SparkRA: A Retrieval-Augmented Knowledge Service System Based on Spark Large Language Model

Dayong Wu[1], Jiaqi Li[1,2], Baoxin Wang[1,3], Honghong Zhao[1], Siyuan Xue[1], Yanjie Yang[1], Zhijun Chang[4], Rui Zhang[1,5], Li Qian[4], Bo Wang[1,5], Shijin Wang[1], Zhixiong Zhang[4], Guoping Hu[1]

1. State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
2. University of Science and Technology of China, Hefei, China.
3. Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, Harbin, China.
4. National Science Library, Chinese Academy of Sciences, China.
5. iFLYTEK AI Research (Hebei), Langfang, China.

Abstract

Large language models (LLMs) have shown remarkable achievements across various language tasks. To enhance the performance of LLMs in scientific literature services, we developed the scientific literature LLM (SciLit-LLM) through pre-training and supervised fine-tuning on scientific literature, building upon the iFLYTEK Spark LLM. Furthermore, we present a knowledge service system Spark Research Assistant (SparkRA) based on our SciLit-LLM. SparkRA is accessible online[1] and provides three primary functions: literature investigation, paper reading, and academic writing. As of July 30, 2024, SparkRA has garnered over 50,000 registered users, with a total usage count exceeding 1.3 million.

1 Introduction

Large language models (LLMs) have achieved significant success in natural language processing, including text generation and language understanding (Brown et al., 2020; Chowdhery et al., 2023). Owing to their strong capabilities, LLMs have shown immense potential across many downstream fields, such as education, medicine, and finance (Kasneci et al., 2023; Thirunavukarasu et al., 2023; Clusmann et al., 2023; Shah et al., 2023).

As the performance of LLMs in scientific literature does not fully meet the needs of scholars, we developed a Scientific Literature LLM (SciLit-LLM). We began by collecting a large dataset of scientific literature, including academic papers and patents, and performed data cleaning to ensure high-quality academic text. We then continued pre-training the open-source iFLYTEK Spark LLM (13B)[2] using an autoregressive training task, followed by supervised fine-tuning, to create our SciLit-LLM.

[1]https://paper.iflytek.com/

[2]https://gitee.com/iflytekopensource/iFlytekSpark-13B

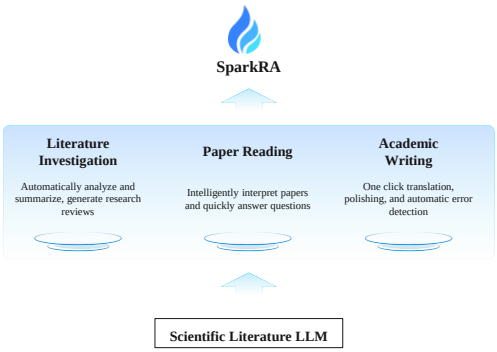

Figure 1: The process of building SparkRA system.

Traditional knowledge service systems generally provide limited functionalities, such as the retrieval of scholarly articles and assistive reading services. In this paper, we introduce the Spark Research Assistant (SparkRA), a knowledge service system based on our scientific literature LLM. SparkRA offers a comprehensive, one-stop solution for scientific literature services. Figure 1 depicts the process of constructing the SparkRA system. The features of SparkRA are as follows:

- Literature investigation: this sub-system can automatically analyze and summarize research areas, and generate research reviews.
- Paper reading: this sub-system can intelligently interpret papers and quickly answer questions.
- Academic writing: this sub-system can provide the functions for writing academic papers including one-click translation, polishing, and automatic error detection.

Experimental evaluation demonstrates that SparkRA outperforms existing models, including GPT-3.5 and Llama3-8B, across all tasks, establishing its efficacy in enhancing the productivity and accuracy of academic research activities.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing
System Demonstrations, pages 382–389
November 12-16, 2024

2 Scientific Literature LLM

2.1 Base model

To build the LLM for scientific literature services, we selected the Spark LLM as the foundation model for building our scientific literature LLM (ScLlit-LLM). The Spark LLM, developed by iFLYTEK Research, demonstrates impressive performance in processing both English and Chinese languages. iFlytekSpark-13B has consistently ranked among the top in numerous well-known public benchmarks, demonstrating its superiority. Its performance is notably superior to other open-source models of equivalent size.

