. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 5822–5834.
16. Gao, Y.; Hou, F.; Wang, R. A Novel Two-step Fine-tuning Framework for Transfer Learning in Low-Resource Neural Machine Translation. Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024; Duh, K.; Gómez-Adorno, H.; Bethard, S., Eds. Association for Computational Linguistics, 2024, pp. 3214–3224. doi:10.18653/V1/2024.FINDINGS-NAACL.203.
17. Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilic, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; Tow, J.; Rush, A.M.; Biderman, S.; Webson, A.; Ammanamanchi, P.S.; Wang, T.; Sagot, B.; Muennighoff, N.; del Moral, A.V.; Ruwase, O.; Bawden, R.; Bekman, S.; McMillan-Major, A.; Beltagy, I.; Nguyen, H.; Saulnier, L.; Tan, S.; Suarez, P.O.; Sanh, V.; Laurençon, H.; Jernite, Y.; Launay, J.; Mitchell, M.; Raffel, C.; Gokaslan, A.; Simhi, A.; Soroa, A.; Aji, A.F.; Alfassy, A.; Rogers, A.; Nitzav, A.K.; Xu, C.; Mou, C.; Emezue, C.; Klamm, C.; Leong, C.; van Strien, D.; Adelani, D.I.; et al.. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR* **2022**, *abs/2211.05100*, [2211.05100]. doi:10.48550/ARXIV.2211.05100.
18. Zhou, Y.; Geng, X.; Shen, T.; Pei, J.; Zhang, W.; Jiang, D. Modeling event-pair relations in external knowledge graphs for script reasoning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* **2021**.
19. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
20. Fan, C.; Yan, Z.; Wu, Y.; Qian, B. Span prompt dense passage retrieval for Chinese open domain question answering. *J. Intell. Fuzzy Syst.* **2023**, *45*, 7285–7295. doi:10.3233/JIFS-231328.
21. Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; Grave, E. Towards Unsupervised Dense Information Retrieval with Contrastive Learning. *CoRR* **2021**, *abs/2112.09118*, [2112.09118].
22. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. *CoRR* **2023**, *abs/2302.13971*, [2302.13971]. doi:10.48550/ARXIV.2302.13971.
23. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.;

Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A.M.; Pillai, T.S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; Fiedel, N. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.* **2023**, *24*, 240:1–240:113.

24. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3445–3456.
25. Zhou, Y.; Tao, W.; Zhang, W. Triple sequence generative adversarial nets for unsupervised image captioning. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7598–7602.
26. Zhou, Y. Sketch storytelling. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 4748–4752.
27. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021; Elish, M.C.; Isaac, W.; Zemel, R.S., Eds. ACM, 2021, pp. 610–623. doi:10.1145/3442188.3445922.
28. Bhattarai, M.; Santos, J.E.; Jones, S.; Biswas, A.; Alexandrov, B.S.; O'Malley, D. Enhancing Code Translation in Language Models with Few-Shot Learning via Retrieval-Augmented Generation. *CoRR* **2024**, *abs/2407.19619*, [2407.19619]. doi:10.48550/ARXIV.2407.19619.
29. Zhou, Y.; Long, G. Multimodal Event Transformer for Image-guided Story Ending Generation. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3434–3444.
30. Zhou, Y.; Long, G. Style-Aware Contrastive Learning for Multi-Style Image Captioning. Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 2257–2267.
31. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Zhong, S.; Yin, B.; Hu, X. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data* **2024**, *18*, 1–32.
32. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada; Wooldridge, M.J.; Dy, J.G.; Natarajan, S., Eds. AAAI Press, 2024, pp. 17754–17762. doi:10.1609/AAAI.V38I16.29728.
33. Cheng, Q.; Li, X.; Li, S.; Zhu, Q.; Yin, Z.; Shao, Y.; Li, L.; Sun, T.; Yan, H.; Qiu, X. Unified Active Retrieval for Retrieval Augmented Generation. *CoRR* **2024**, *abs/2406.12534*, [2406.12534]. doi:10.48550/ARXIV.2406.12534.
34. Liu, J.; Jin, J.; Wang, Z.; Cheng, J.; Dou, Z.; Wen, J. RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit. *CoRR* **2023**, *abs/2306.05212*, [2306.05212]. doi:10.48550/ARXIV.2306.05212.
35. Chen, J.; Lin, H.; Han, X.; Sun, L. Benchmarking large language models in retrieval-augmented generation. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 17754–17762.
36. Zhou, H.; Liu, F.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S.S.; Zhou, P.; Liu, J.; Hua, Y.; Mao, C.; You, C.; Wu, X.; Zheng, Y.; Clifton, L.; Li, Z.; Luo, J.; Clifton, D.A. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge, 2024, [arXiv:cs.CL/2311.05112].
37. Liu, L.; Yang, X.; Lei, J.; Liu, X.; Shen, Y.; Zhang, Z.; Wei, P.; Gu, J.; Chu, Z.; Qin, Z.; Ren, K. A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions. *CoRR* **2024**, *abs/2406.03712*, [2406.03712]. doi:10.48550/ARXIV.2406.03712.
38. Wang, G.; Yang, G.; Du, Z.; Fan, L.; Li, X. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation, 2023, [arXiv:cs.CL/2306.09968].

 doi:10.20944/preprints202410.1658.v1

39. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.A.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B.A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S.S.; Barral, J.K.; Webster, D.R.; Corrado, G.S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; Natarajan, V. Towards Expert-Level Medical Question Answering with Large Language Models. *CoRR* **2023**, *abs/2305.09617*, [2305.09617]. doi:10.48550/ARXIV.2305.09617.
40. Wu, J.; Zhu, J.; Qi, Y. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation. *CoRR* **2024**, *abs/2408.04187*, [2408.04187]. doi:10.48550/ARXIV.2408.04187.