Harmonizing Offline Reinforcement Learning with Language Models Analysis of Human Responses

*Dr. N Krishnaiah[1], Vastrala Vaishnavi[2], Ch. Thanmai[3], D. Pranathi[4], G. Sunaina[5], K. Gayathri Bhavya[6], M. Aditi[7]
[1]Professor, St. Martin's Engineering College, Secunderabad, Telangana-500100
[2]UG Student, B V C Engineering College, Odalarevu, Amalapuram, A. P
[3,4,5,6,7]UG Students, St. Martin's Engineering College, Secunderabad, Telangana-500100
*Corresponding Author
Email: drnkrishnaiahit@smec.ac.in

ABSTRACT

Learning from human preferences is vital for language models (LMs) to successfully cater to human wants and social values. Using human feedback to motivate compliance with instructions, previous studies have made significant improvement. Proximal Policy Optimization (PPO) and other online RL methods are heavily relied upon in these methods, although they have shown to be unstable and difficult to tune for language models. The complexity involved in implementing a distributed system for PPO also reduces the effectiveness of distributed training on a broad scale. To align LMs without engaging with RL settings, we offer an offline method called reinforcement learning from human feedback (RLHF). To better align language models with user preferences, we investigate the use of maximum likelihood estimation (MLE) with filtering, reward-weighted regression (RWR), and Decision Transform (DT). Our methods use a loss function analogous to supervised fine tuning to guarantee more consistent model training than PPO while making do with a minimalist machine learning system (MLSys) and significantly less computational resources (by about 12.3%). The experimental data show that DT alignment performs better than PPO and other Offline RLHF techniques.

Keywords: Human Preferences, Language Models, RLHF, MLSys.

1. Introduction

Since its release by the US company OpenAI in November 2022, a chatbot, ChatGPT, has stunned the world with its outstanding performance in conversation with humans [1]. Bill Gates acclaimed that the new generation of conversational agents 'will change the way people work, learn, travel, get health care, and communicate with each other', bringing in significant productivity improvement and reducing some of the world's worst inequities, particularly for health [1]. The White House media release acclaimed that 'From cancer prevention to mitigating climate change to so much in between, AI—if properly managed— can contribute enormously to the prosperity, equality, and security of all' [2]. ChatGPT is a representative example of generative artificial intelligence (AI) technology. Generative AI refers to a subset of AI technologies that learn to predict the next word or sequence of words giving the preceding context. They can generate new content, such as text, images, music, speech, video, or code. Their huge success has attracted unprecedented speed of adoption, excitement, and controversy. Generative AI models use advanced deep learning and transfer learning algorithms and machine learning techniques to learn patterns and relationships from the existing data and generate new content similar in style, tone, or structure. Deep learning is a subset of machine learning that uses neural networks with multiple layers of processing nodes to analyze various factors of data for complex pattern recognition and prediction. Transfer

learning is a machine learning technique that adapts a pre-trained model to a new but related task, leveraging knowledge from the initial task to improve new task performance.

Generative AI models are a subset of large language models (LLMs), e.g., generative pre-trained transformer (GPT). For example, GPT-3 is trained on 175 billion parameters, while GPT-4 is trained on one trillion parameters. An intermediary version, GPT-3.5, is specifically trained to predict the next word in a sequence using a large dataset of Internet text. It is the model that underpins the current version of ChatGPT [3]. After being pretrained on huge amounts of data to learn intricate patterns and relationships, these LLMs have developed capabilities to imitate human language processing [4]. Upon receiving a query or request in a prompt, ChatGPT can generate relevant and meaningful responses and answer questions drawing from its learned language patterns and representations [5]. These LLMs are often referred to as the "foundation model" or "base model" for generative AI, as they are the starting point for the development of more advanced and complex models.

