Research Proposal on Reliable Biomedical Large Language Models

Boyang Gu
Imperial College London
boyang.gu19@imperial.ac.uk

2023

1 Introduction

Over the past few years, a wide range of general large language models (LLMs), such as PaLM [11], LLaMA [72, 73], GPT-series [56, 6, 49], and ChatGLM [14, 89] have emerged and advanced the state-of-the-art in various natural language processing (NLP) tasks, including text generation, text summarization, and question answering. Inspired by the great success of general LLMs, the development and application of medical LLMs have gained growing research interests as they aim to assist medical professionals and improve patient care [3, 70, 53]. To this end, several endeavors have been made to adapt general LLMs to the healthcare domain, leading to the emergence of medical LLMs [64, 65, 23, 75, 71, 38, 82]. For example, based on PaLM [11], MedPaLM [64] and MedPaLM-2 [65] have achieved a competitive score of 86.5 compared to human experts (87.0 [81]) in the United States Medical Licensing Examination (USMLE) [31]; based on publicly available LLMs, e.g., LLaMA [72, 73], several medical LLMs, including ChatDoctor [38], MedAlpaca [23], PMC-LLaMA [81], BenTsao [75], and Clinical Camel [71], have been introduced.

Although existing medical LLMs have achieved promising results, there are some key issues in their development and application that need to be addressed. Firstly, many of these models primarily focus on biomedical natural language processing (NLP) tasks, such as dialogue and question answering, often overlooking their practical utility in clinical practice [70]. Recent research has begun to explore the potential of medical LLMs in various clinical scenarios, including electronic health records (EHRs) [83, 80], discharge summary generation [53], health education [59], and care planning [16]. Moreover, most existing medical LLMs evaluate their performances mainly on medical question answering and dialogue generation tasks, neglecting other biomedical tasks, such as text summarization, relation extraction, and information retrieval. These challenges pose a gap in the current research and application of LLMs in healthcare.

2 Related work

In this section, we will discuss some widely studied downstream tasks in the biomedical NLP domain. There are many more downstream tasks in this domain, such as Natural Language Inference, Text Simplification, Semantic Textual Similarity, and Relation Extraction, but those tasks we won't introduce here have similar natures to the tasks we will discuss below.

2.1 Entity Extraction

To perform the NER task, for each input token, the LLMs output a dense representation, which not only embeds the tokens but also includes its relation with other tokens in the text. Therefore, with the dense representations of input tokens extracted as vectors, additional layers are applied to the last Transformer layer to fit the downstream entity extraction task. The widely-used additional layers are softmax, BiLSTM [27], CRF [67, 88], and their combinations [55, 67, 17, 88, 7, 33]. Gu et al. [20] proposed a distillation method that uses few-shot GPT-3.5 to extract the correct entities and create the training set for the student model. This method shows that GPT models can be used to label the dataset first to achieve unsupervised learning. Wang et al. [77] defined a new prompting method: self-questioning prompting (SQP). The idea of this prompting method is to let the GPT model ask questions about the given text and

then asks the GPT to answer those question to extract useful information for specific tasks. Wang et al. [77] further evaluated the performances of BARD, GPT-3.5, and GPT-4 with SQP and achieved some SOTA results.

2.2 Text Classification

It is a classic classification problem: assigning predefined labels to a text, and it is common for a text to have multiple labels to describe it. Since it is a text-level task, the [CLS] vector is augmented when using transformer-based models. [CLS] does not correspond to any existing token in the sentence. Instead, it is manually augmented for the model to learn the information of all tokens, hence further being used to classify the relation. Another approach to distilling the overall information is to use the weighted sum of final attention layer outputs. There are works showing that adding a custom attention layer after the original model (BERT) improves the performance [69, 79]. There are also some works combining graph-based models with GPT models. ChatGraph [61] used ChatGPT to extract text information and apply it to a graph-based model that outperforms GPT models. CohortGPT [21] used Chain-of-Thought (CoT) prompting and knowledge graph to outperform few-shot ChatGPT and GPT-4. McCreery et al. [44] double-fine-tuned the BERT model in the sense that they first fine-tuned the model on a general dataset and then fine-tuned it on a medical-specific dataset.