2.2 Continual pre-training

While the Spark LLM exhibits strong capabilities in language comprehension and text generation, it may struggle to directly provide accurate responses to scholarly inquiries without targeted training in the scientific domain. Consequently, we have designed a Scientific literature LLM that is specifically oriented towards parsing and understanding scientific literature.

Inspired by the existing research (Beltagy et al., 2019; Hong et al., 2022), we have further pretrained the spark model on an extensive corpus of academic texts to enhance the model's performance in processing and generating scientific literature

Data preparation. To enhance the foundational large language model (LLM), it is imperative to amass a vast corpus of high-quality data, which includes kinds of scholarly literature like papers and patents. We collected a vast number of academic papers from various publicly accessible websites, such as arXiv[3].

Given that academic documents are predominantly archived in PDF format, it is crucial to convert these PDFs into text while meticulously eliminating any extraneous elements. For this purpose, we employed a sophisticated PDF parsing tool developed by iFLYTEK. In the process of advancing our scientific literature LLM, we have incorporated a dataset comprising over 10M academic papers.

To prevent LLM from losing its general capabilities, we also incorporated a significant amount of general corpora. This strategy ensures that after continual pre-training, the scientific literature LLM performs better in the field of science while maintaining the general capabilities.

[3]https://arxiv.org/

Pre-training. Similar to the traditional LLM pretraining process, the scientific literature LLM employs the same next-word prediction task for its continual pre-training on a corpus of scientific literature comprising billions of tokens.

Upon evaluation, the scientific literature LLM, continual pre-training, exhibits improved performance on general scholarly inquiries. Moreover, for specialized academic queries without provided context, the scientific literature LLM demonstrates a higher rejection tendency, effectively reducing instances of hallucination.

2.3 Supervised fine-tuning

Supervised fine-tuning (SFT) is a technique used to enhance large language models (LLMs) by further training a pre-trained model to improve its accuracy and relevance for specific tasks or domains. The efficacy of SFT in refining LLMs is well-documented (Wei et al., 2022; Ouyang et al., 2022). This process involves utilizing a carefully curated dataset with labeled examples that illustrate the desired output. During SFT, the model learns from these examples to comprehend the intricacies of the task more thoroughly. Consequently, SFT enables the model to retain its broad knowledge base while acquiring specialization in targeted areas, resulting in enhanced user experiences and more precise information delivery.

Data preparation. In the construction of our datasets for supervised fine-tuning, each instance within datasets is composed of three elements: an instruction, an input, and an output. We utilize a dual approach in formulating instructions, leveraging both Self-instruct (Wang et al., 2023b) and human writing.

To exemplify, consider the instruction: "Please translate the input English sentence into Chinese"; here, the input component would be an English sentence. For the generation of outputs corresponding to given instructions and inputs, we employ meticulously devised manual methods to craft expert responses.

Training. Upon completing the construction of SFT datasets, we commenced the Supervised Fine-Tuning (SFT) of scientific literature LLM. The instances within the dataset serve as labeled data for the SFT of the model. Since each instance is meticulously crafted by experts, they are of higher quality compared to the generic data used during the