Distinct from traditional AI systems, which are typically rule-based or rely on predefined datasets, generative AI models possess the unique ability to create new content that is original and not explicitly programmed. This can result in outputs that are similar in style, tone, or structure to the prompt instruction. Therefore, if designed thoughtfully and developed responsibly, generative AI has the potential to amplify human capabilities in various domains of information management. These may include support for decision-making, knowledge retrieval, question answering, language translation, and automatic report or computer code generation [4].

It is not surprising that a significant area for generative AI and LLM to revolutionize is healthcare and medicine, a human domain in which language is key for effective interactions for and between clinicians and patients [6]. It is also an information-rich field where every assessment, diagnosis, treatment, care plan, and outcome evaluation must be documented in specific terms or natural language in electronic health records (EHR). Once the LLM is exposed to the relevant EHR data set in a specific healthcare field, the model will learn the relationships between the terms and extend its model to represent the knowledge in this field. With the further advancement of generative AI technologies, including video and audio technologies, the dream is not far away for healthcare providers to audit instead of simply typing data into EHR. Clinicians may orally request computers to write prescriptions or order lab tests and ask the generative AI models integrated with EHR systems to automatically retrieve data, generate shift hand-over reports and discharge summaries, and support diagnostic and prescription decision-making. Therefore, generative AI can be 'a powerful tool in the medical field' [7].

Generative AI and LLMs have also sparked intense debates and discussions regarding their potential benefits, future perspectives, and critical limitations for healthcare and medicine. In Sallam's seminal systematic review of 60 selected papers that assess the utility of ChatGPT in healthcare education, research, and practice, 85% (51/60) of papers cited benefits/applications, while an overwhelming 97% (58/60) raised concerns or possible risks associated with ChatGPT use [8]. These findings suggest that with proper handling of ethical concerns, transparency, and legal matters, these technologies could not only expedite research and innovation but also foster equity in healthcare.

2. Methods

The scoping literature review addressed questions a health or medical scholar without adequate machine learning background yet keen on the generative AI and LLM field might ask. To identify the pertinent literature, our primary search was structured into two steps using Boolean logic with the keywords listed in Table 1.

Table 1. The keywords used for the literature search.

Aim for Literature Search	Search Keywords
Step 1: Credible information about generative AI and LLM.	("generative artificial intelligence" or "generative AI") or/and ("large language model" or "LLM")
Step 2: Application of generative AI and LLM in healthcare or medicine.	("generative artificial intelligence" or "generative AI") or/and ("large language model" or "LLM") and (("healthcare" or "health care") or "medicine")
Step 3: AI limitations, regulation, and ethics.	("regulation" and "generative artificial intelligence" or "generative AI") or/and ("large language model" or "LLM") and (("healthcare" or "health care") or "medicine") and ("limit*" or "align*" or "ethics")

Step 1 aimed to grasp the scope of generative AI and LLM, initiating with Google Scholar because the significant articles that were pertinent to our inquiries, e.g., the development of Open AI's GPT models and Google's PaLM models, were published in arXiv, a free repository for academic pre-prints. One query often led to subsequent queries, guided by the referenced literature; therefore, we further assessed these references. Once wellinformed about generative AI and LLM, we proceeded to Step 2, exploring the literature detailing their applications in healthcare or medicine in PubMed. The search period was from 1 March to 15 July 2023.

Article titles and abstracts were scanned to assess their relevance to our research questions. Noting the focus on GPT's limitations and performance (see Table 2), we extended keywords in Step 3, addressing ethical and regulatory considerations for generative AI. Drawing from official websites, we compiled regulatory perspectives from the US and UK governments on generative AI. This iterative approach, utilizing the keywords in Table 1, resulted in two distinct concept clusters relevant to our enquiry from the 88 analyzed article titles and abstracts (see Figure 1).

Informed by the sharpened research questions and insights, we crafted the outline of our article, delineating fundamental research issues, concepts, and their causal interconnections. The iterative practices of evidence evaluation, question adjustment, and conceptual mapping persisted until we achieved satisfaction with the content. After this, we further polished and finalized the manuscript.