2.3 Information Retrieval

Information retrieval (IR) plays an important role in the clinical area. It is the process of retrieving relevant knowledge or information related to the query from a number of unstructured data. Jin et al. [32] developed a BERT-based model BioCPT that encodes the query and articles for the ranking task. They split the method into training, inference, and evaluation steps. In the training step, they introduced query-to-document loss and document-to-query loss to train the encoders for the query and articles. In the inference step, they concatenate the encoding of the query and its best-fit article d_1^q together with $k-1$ non-relevant articles $(d_2^q, d_3^q, \cdots, d_k^q)$ found by maximum inner product search (MIPS) to further train the model to rank d_1^q at the top. Sun et al. [68] applied zero-shot ChatGPT to rank the most relevant documents without abnormal prompting. They also further used GPT-4 to re-rank the top 30 documents retrieved by ChatGPT. Abonizio et al. [1] introduced two LLM-based data augmentation methods, namely InPars and Promptagator, for IR. For InPars, they used GPT-3 and GPT-J to generate a new query for a randomly selected document. They used few-shot prompting which provides the model with good query examples or bad query examples. For Promptagator, the major difference is that a more dataset-specific prompting is applied. Ateia and Kruschwitz [4] proposed a query expansion technique that expands the current query into a more comprehensive query, which consistently improves the performance of any successive tasks. It is done purely by the GPT model with regular instructional prompting. Similarly, Wang et al. [76] used ChatGPT to generate more refined Boolean queries for systematic reviews. They showed that ChatGPT is able to generate or refine queries with higher precision.

2.4 Text Summarization

For extractive summarization, the key point is to define some scoring system that scores all sentences and hence finds the most important ones. Moradi et al. [47] used a clustering-based method to summarize medical texts. They vectorize the tokens of the text by BERT and cluster the sentence vector into k clusters, then they define an informativeness score that chooses one sentence from each cluster to form a summary. Moradi et al. [46] used the graph-based model to summarize. They treated sentences as nodes and relations as edges. The relations are measured by calculating the cosine similarity of vectors representing the sentences. They then used different graph ranking algorithms to choose important sentences as a summary. Du et al. [13] used purely the transformer-based model as the scoring system. They tokenized the whole text with [CLS] and [SEP] augmented. They further augmented the corresponding sentence and token positions to each token and fed the whole vector into the model. A sigmoid layer is added to the model so the output is between zero and one, and the output is considered the score for each sentence. Since the transformer automatically extracts relations of one sentence to others, Du et al. don't need to design a score manually. Chen et al. [9] also relied only on the model to score the sentences. The difference is that they used AlphaBERT and the training is split into pre-training and fine-tuning. McInerney et al. [45] combined the summarization model with a query to output more specific summaries. Pang et al. [52] proposed a principled inference framework with top-down and bottom-up inference techniques to improve summarization models. There are also works about summarizing more than one text at a time. Those works mainly relied on graph-based or transformer-based models to extract relations between different texts [8]. Zhang et al. [90] applied one-shot GPT-3.5, using dialogues and summaries from the same category as

prompts to generate abstractive summarization. More studies are needed to qualitatively analyze the performance of GPT models on biomedical text summarization [60].

2.5 Question Answering

For multiple-choice-orientated QA, the question is considered a classification problem. Similar to TC (see Sec 2.2), a common approach is to apply an MLP layer on the final [CLS] vector for classification. For free-text-orientated QA, the common approach is to use IR (see Sec 2.3). Models need to extract the most relevant tokens from the related documents to generate answers. As a result, some QA models are multi-tasked [2]. Since biomedical QA datasets are relatively small in size compared to general datasets, using pre-trained models from general datasets and then finetuning them on the biomedical data improves the performance [66, 87]. Masking strategy is also well-used in QA. For a document, the model will randomly mask some tokens and try to predict the masked unknown tokens. The trained models then can be used to predict relevant tokens about the query. Pergola et al. [54] proposed a biomedical-specific masking method. Instead of masking tokens randomly, the model will identify biomedical-related tokens and mask them to focus more on in-domain learning. Like what we discussed in the IR subsection, hallucination also threatens the quality of outputs in QA. One way is to use biomedical search systems like Almanac [26]. Another approach is to use extra datasets as an augmentation for the QA task. There are many chatbots based on LLM QA, such as Clinical Camel [71], DoctorGLM [82], ChatDoctor [38], HuaTuo [75], HuaTuoGPT [91], and MedAlpaca [23]. Except for a few [36], most chatbots are a black box the the consumers, so further studies of those chatbots are required.