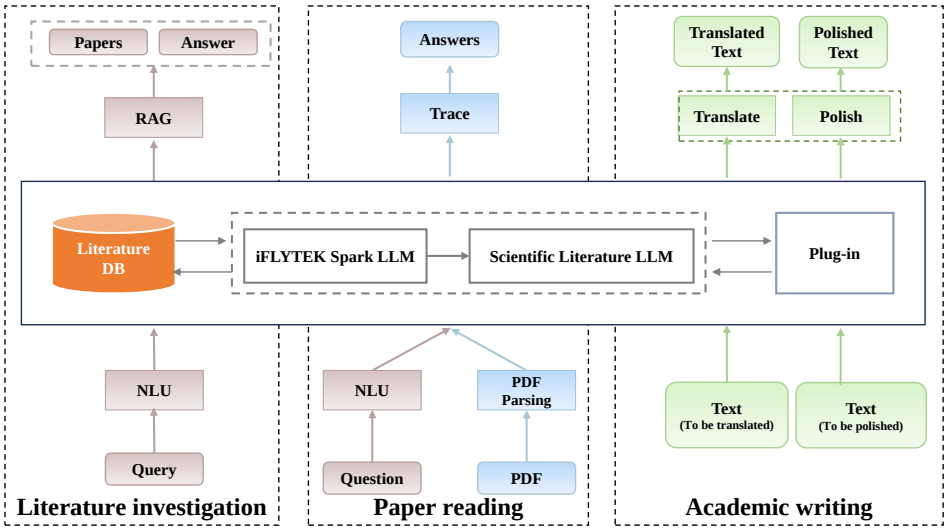

Figure 2: The system architecture of SparkRA integrates iFLYTEK Spark LLM and Scientific Literature LLM to facilitate literature investigation, paper reading, and academic writing.

pre-training phase. Moreover, these labeled data enhance the LLM's ability to answer questions. The scientific literature LLM that has undergone SFT with domain-specific data can learn from experts' responses to research-related inquiries and generalize this knowledge to a broader array of questions.

3 SparkRA

Based on our SciLit-LLM, we developed a literature services system SparkRA. This platform is comprised of three functions: literature investigation, paper reading, and academic writing. Notably, SparkRA is equipped to process inputs in both Chinese and English, thereby catering to a diverse linguistic user base. The architecture of SparkRA is shown in Figure 2 and the demonstration video has been published on YouTube[4].

3.1 Literature investigation

This function is designed to facilitate the exploration of academic literature and is comprised of three integral components: an investigation copilot, a research topic search engine, and a review generation module. The architecture and screenshot of the literature investigation function are respectively shown in Figure 3 and Figure 4.

Investigation copilot. This copilot assists users in deepening their understanding of specific research domains and various scholars through interactive natural language dialogue.

[4]https://youtu.be/bdUMTr3pMfY

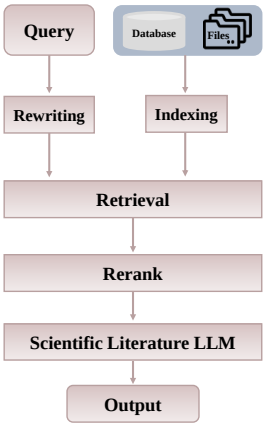

Figure 3: The architecture of RAG-based literature investigation.

(1) Area-based survey. Users can easily obtain the summarization and papers of a specific research area. For example, the user can send the query "What are the recent papers of fake news section in 2023". SparkRA will show the papers and give a summary.

(2) Scholar-based survey. This function can output the papers of the input scholar and divide the papers into different research areas. For example, the user can send the query "What research has Chris Manning from Stanford University conducted".

Topic search engine. The search interface accommodates queries pertaining to research topics in both Chinese and English. Upon receiving a specified topic, SparkRA retrieves relevant papers from an extensive academic library and provides