Table 2. Comparison of key concepts included in the network of terms derived from the initially scanned 88 academic articles versus the 55 articles employed in this scoping review. Analysis was conducted on article titles and abstracts.

Terms from the Primary 88 Articles Scanned			Terms from the 55 Referenced Articles		
Term	[1] Occurrences	[2] Relevance	Term	Occurrences	Relevance
gpt	20	2.46	response	55	2.79
limitation	14	1.47	physician	17	2.73
performance	24	1.41	question	35	1.02
llm	27	1.13	patient message	11	1
response	58	0.85	chatgpt	44	0.86
patient message	11	0.81	research	19	0.59
physician	17	0.79	study	29	0.29
manuscript	10	0.79	large language model	18	0.26
chatbot	14	0.54	patient	11	0.24
research	30	0.48	model	18	0.22
question	40	0.28			

[1] Occurrences: the number of times a specific term appears in the data set. [2] Relevance: normalized frequency of co-occurrence of the terms with the other terms in the data set.

3. Results

We present our findings from seven aspects: technological approaches to generative AI applications, methods to train LLM, model evaluation, current applications of generative AI and LLM in healthcare and medicine, benefits, ethical and regulatory considerations, and future research and development directions.

3.1. Technological Approaches to the Application of Generative AI and LLMs

Generative AI and LLMs are powered by a suite of deep learning technologies. For example, ChatGPT is a series of deep learning models that utilize transformer architecture that resorts to self-attention mechanisms to process large human-generated text datasets (GPT-4 response, 23 August 2023). These AI technologies work in harmony to power ChatGPT, enabling it to handle a wide range of tasks, including natural language understanding, language generation, text completion, translation, summarization, and much more.

There are three key factors in choosing LLMs to solve practical problems: models, data, and downstream tasks [1], which also apply to solve healthcare and medicine problems.

3.1.1. Models

Based on model training strategies, architectures, and use cases, LLMs are classified into two types [1]: (1) encoder–decoder or encoder-only language models and (2) decoderonly models. The encoder-only models represented by BERT family models have started to phase out after the debut of ChatGPT. Encoder–decoder models, e.g., Meta's BART, remain promising as most of them are open-sourced, providing opportunities for the global software community to continuously explore and develop. Decoder-only models, represented by the GPT family models, Pathways Language Model (PaLM) introduced by Google [10], and LLaMA models from Meta, have and will continue to dominate the LLM space because they are the foundation models for generative AI technologies.

3.1.2. Data

The impact of data on the models' effectiveness starts from pre-training data and continues through to the training, test, and inference data [6]. The quality, quantity, and diversity of pre-training data significantly influence the performance of LLMs [1]. Therefore, pre-training base models on data from a specific healthcare or medical field to produce instruction fine-tuned models are the recommended development method for downstream machine learning tasks for these fields [13]. Of course, with abundant annotated data, both base LLM and instruction fine-tuned models can achieve satisfactory performance on a particular task and meet the important privacy constraint for healthcare and medical data [14].

3.1.3. Task

LLMs can be applied to four types of tasks: natural language understanding (NLU), natural language generation, knowledge-intensive tasks, and reasoning [1]. Traditional natural language understanding tasks include text classification, concept extraction or named entity recognition (NER), relationship extraction, dependency parsing, and entailment prediction. Many of these tasks are intermediate steps in large AI systems, such as NER for knowledge graph construction. Using the decoder LLMs may directly complete inference tasks and remove these intermediate tasks.

Natural language generation includes two major types of tasks: (1) converting input texts into new symbol sequences, such as text summarization and machine translation, and (2) "open-ended" generation, which aims to generate new text or symbols in response to the input prompt, e.g., question answering, crafting emails, composing news articles, and writing computer codes [1]. This capability is useful for many tasks in healthcare and medicine.