3 Research Questions and Methodology

LLMs have emerged as promising tools for applications in the medical and healthcare fields. However, their integration is not without challenges. This section covers the challenges of using LLMs in the medical field.

3.1 Hallucination

Hallucination of LLMs refers to the phenomenon where the generated output contains inaccurate or nonfactual information. It can be categorized into intrinsic and extrinsic hallucinations [30, 74, 57]. Intrinsic hallucination refers to generating outputs logically contradicting factual information - such as LLMs generating wrong calculations of mathematical formulas [30]. Extrinsic hallucination happens when the output generated cannot be verified - typical examples include LLMs 'faking' citations that do not exist or 'dodging' the question. When integrating LLMs into the medical domain, fluent but nonfactual LLM hallucinations can lead to the dissemination of incorrect medical information, which can cause misdiagnoses, inappropriate treatments, and harmful patient education. Given the criticality of the medical domain, it is vital to ensure the accuracy of LLM outputs.

Potential Solutions Current solutions to mitigate LLM hallucination can be categorized into training-time correction, generation-time correction, and retrieval-augmented correction. The first solution, training-time correction, aims to mitigate hallucination by adjusting model weights and thus reducing the probability of generating hallucinated outputs. Examples of training-time correction include factually consistent reinforcement learning [58] and contrastive learning[10]. Another solution to reduce hallucination is to add a 'reasoning' process to the LLM inference to ensure reliability. Methods include drawing multiple samples [43] or using a confidence score to identify hallucination before the final generation. The third approach is the retrieval-augmented correction method, which utilizes external resources to help mitigate hallucination. For example, using factual documents as prompts [63] or chain-of-retrieval prompting technique [12].

3.2 Lack of Evaluation Benchmarks and Metrics

With the emerging ability of general-purpose LLMs, current benchmarks and metrics fail to evaluate LLM's overall capabilities, especially in the medical domain. Current benchmarks such as MedQA (USMLE) [31] and MedMCQA [51] offer extensive coverage on question-answering tasks but fail to evaluate important LLM-specific metrics such as trustworthiness, faithfulness, helpfulness and explainability. The need for more domain and LLM-specific benchmarks and metrics is emerging.

Potential Solutions New benchmarks should include capabilities like sourcing from authoritative medical references, adapting to the evolving landscape of medical knowledge, and clearly communicating uncertainties [64]. Additionally, considering the sensitive nature of healthcare, these benchmarks should also assess factors such as fairness, ethics, and equity, which, though crucial, pose quantification challenges [64]. The aim is to create benchmarks that more effectively mirror actual clinical scenarios, thus providing a more accurate measure of LLMs' suitability for medical advisory roles. Current LLM research in medicine has largely focused on general medicine, likely due to the greater availability of data in this area [64, 65, 24]. However, this focus has resulted in the underrepresentation of LLM applications in specialized fields like 'rehabilitation therapy' and 'sports medicine'. Despite initiatives to incorporate physical activity (PA) into healthcare systems, implementation remains challenging, particularly in developing countries with limited PA education among healthcare providers [42]. We could disseminate accurate PA knowledge and aid in the creation of personalized PA programs. Such applications could significantly enhance PA levels, improving global health outcomes, especially in resource-constrained environments.

3.3 Domain Data Limitations

Current datasets in the medical domain remain relatively small compared to datasets used to train general-purpose LLMs. The medical knowledge domain is vast; existing datasets are limited and do not cover the entire space [64]. This results in LLMs exhibiting extraordinary performance on open benchmarks with extensive data coverage yet falling short on real-life tasks such as differential diagnosis and personalized treatment planning [65].