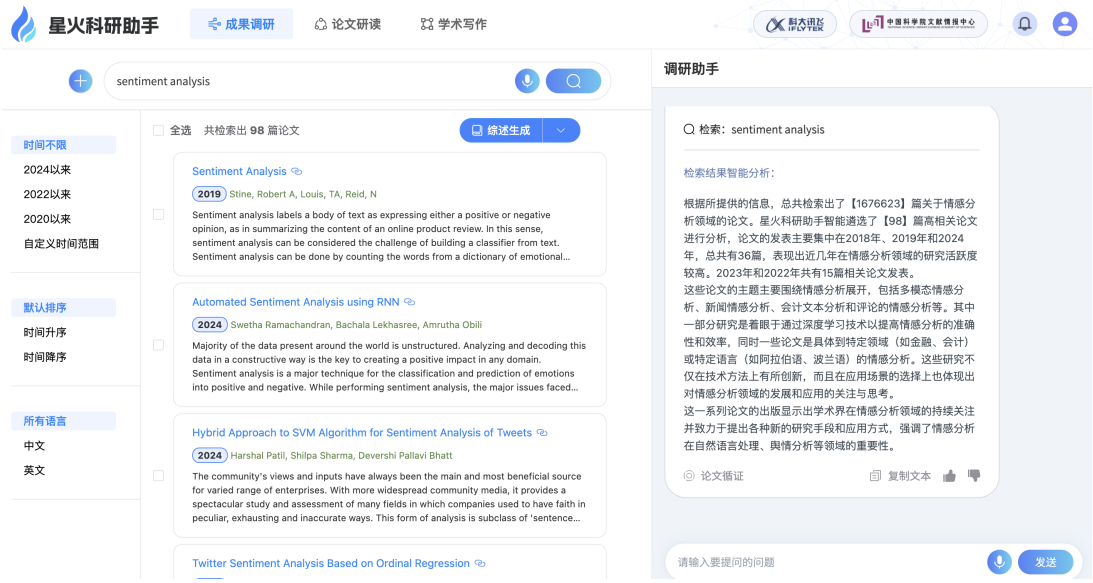

Figure 4: Literature investigation page.

concise summaries of their content.

(1) Query rewriting. There is a diversity of user retrieval query formats and the occasional inclusion of noise, such as "In the library, what LLM technologies can assist users in improving the efficiency of finding books?". Upon receiving a user's query, scientific literature LLM is used to revise the query into a format more suited for retrieval, like "Applications of large models in library search domain". This strategy can significantly enhance the system's ability to locate the desired literature.

(2) Precise Retrieval. Upon completion of the rewriting process, the revised query is subjected to information extraction through natural language understanding technologies, such as Named Entity Recognition (NER). The extracted information encompasses scholars, institutions, dates, domains, and keywords, among others. Based on the extracted content, the corresponding search plugin interfaces are invoked to obtain precise search results.

(3) Literature-based summary. Building on the retrieval outcomes, the scientific literature LLM synthesizes findings, encompassing the distribution of publication years, trends in literature popularity, recent focal topics, and potential future directions of development.

Review generation. This function enables the generation of a report based on a selection of papers, with a maximum limit of 30 papers. The generated report facilitates an expedited comprehension of a substantial volume of literature within a specific domain or authored by an individual.

In this function, we leveraged the clustering capabilities and inductive summarization prowess of LLM. Through the clustering of dozens of literature papers, the model structured the introduction, body, and conclusion of a comprehensive review, including the formulation of pertinent headings. Subsequently, the model demonstrated its robust capacity for inductive reasoning and summarization. It also featured the capability to annotate the analytical text with hyperlinks, serving as citations that facilitate reference validation at the end of the review and enable user verification.

3.2 Paper reading

This function can assist scholars and students in reading academic papers. With the rapid development of artificial intelligence technology, a large number of cutting-edge papers emerge every day. It is necessary to develop an intelligent system to help people understand papers.

For paper reading, LLMs with longer context windows are required because the full article of paper is usually long. However, training an LLM with long context windows from scratch requires significantly larger investments. To facilitate this, we employ a retrieval-augmented approach to enhance the effectiveness of the large model's answers. We initiate text splitting as a primary step and engage in chapter recognition to preserve the semantic integrity of segments. For the cross-language retrieval embedding model, firstly, we generate questions from paper segments using an LLM and con-

struct a large set of (question, positive sample, negative samples) pairs for training. Subsequently, we use XLM-RoBERTa (Conneau et al., 2020) as the language encoder and fine-tune the model via contrastive learning. The input question and retrieved segments are finally fed into the SciLit-LLM to generate answers.