3.2. Methods to Train LLMs

3.2.1. Fine-Tuning LLMs

LLMs can be fine-tuned by various strategies, e.g., modifying the number of parameters [18], size of the training data set, or the amount of computing used for training [1]. Fine-tuning LLMs will scale up the pretrained LLMs and significantly improve their performance in reasoning beyond the power-law rule to unlock unprecedented, fantastic emergent abilities [6,19]. Emergent abilities refer to specific competencies that do not exist in smaller models but become salient as the model scales. These include but are not limited to nuanced concept understanding, sophisticated word manipulation, advanced logical reasoning, and complex

coding tasks [6]. For instance, when the PaLM model was scaled from 8 billion parameters to 540 billion parameters, it exhibited emergent abilities that essentially doubled its performance. The scaled Med-PaLM model achieved an accuracy of 67.2% in answering questions from the United States Medical Licensing Exam (USMLE) dataset.

3.2.2. Reinforcement Learning from Human Feedback (RLHF)

RLHF refers to methods that combine three interconnected model training processes: feedback collection, reward modeling, and policy optimization [22]. RLHF has been implemented as instruction prompts to train LLMs to achieve remarkable performance across many NLP tasks [6,16,18]. It not only improves model accuracy, factuality, consistency, and safety and mitigates harm and bias within medical question-answering tasks [6], but also bridges the gap between LLM-generated answers and human responses. Therefore, RLHF brings LLMs considerably closer to practical applications within real-world clinical settings.

3.2.3. Prompt Engineering

Prompt engineering refines prompts for generative AI to generate text or images, often through an iterative refinement process. To date, five instruction prompts have been reported: zero-shot, few-shot, chain-of-thought, self-consistency, and ensemble refinement learning.

Zero-shot learning enables the training of LLMs for specific NLP tasks through singleprompt instructions, eliminating the need for annotated data [23]; e.g., people enter instructions into 'prompt' to seek answers from ChatGPT. This approach avoids the issue of catastrophic forgetting often encountered in fine-tuned neural networks, as it does not require model parameter updates [24]. Recent studies, such as those by Zhong et al., affirm the efficacy of LLM's zero-shot learning in various traditional natural language understanding tasks [25].

3.3. Model Evaluation

Three challenges impede the application of LLMs in modeling real-world tasks [1]: (1) noisy/unstructured real-world input data that are often messy, e.g., containing typos, colloquialisms, and mixed languages; (2) ill-defined practical tasks that are difficult to classify into predefined NLP task categories; and (3) ambiguous instructions that may contain multiple implicit intents. These ambiguities cause difficulty in predictive modelling without follow-up probing questions. Despite performing better than the fine-tuned models in addressing the above three challenges, the effectiveness of foundation models in handling real-world input data is yet to be evaluated [1,6]; therefore, Bommasani et al. calls for a holistic evaluation of LLMs [31].

3.4. Current Applications of Generative AI and LLMs in Healthcare and Medicine

There is tremendous potential for LLMs to innovate information management, education, and communication in healthcare and medicine [7]. Li et al. proposed a taxonomy to classify ChatGPT's utility in healthcare and medicine based on two criteria: (1) the nature of medical tasks that LLMs address and (2) the targeted end users [34]. According to the first criterion, seven types of ChatGPT applications were outlined: triage, translation, medical research, clinical workflow, medical education, consultation, and multimodal. Conversely, the second criterion delineates seven categories of end users: patients/relatives, healthcare professionals/clinical centers, payers, researchers, students/teachers/exam agencies, and lawyers/regulators.

A use case of LLMs to support the medical task of triage [34] is assisting healthcare professionals in condensing a patient's hospital stay into succinct summaries based on their medical records, then generating discharge letters [35], benefiting from these models' strong ability to summarize data from heterogeneous sources [36]. A useful application of LLM to improve clinical workflow is to significantly reduce the documentation burden that has long plagued doctors and nurses, a problem that persisted even after the transition from paper to electronic health records [37]. Importantly, LLM can improve interpretability [1], a vital goal

in health data management. Therefore, they have the potential to lead to remarkable improvements in healthcare safety, quality, and efficiency.

