MEDALIGN: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records

Scott L. Fleming[1,2,*], Alejandro Lozano[1,*], William J. Haberkorn[3,4,*], Jenelle A. Jindal[5,*], Eduardo Reis[6,7,8,*], Rahul Thapa[9], Louis Blankemeier[10], Julian Z. Genkins[11,12], Ethan Steinberg[2,5], Ashwin Nayak[13], Birju Patel[5], Chia-Chun Chiang[14,15], Alison Callahan[5,13], Zepeng Huo[5], Sergios Gatidis[6], Scott Adams[6], Oluseyi Fayanju[13], Shreya J. Shah[13], Thomas Savage[1,16], Ethan Goh[5,17], Akshay S. Chaudhari[1,6,15], Nima Aghaeepour[1,3,4], Christopher Sharp[13,15], Michael A. Pfeffer[9,13], Percy Liang[2,15], Jonathan H. Chen[5,15,16,17], Keith E. Morse[4], Emma P. Brunskill[2,15,†], Jason A. Fries[5,†], Nigam H. Shah[9,13,15,17,†]

[1] Department of Biomedical Data Science, Stanford School of Medicine, Stanford, CA, USA
[2] Department of Computer Science, Stanford School of Engineering, Stanford, CA, USA
[3] Department of Anesthesiology, Peri-operative, and Pain Medicine, Stanford School of Medicine, Stanford, CA, USA
[4] Department of Pediatrics, Stanford School of Medicine, Stanford, CA, USA
[5] Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[6] Department of Radiology, Stanford School of Medicine, Stanford, CA, USA
[7] Center for Artificial Intelligence in Medicine and Imaging (AIMI), Stanford University, Stanford, CA, USA
[8] Hospital Israelita Albert Einstein, Sao Paulo, SP, Brazil
[9] Technology and Digital Solutions, Stanford Health Care, Palo Alto, CA, USA
[10] Department of Electrical Engineering, Stanford School of Engineering, Stanford, CA
[11] Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA
[12] Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA
[13] Department of Medicine, Stanford School of Medicine, Stanford, CA, USA
[14] Department of Neurology, Mayo Clinic, Rochester, MN, USA
[15] Human-Centered Artificial Intelligence Institute, Stanford University, Stanford, CA, USA
[16] Division of Hospital Medicine, Stanford University, Stanford, CA, USA
[17] Clinical Excellence Research Center, Stanford School of Medicine, Stanford, CA, USA
{scottyf, lozanoe}@stanford.edu

Abstract

The ability of large language models (LLMs) to follow natural language instructions with human-level fluency suggests many opportunities in healthcare to reduce administrative burden and improve quality of care. However, evaluating LLMs on realistic text generation tasks for healthcare remains challenging. Existing question answering datasets for electronic health record (EHR) data fail to capture the complexity of information needs and documentation burdens experienced by clinicians. To address these challenges, we introduce MEDALIGN, a benchmark dataset of 983 natural language instructions for EHR data. MEDALIGN is curated by 15 clinicians (7 specialities), includes clinician-written reference responses for 303 instructions, and provides 276 longitudinal EHRs for grounding instruction-response pairs. We used MEDALIGN to evaluate 6 general domain LLMs, having clinicians rank the accuracy and quality of each LLM response. We found high error rates, ranging from 35% (GPT-4) to 68% (MPT-7B-Instruct), and 8.3% drop in accuracy moving from 32k to 2k context lengths for GPT-4. Finally, we report correlations between clinician rankings and automated natural language generation metrics as a way to rank LLMs without human review. MEDALIGN is provided under a research data use agreement to enable LLM evaluations on tasks aligned with clinician needs and preferences.

Introduction

Large language models (LLMs) have revolutionized natural language processing in tasks such as reading comprehension, reasoning, and language generation (Zhao et al. 2023), prompting researchers to explore applications in healthcare (Thirunavukarasu et al. 2023). Recent LLMs like MedPalm (Singhal et al. 2023) and GPT-4 (Nori et al. 2023) have demonstrated expert-level performance on medical question-answering benchmarks including MedQA (Jin et al. 2021), MMLU (Hendrycks et al. 2020), and the USMLE (Kung et al. 2023). However, these

[*] These authors contributed equally.
[†] Equal leadership.