Reading Copilot enhances paper comprehension through natural language interactions. Questions fall into two categories: those within the paper, which SciLit-LLM answers using the input paper alone, and those outside the paper, which require a search engine plugin to retrieve relevant information. For the latter, answers are generated through retrieval-augmented generation using SciLit-LLM.

Multi-Document Comparison allows for the comparison of two to five papers. For each selected paper, SparkRA provides the abstract and contributions separately. It also generates a comparative analysis table that highlights the proposed approaches and advantages of each paper. SparkRA can identify and output both the similarities and differences among the selected papers.

3.3 Academic writing

This function is directly powered by SciLit-LLM and includes polishing and translation.

Paper polishing. This function is used to assist the scholar and students in polishing the academic paper draft. We construct a large corpus of texts requiring polishing based on a multitude of well-written academic papers, utilizing few-shot learning and chain-of-thought (COT) prompting methodologies, followed by supervised learning for instruction fine-tuning.

Academic translation. In order to accurately translate domain-specific terminology, we have implemented a dynamic perception prompts approach to guide the model in completing translation tasks. Based on the user's input prompts, we obtain prompts with professional terminology translations from a terminology translation lexicon in the knowledge base, which are then fed into the large language model.

4 Experiments

4.1 Experiment setting

To validate the results of SparkRA, we adopt the following LLMs as the baseline models:

- Llama: a large-scale language model developed and open-sourced by Meta, was compared to SciLit-LLM using three versions: Llama2-7B, Llama2-13B, and Llama3-8B.
- ChatGPT (GPT-3.5): it is a large-scale language model in the field of artificial intelligence developed by OpenAI.
- GPT-4: GPT-4 Turbo serves as our baseline model, consistently outperforming in a range of NLP tasks.

We evaluate the performance of models using the mean opinion score (MOS) on a scale of 1 (poorest) to 5 (optimal), with evaluations conducted by more than five individuals per task. For the machine translation task, we also use the BLEU metric (Papineni et al., 2002) for model evaluation. We gathered 100 academic parallel paragraphs from public Chinese journals with Chinese and English abstracts to serve as test sets. The highest results in the table are highlighted in bold, and the second-highest results are underlined.

To assess paper reading performance, we employ following two measures:

- Factuality: evaluates the accuracy of the system's response to factual information;
- Informativeness: assesses the completeness of the system's response.

To evaluate paper polishing and academic translation performance, we use three criteria:

- Fluency: assesses the language coherence of model's outputs;
- Fidelity: measures content faithfulness to the original text;
- Academic: evaluates adherence to academic language standards.

4.2 Results

The results of the paper reading are shown in Table 1. SparkRA outperforms other models across all metrics. It achieves the highest score in Factuality with a score of 4.68, surpassing the closest competitor, GPT-4, which scores 4.67. In terms of Informativeness, SparkRA attains a score of 4.45, again leading over GPT-4, which scores 4.43. Overall, SparkRA achieves the highest average score of 4.57, demonstrating superior performance compared to other models like Llama3-8B and Spark

	Factuality	Informativeness	Avg.
Llama2-7B	3.98	3.50	3.74
Llama2-13B	4.47	3.72	4.10
Llama3-8B	4.63	4.19	4.41
GPT-3.5	4.20	3.97	4.09
GPT-4	4.67	4.43	4.55
SparkRA	**4.68**	**4.45**	**4.57**

Table 1: Results of paper reading task.

	Fluency	Fidelity	Academic	Avg.
Llama2-7B	**4.59**	3.94	4.44	4.32
Llama2-13B	**4.59**	3.53	4.06	4.06
Llama3-8B	4.56	3.97	4.47	4.33
GPT-3.5	4.26	4.23	4.38	4.29
GPT-4	4.26	4.29	4.41	4.32
SparkRA	4.41	**4.45**	**4.61**	**4.49**

Table 2: Results of paper polishing task.

v3. These results underscore SparkRA's effectiveness in producing factually accurate and informative text, establishing it as a state-of-the-art model in the paper reading task.