Potential Solutions Although the volume of medical and health data is large, most require extensive ethical, legal, and privacy procedures to be accessed. In addition, these data are often unlabeled, and solutions to leverage these data, such as human labelling and unsupervised learning [39], face challenges due to the lack of human expert resources and small margins of error. Current state-of-the-art approaches [64], [65], [38], prefer to fine-tune on smaller open-sourced datasets to improve models' domain-specific performances. Another solution is to generate high-quality synthetic datasets using LLMs to broaden the knowledge coverage [22]. However, several works have discovered that training on generated datasets causes models to forget [62]. Therefore, future research is needed to validate the effectiveness of using synthetic data for LLMs in the medical field. Also, current datasets fail to capture the differences in race and language. We could work on such new datasets.

3.4 New Knowledge Adaptation

LLMs are trained on extensive data to learn knowledge. Once the LLM is trained, injecting new knowledge through re-training is expensive and inefficient. Two problems occur when a knowledge update is required (for example, a new adverse effect of a medication, or a novel disease): The first problem is how to make LLMs 'forget' the old knowledge - it is almost impossible to remove all 'old knowledge' from the training data, and the discrepancy between new and old knowledge can cause unintended association and bias [28]. The second problem is the timely addition of knowledge - how do we ensure the model is updated in real-time? These problems pose significant barriers to using LLMs in medical fields, where accurate and timely update of up-to-date medical knowledge is crucial in real-world implementations.

Potential Solutions We can categorize current solutions into model editing and retrieval-augmented generation. Model editing [85] refers to altering the model's knowledge by modifying the model's parameters. These methods do not generalize, and their effectiveness varies across different model architectures. The second solution is retrieval-augmented generation, which provides external knowledge sources as prompts during model inference. For example, Lewis et al. [35] enabled model knowledge updates by updating the model's external knowledge memory.

3.5 Behaviour Alignment

Behaviour alignment refers to the process of ensuring that the LLM's behaviours align with the objectives of its task. While efforts are spent aligning LLMs with human behaviour, the behaviour discrepancy between general humans and medical professionals remains challenging for adopting LLMs in the medical domain. For example, ChatGPT's answers for medical consultations are not as concise and professional as the human expert's answers [50]. In addition, misalignment may introduce unnecessary harm and ethical concerns [25] that lead to undesirable consequences in the medical domain.

Potential Solutions Current solutions include instruction fine-tuning, reinforcement learning from human feedback (RLHF) [5, 50], and prompt tuning [34, 41]. Instruction fine-tuning [78] refers to improving the performance of LLMs on specific tasks based on explicit instructions. For example, Ouyang et al. [50] used this technique to help LLMs generate less toxic and more suitable outputs. RLHF is a reinforcement learning technique that uses human feedback to evaluate and align the outputs of LLMs. It has proven effective in multiple tasks, such as helping LLMs become helpful chatbots [19] and decision-making agents [48]. Prompt tuning can also align LLMs to the expected output format. For example, Liu et al. [40] uses a prompting strategy, chain of hindsight, to enable the model to detect and correct its errors, which aligns the generated output with human expectations.

3.6 Multimodal LLM Integrated with Time-Series, Visual, and Audio Data

Multimodal LLMs (MLLMs), or Large Multimodal Models (LMMs), are LLM-based models designed to perform multimodal tasks [86]. While LLMs primarily address NLP tasks, MLLMs support a broader range of tasks, such as comprehending the underlying meaning of a meme [84] and generating website codes from images [92]. This versatility suggests promising applications of MLLMs in medicine. For example, recent works have introduced an MLLM-based framework that integrates vision, audio, and language inputs for automated diagnosis in dentistry [29]. However, there are only very few medical LLMs that can process time series data, such as electrocardiograms (ECGs) [37] and sphygmomanometers (PPGs) [15]. These time series data are important for medical diagnosis and monitoring. We could inherit the knowledge of audio models into time series data. Moreover, like LLMs, MLLMs are associated with data privacy and quality challenges. The multimodal nature of MLLM also introduces unique issues, including limited perception capabilities [29][92], fragile reasoning chains [18], sub-optimal instruction-following ability [18], and object hallucination [92]. Therefore, more research is needed to address these issues, ensuring a safe and effective application of MLLM in medicine.