3.5. The Benefits of Generative AI and LLMs for Healthcare and Medicine

The application of generative AI and LLMs in healthcare and medicine remains predominantly within the academic research stage [43]. There are various cases delineate preliminary efforts in the exploration of generative AI within these fields.

- Creating Synthetic Patient Health Records to Improve Downstream Clinical Text Mining
- Using Chatbot Underpinned by LLMs to Assist Health Communication
- Potential to Address Routine Patient Queries following Routine Surgery
- Improving Accuracy in Medical Image Analysis
- Potential to Provide Ongoing Clinical Decision Support throughout the Entire Clinical Workflow
- Fine-Tuning Local Large Language Models for Pathology Data Extraction and Classification

4. Conclusions

This article examines the transformative potential of generative AI and LLMs in healthcare and medicine. It delves into the foundational mechanisms, diverse applications, learned insights, and the ethical and legal considerations associated with these technologies, highlighting the unique role of RLHF in model development. A limitation of this scoping review is its non-exhaustive nature, as it does not conduct a comprehensive, systematic appraisal of all the extant literature during the specific time period. The inclusion of numerous papers from arXiv that have not undergone rigorous peer review could potentially reduce the research's rigor.

Unlike traditional rule-based AI, these contemporary technologies empower domain experts and necessitate collaborative co-design process involving both clinicians and consumers. Global efforts are centered on exploring numerous opportunities and challenges in areas such as ethics, transparency, legal implications, safety, and bias mitigation. The promise for improving healthcare quality, safety, and efficiency is significant. Healthcare organizations should actively engage with these technologies while upholding ethical standards.

References

1. Yang, J.; Jin, H.; Tang, R.; Han, X.; Feng, Q.; Jiang, H.; Yin, B.; Hu, X. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. arXiv 2023, arXiv:2304.13712.
2. The White House. Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI; The White House: Washington, DC, USA, 2023.
3. OpenAI. Aligning Language Models to Follow Instructions. 2022. Available online: https://openai.com/research/instructionfollowing (accessed on 30 June 2023).
4. Zhao, Z.; Wallace, E.; Feng, S.; Klein, D.; Singh, S. Calibrate before use: Improving few-shot performance of language models. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; PMLR 139, pp. 12697–12706.
5. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios. J. Med. Syst. 2023, 47, 33. [CrossRef] [PubMed]
6. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S. Large language models encode clinical knowledge. Nature 2023, 620, 172–180. [CrossRef] [PubMed]
7. Harrer, S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. eBioMedicine 2023, 90, 104512. [CrossRef] [PubMed]
8. Sallam, M. ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. Healthcare 2023, 11, 887. [CrossRef] [PubMed]
9. van Eck, N.J.; Waltman, L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics 2017, 111, 1053–1070. [CrossRef] [PubMed]
10. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S. Palm: Scaling language modeling with pathways. arXiv 2022, arXiv:2204.02311.

11. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E. On the opportunities and risks of foundation models. arXiv 2021, arXiv:2108.07258.
12. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S. Scaling instructionfinetuned language models. arXiv 2022, arXiv:2210.11416.
13. Wang, B.; Xie, Q.; Pei, J.; Chen, Z.; Tiwari, P.; Li, Z.; Fu, J. Pre-trained language models in biomedical domain: A systematic survey. ACM Comput. Surv. 2021, 56, 1–52. [CrossRef]
14. Tang, R.; Han, X.; Jiang, X.; Hu, X. Does synthetic data generation of llms help clinical text mining? arXiv 2023, arXiv:2303.04360.
15. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digit. Health 2023, 2, e0000198. [CrossRef] [PubMed]
16. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D. Towards expert-level medical question answering with large language models. arXiv 2023, arXiv:2305.09617.
17. Williams, T.; Szekendi, M.; Pavkovic, S.; Clevenger, W.; Cerese, J. The reliability of AHRQ Common Format Harm Scales in rating patient safety events. J. Patient Saf. 2015, 11, 52–59. [CrossRef] [PubMed]
18. Umapathi, L.K.; Pal, A.; Sankarasubbu, M. Med-HALT: Medical domain hallucination test for large language models. arXiv 2023, arXiv:2307.15343.
19. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. arXiv 2020, arXiv:2001.08361.
20. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med. Educ. 2023, 9, e45312. [CrossRef] [PubMed]
21. Ge, Y.; Hua, W.; Ji, J.; Tan, J.; Xu, S.; Zhang, Y. OpenAGI: When LLM meets domain experts. arXiv 2023, arXiv:2304.04370.
22. Casper, S.; Davies, X.; Shi, C.; Gilbert, T.K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P. Open problems and fundamental limitations of reinforcement learning from human feedback. arXiv 2023, arXiv:2307.15217.
23. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv. 2023, 55, 1–35. [CrossRef]
24. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; GrabskaBarwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. USA 2017, 114, 3521–3526. [CrossRef] [PubMed]
25. Zhong, Q.; Ding, L.; Liu, J.; Du, B.; Tao, D. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv 2023, arXiv:2302.10198.
26. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 2020, 33, 1877–1901.
27. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. 2022, 35, 24824–24837.
28. Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R. Training verifiers to solve math word problems. arXiv 2021, arXiv:2110.14168.
29. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. arXiv 2022, arXiv:2203.11171.
30. Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. Self-refine: Iterative Refinement with Self-Feedback. arXiv 2023, arXiv:2303.17651.
31. Bommasani, R.; Liang, P.; Lee, T. Language Models are Changing AI: The Need for Holistic Evaluation. Available online: https://crfm.stanford.edu/2022/11/17/helm.html (accessed on 30 June 2023).
32. Siru, L.; Allison, B.M.; Aileen, P.W.; Babatunde, C.; Julian, Z.G.; Sean, S.H.; Josh, F.P.; Bryan, S.; Adam, W. Leveraging large language models for generating responses to patient messages. medRxiv 2023. [CrossRef]
33. Chowdhury, M.; Lim, E.; Higham, A.; McKinnon, R.; Ventoura, N.; He, Y.; De Pennington, N. Can Large Language Models Safely Address Patient Questions Following Cataract Surgery; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 131–137.
34. Li, J.; Dada, A.; Kleesiek, J.; Egger, J. ChatGPT in healthcare: A taxonomy and systematic review. medRxiv 2023. [CrossRef]
35. Arora, A.; Arora, A. The promise of large language models in health care. Lancet 2023, 401, 641. [CrossRef]
36. Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; Zhu, C. Gpteval: Nlg evaluation using gpt-4 with better human alignment. arXiv 2023, arXiv:2303.16634.

37. Moy, A.J.; Schwartz, J.M.; Chen, R.; Sadri, S.; Lucas, E.; Cato, K.D.; Rossetti, S.C. Measurement of clinical documentation burden among physicians and nurses using electronic health records: A scoping review. J. Am. Med. Inform. Assoc. 2021, 28, 998–1008. [CrossRef] [PubMed]
38. Sorin, V.; Klang, E.; Sklair-Levy, M.; Cohen, I.; Zippel, D.B.; Balint Lahat, N.; Konen, E.; Barash, Y. Large language model (ChatGPT) as a support tool for breast tumor board. NPJ Breast Cancer 2023, 9, 44. [CrossRef] [PubMed]
39. Lahat, A.; Shachar, E.; Avidan, B.; Shatz, Z.; Glicksberg, B.S.; Klang, E. Evaluating the use of large language model in identifying top research questions in gastroenterology. Sci. Rep. 2023, 13, 4164. [CrossRef] [PubMed]
40. Rao, A.; Kim, J.; Kamineni, M.; Pang, M.; Lie, W.; Succi, M.D. Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv 2023. [CrossRef]