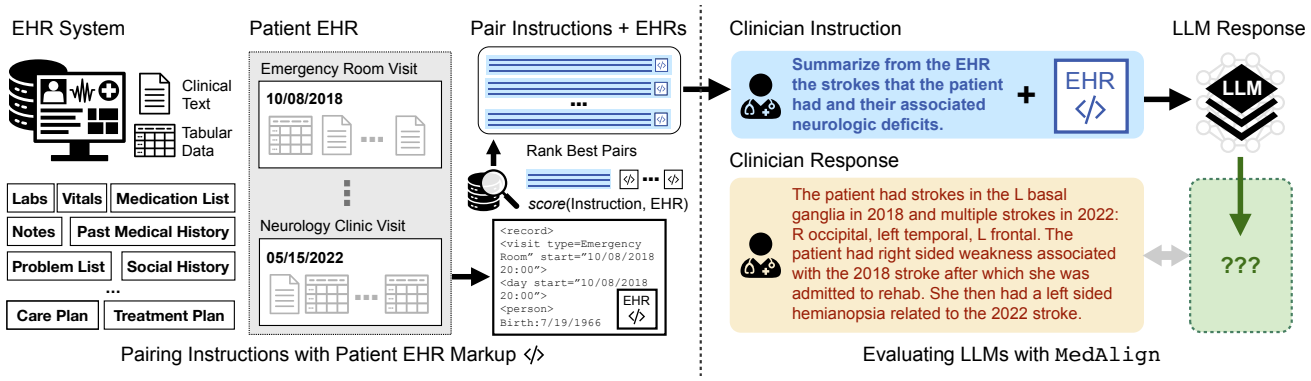

Figure 1: In MEDALIGN, patient EHRs are transformed into XML markup and paired with clinician-generated instructions using a retrieval-based (BM25) scoring metric. The resulting set of instruction + EHR pairs is then reviewed by clinicians to write gold responses, which are used to evaluate EHR instruction following in large language models

benchmarks employ multiple-choice, exam-style evaluations where question stems summarize key information and a single answer choice is best. It is not known if performance on these tasks will translate when a model is deployed in the complex clinical environments.

To be useful, LLMs need to perform well on the specific information-related tasks that clinicians currently complete themselves while caring for patients. These tasks are a significant burden on clinicians, who spend 45% of their day interacting with computers instead of patients (Toscano et al. 2020) and 10 hours a week generating documentation (Gaffney et al. 2022), in part contributing to professional burnout (Muhiyaddin et al. 2021). Examples of these tasks include summarizing a patient's asthma treatment history from different specialists the patient has visited, generating a differential diagnosis based on partially resulted laboratory data, or searching through the clinical notes for mentions of a patient's family support system in order to create the best plan for the patient's hospital discharge (see Table 2). Such tasks could be passed as instructions to an LLM in the form of questions or imperatives (e.g., "Write a discharge summary") grounded in a patient's Electronic Health Record (EHR, an electronic representation of a patient's medical history). However, despite the excitement about LLMs to transform the practice of medicine, evaluations to date have not authentically represented the variety of tasks and idiosyncrasies of EHR data that clinicians face in the real world.

Given the recent emergence of instruction-following capabilities in LLMs (Wei et al. 2022a), there is potential for LLMs to ameliorate such administrative burden. Hand-curated exemplars of instructions and responses have been critical to improve performance of models (Chung et al. 2022), especially on clinical reasoning and knowledge recall tasks in the healthcare domain (Singhal et al. 2023). Thus, a high quality dataset of instruction-EHR-response tuples that represents the breadth of clinical tasks is essential not only as a shared benchmark, but potentially to accelerate the training of specialized LLMs for healthcare (Shah, Entwistle, and Pfeffer 2023).

However, building such a dataset requires an extraordinary effort from a multidisciplinary collaboration. In particular, generating an instruction-following benchmark dataset with representative EHR-based tasks and expert responses is challenging due to the substantial cost and logistical complexity of clinician review. There is a need for an EHR dataset that (1) contains a diverse set of questions and instructions generated by practicing clinicians; (2) pairs these queries with EHRs from both inpatient and ambulatory care settings; (3) leverages both structured and unstructured data from the longitudinal EHR; and (4) is available to the broader academic community.

In light of these challenges and opportunities, we present three contributions:

1. **MEDALIGN Dataset:** We introduce a benchmark dataset called MEDALIGN consisting of 983 questions and instructions submitted by 15 practicing clinicians spanning 7 medical specialties. For 303 of these instructions, we provide a clinician-written reference answer and paired EHR for grounding prompts. Each clinician evaluated and ranked outputs from 6 different LLMs on these 303 instructions and wrote "gold standard" answers. To our knowledge, MEDALIGN is the first dataset of EHR-based instruction-answer pairs (including question *and* imperative instructions) written by clinicians, with clinician evaluations of LLM-generated outputs. Table 1 summarizes MEDALIGN and its distinction from existing datasets for clinical information needs.
2. **Automated Instruction-EHR Matching:** We demonstrate the feasibility of a simple retrieval-based approach to pair an instruction with a relevant patient EHR. By isolating the process of instruction solicitation, we were able to scale and diversify the set of clinicians who submitted instructions. Furthermore, we show that our process for matching instructions to relevant EHRs produces a relevant pairing 74% of the time — at least twice as frequently as randomly pairing instructions to EHRs.
3. **Automated Evaluation of LLM Responses:** We analyze the correlation between clinician rankings and automated natural language generation (NLG) metrics as a