4 Conclusion

For potential research in the biomedical NLP domain, we introduced some interesting areas that still need further research, including Hallucination, Evaluation of Benchmarks and Metrics, Domain Data Limitations, New Knowledge Adaptation, Behaviour Alignment, and Multimodal LLM. We plan to focus first on the introduction of new benchmarks, and then use the new benchmarks to deal with the hallucination issue. Finally, we will turn our model multimodal with the time-series data.

References

[1] Hugo Abonizio, Luiz Bonifacio, Vitor Jeronymo, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. Inpars toolkit: A unified and reproducible synthetic data generation pipeline for neural information retrieval. *arXiv preprint arXiv:2307.04601*, 2023.

[2] Arda Akdemir and Tetsuo Shibuya. Transfer learning for biomedical question answering. In *CLEF (Working Notes)*, 2020.

[3] Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023.

[4] Samy Ateia and Udo Kruschwitz. Is chatgpt a biomedical expert?–exploring the zero-shot performance of current gpt models in biomedical tasks. *arXiv preprint arXiv:2306.16108*, 2023.

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Miao Chen, Fang Du, Ganhui Lan, and Victor S Lobanov. Using pre-trained transformer deep learning models to identify named entities and syntactic relations for clinical protocol analysis. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, pages 1–8, 2020.

[8] Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. Capturing relations between scientific papers: An abstractive model for related work section generation. Association for Computational Linguistics, 2021.

[9] Yen-Pin Chen, Yi-Ying Chen, Jr-Jiun Lin, Chien-Hua Huang, Feipei Lai, et al. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (alphabert): development and performance evaluation. *JMIR medical informatics*, 8(4):e17787, 2020.

[10] I-Chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, Graham Neubig, et al. Improving factuality of abstractive summarization via contrastive reward learning. *arXiv preprint arXiv:2307.04507*, 2023.

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[12] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

[13] Yongping Du, Qingxiao Li, Lulin Wang, and Yanqing He. Biomedical-domain pre-trained language model for extractive summarization. *Knowledge-Based Systems*, 199:105964, 2020.

[14] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[15] Zachary Englhardt, Richard Li, Dilini Nissanka, Zhihan Zhang, Girish Narayanswamy, Joseph Breda, Xin Liu, Shwetak Patel, and Vikram Iyer. Exploring and characterizing large language models for embedded system development and debugging. *arXiv preprint arXiv:2307.03817*, 2023.

[16] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv preprint arXiv:2308.14089*, 2023.

[17] Kathleen C Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. Extracting umls concepts from medical text using general and domain-specific deep learning models. *arXiv preprint arXiv:1910.01274*, 2019.

[18] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[19] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

[20] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*, 2023.

[21] Zihan Guan, Zihao Wu, Zhengliang Liu, Dufan Wu, Hui Ren, Quanzheng Li, Xiang Li, and Ninghao Liu. Cohortgpt: An enhanced gpt for participant recruitment in clinical study. *arXiv preprint arXiv:2307.11346*, 2023.

[22] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

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

[24] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2023.

[25] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

[26] William Hiesinger, Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex Dalal, Jennifer Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, et al. Almanac: Retrieval-augmented language models for clinical medicine. 2023.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[28] Jason Hoelscher-Obermaier, Julia Persson, Esben Kran, Ioannis Konstas, and Fazl Barez. Detecting edit failures in large language models: An improved specificity benchmark. *arXiv preprint arXiv:2305.17553*, 2023.

[29] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023.

[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[31] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[32] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, John Wilbur, and Zhiyong Lu. Biocpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *arXiv preprint arXiv:2307.00589*, 2023.

[33] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. Umls-based data augmentation for natural language processing of clinical research literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, 2021.

[34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021.

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

[36] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023.

[37] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.

[38] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023.