Table 2 shows the results of the paper polishing task. While Llama2-13B generates coherent text, it struggles with fidelity due to non-existent elements. Although Spark v3 performs well across tasks, our SparkRA model, pre-trained on scientific literature and fine-tuned with 13 billion parameters, shows even greater improvement. SparkRA achieves state-of-the-art results compared to widely used LLMs like GPT-3.5 and GPT-4 across all evaluation metrics, excelling particularly in academic relevance.

Table 3 presents the academic translation results. SparkRA excels with the highest fidelity score (4.91) and the second-highest academic quality (4.75), showcasing its superior ability to preserve meaning and produce contextually appropriate translations. Additionally, SparkRA's BLEU score of 0.198 reflects its robustness in both human and automatic evaluations. Despite lower human evaluation scores than GPT-4, SparkRA's 13B parameter size offers flexibility, ease of training, and cost-effectiveness.

5 Related Work

Scientific literature pre-trained language model Since the release of the pre-trained models (Vaswani et al., 2017; Radford et al., 2018; Devlin et al., 2019), the language models for scientific literature have attracted the attention of scholars. These models are trained on various scientific datasets, with SciBERT on PubMed Central (Beltagy et al., 2019), BioBERT and BioMegatron on biomedical literature (Lee et al., 2020; Shin et al., 2020), Galactica on multilingual articles (Taylor et al., 2022), and ScholarBERT on ACL Anthology Corpus (Hong et al., 2022).

	Fluency	Fidelity	Academic	Avg.	BLEU
Llama2-7B	4.53	3.93	4.13	4.20	0.104
Llama2-13B	**4.73**	4.03	4.33	4.36	0.116
Llama3-8B	4.64	4.46	4.43	4.51	0.168
GPT-3.5	4.41	4.75	4.54	4.57	0.193
GPT-4	4.50	4.88	**4.84**	**4.74**	0.180
SparkRA	4.34	**4.91**	4.75	4.67	0.198

Table 3: Results of academic translation task.

Retrieval augmented generation to LLM Retrieval-Augmented Generation (RAG), introduced by Lewis et al. (2020), mitigates hallucinations in Large Language Models (LLMs) by integrating external data. Ma et al. (2023) advanced RAG with query rewriting, while Chen et al. (2023) benchmarked its effects, creating the RGB. Lyu et al. (2023) developed an algorithm for assessing retrieved data significance.

AI for science Artificial intelligence has significantly impacted scientific research, enhancing efficiency and literature growth (Merchant et al., 2023; Szymanski et al., 2023). Wang et al. (2023a) proposed an AI-based scientific research method that can automatically extract useful information from a large amount of data and then use this information to conduct scientific research and discovery. Artificial intelligence technology has great potential in scientific research and discovery.

6 Conclusion

The SparkRA system, built on the SciLit-LLM, provides a comprehensive solution for academic tasks, including literature investigation, paper reading, and academic writing. Through extensive experiments, SparkRA demonstrated superior performance compared to existing models like ChatGPT, and even surpassed GPT-4 in specific tasks such as paper polishing, demonstrating its potential to enhance productivity for researchers and students with its precise and context-aware support for academic activities.

Acknowledgements

We thank anonymous reviewers for their helpful comments. Thanks to Shichuan Sun, Qingye Meng, Qirui Song, Hao Zhang, Tao Song, Bowen Fang and Chi Yu for their support for SparkRA system and this paper.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Carl Malamud, Roger Magoulas, Kyle Chard, and Ian Foster. 2022. Scholarbert: Bigger is not always better. *arXiv preprint arXiv:2205.11342*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiaozhong Lyu, Stefan Grafberger, Samantha Biegel, Shaopeng Wei, Meng Cao, Sebastian Schelter, and Ce Zhang. 2023. Improving retrieval-augmented large language models via data importance learning. *arXiv preprint arXiv:2307.03027*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, pages 1–6.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nigam H Shah, David Entwistle, and Michael A Pfeffer. 2023. Creation and adoption of large language models in medicine. *Jama*, 330(9):866–869.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Biomegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. 2023. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, pages 1–6.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023a. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.