[39] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. Auto-encoding knowledge graph for unsupervised medical report generation. In *Advances in Neural Information Processing Systems*, 2021.

[40] Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. *arXiv preprint arXiv:2302.02676*, 3, 2023.

[41] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[42] Felipe Lobelo, Mark Stoutenberg, and Adrian Hutber. The exercise is medicine global health initiative: a 2014 update. *British journal of sports medicine*, 48(22):1627–1633, 2014.

[43] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

[44] Clara McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Domain-relevant embeddings for medical question similarity. *arXiv preprint arXiv:1910.04192*, 2019.

[45] Denis Jered McInerney, Borna Dabiri, Anne-Sophie Touret, Geoffrey Young, Jan-Willem Meent, and Byron C Wallace. Query-focused ehr summarization to aid imaging diagnosis. In *Machine Learning for Healthcare Conference*, pages 632–659. PMLR, 2020.

[46] Milad Moradi, Maedeh Dashti, and Matthias Samwald. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. *Journal of Biomedical Informatics*, 107:103452, 2020.

[47] Milad Moradi, Georg Dorffner, and Matthias Samwald. Deep contextualized embeddings for quantifying the informative content in biomedical text summarization. *Computer methods and programs in biomedicine*, 184:105117, 2020.

[48] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[49] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[51] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[52] Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, and Caiming Xiong. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*, 2022.

[53] Sajan B Patel and Kyle Lam. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108, 2023.

[54] Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. Boosting low-resource biomedical qa via entity-aware masking strategies. *arXiv preprint arXiv:2102.08366*, 2021.

[55] Beatrice Portelli, Edoardo Lenzi, Emmanuele Chersoni, Giuseppe Serra, and Enrico Santus. Bert prescriptions to avoid unwanted headaches: a comparison of transformer architectures for adverse drug event detection. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: main volume*, pages 1740–1747, 2021.

[56] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[57] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[58] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.

[59] Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. The role of large language models in medical education: applications and implications, 2023.

[60] Chantal Shaib, Millicent L Li, Sebastian Joseph, Iain J Marshall, Junyi Jessy Li, and Byron C Wallace. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*, 2023.

[61] Yucheng Shi, Hehuan Ma, Wenliang Zhong, Gengchen Mai, Xiang Li, Tianming Liu, and Junzhou Huang. Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. *arXiv preprint arXiv:2305.03513*, 2023.

[62] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493*, 2023.

[63] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[64] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[65] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[66] Sarvesh Soni and Kirk Roberts. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5532–5538, 2020.

[67] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Biomedical named entity recognition using bert in the machine reading comprehension framework. *Journal of Biomedical Informatics*, 118:103799, 2021.

[68] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*, 2023.

[69] Matthew Tang, Priyanka Gandhi, Md Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. Progress notes classification and keyword extraction using attention-based deep learning models with bert. *arXiv preprint arXiv:1910.05786*, 2019.

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

[71] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.

[72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[73] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

[75] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[76] Shuai Wang, Harrisen Scells, Bevan Koopman, and Guido Zuccon. Can chatgpt write a good boolean query for systematic review literature search? *arXiv preprint arXiv:2302.03495*, 2023.

[77] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

[78] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[79] David A Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, et al. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Medical Imaging with Deep Learning*, pages 811–826. PMLR, 2020.

[80] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135, 2023.

[81] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

[82] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023.

[83] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B Costa, Mona G Flores, et al. A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194, 2022.

[84] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

[85] Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. *arXiv preprint arXiv:2305.13172*, 2023.

[86] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[87] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. Pre-trained language model for biomedical question answering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 727–740. Springer, 2019.

[88] Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. Biobert based named entity recognition in electronic medical record. In *2019 10th international conference on information technology in medicine and education (ITME)*, pages 49–52. IEEE, 2019.

[89] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2022.

[90] Boya Zhang, Rahul Mishra, and Douglas Teodoro. Ds4dh at mediqa-chat 2023: Leveraging svm and gpt-3 prompt engineering for medical dialogue classification and summarization. *medRxiv*, pages 2023–06, 2023.

[91] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

[92] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.