Dataset	Questions	Documents	Patients	Specialties	Labeler	Source
(Raghavan et al. 2018)	5696	71	71	-	Medical Students	Clinical Note
(Pampari et al. 2018)	73111	303	303	-	Programmatic	Discharge Summary
(Fan 2019)	245	138	-	1	Author	Discharge Summary
(Yue et al. 2021)	1287	36	-	-	Medical Experts	Clinical Note
(Soni et al. 2022)	3074	1009	100	1	Clinicians	Radiology Note
MEDALIGN (Ours)	983	37264	276	7	Clinicians	EHR

Table 1: Comparison of our work to existing EHR QA datasets.

way to scalably reproduce such analyses, reducing future needs for clinicians to label and rank LLM responses.

Background and Related Work

The volume of patient care data is growing exponentially, with a compound annual growth rate approaching 36% (Culbertson 2021). Utilizing LLMs to more efficiently interact with patient data holds great potential to help clinicians manage increasingly complicated information needs and circumvent low-usability EHR interfaces (Melnick et al. 2020). However, evaluation of LLMs to improve meaningful outcomes like clinician burnout or patient health has been inadequately studied, mainly due to benchmark datasets which do not represent true clinician needs (Henry et al. 2020), narrowly focus on a specific medical specialty or subset of EHR data (Lehman et al. 2022), and/or are overly simplistic due to templated question construction (Pampari et al. 2018; Yue, Gutierrez, and Sun 2020). These works highlight the challenges in collecting high-quality clinician-generated questions and answers; we consider each in turn.

Questions and instructions in an EHR-based benchmark dataset should be paired with relevant patient EHRs. In order to ensure relevancy, prior works have provided clinicians with specific patient EHRs and asked them to generate questions based on those patients' data (Lehman et al. 2022). Unfortunately, requiring EHRs as context for question generation limits scalability, as medical institutions restrict access to patient data to preserve patient privacy. Pampari et al (2018) attempted to overcome these scalability issues by generating questions via a template-based approach, but this led to issues with question quality and diversity (Yue, Gutierrez, and Sun 2020). Our method of soliciting clinician-generated instructions without a specific patient's EHR as context overcomes these scaling issues, albeit at the cost of potentially less relevant instruction-to-EHR pairings (we discuss our approach to addressing this problem in the Dataset Curation section).

Beyond generating questions, generating expert answers at scale is also prohibitively difficult. Reviewing an EHR to answer patient-specific queries can take 30+ minutes for a single patient (Siems et al. 2020). This excludes any time required to generate a response to the query. Prior works have attempted to overcome the bottleneck of generating responses by extracting answers verbatim from individual clinical notes or discharge summaries (Soni et al. 2022; Oliveira et al. 2021; Fan 2019). However, many clinical tasks require synthesizing information from both structured data and multiple free-text documents to arrive at an adequate response, an aspect not explored in existing EHR QA datasets. In such cases, answers extracted from a single note in the patient's record may not be an adequate; free-text text generation is required. While there is at least one example of an EHR-based question answering dataset in the literature that includes both structured and unstructured data (Raghavan et al. 2018), it neither contains free-text responses nor is publicly available. Finally, all of the aforementioned datasets focus on simple question answering (i.e., providing concise, factoid-style answers) rather than general instruction following, which often requires executing a series of complex directives and commands to accomplish tasks. To the best of our knowledge, there does not exist *any* EHR-based benchmark dataset that incorporates instruction following.

The significant costs of clinician review present barriers not only for *de novo* dataset generation, but also for reliable evaluation of new methods on existing datasets. Automated metrics for evaluating Natural Language Generation (NLG) systems have shown moderate to high correlation with human judgments on tasks like machine translation (Freitag et al. 2022), but it is unclear whether these findings extend to other domains and tasks. While there is precedent (Lehman et al. 2022) for *applying* automated metrics like BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), and BERTScore (Zhang et al. 2020) to NLG tasks in the clinical domain, there is comparatively very little work assessing correspondence between these metrics and human judgment on clinical NLG tasks. Thus not only do we have a poor understanding of how LLMs perform on EHR-based instruction-following tasks, but also we do not know whether it is possible to reliably automate such evaluations. Automation could substantially reduce the "barrier to entry" for research teams with limited resources.

Dataset Curation Process

Electronic Health Records (EHRs) EHR systems are software for managing patient medical record data. From a clinician's view, a patient EHR is accessed via a graphical user interface that provides access to data elements associated with medical care, e.g., medication lists, treatment plans. These data are stored as a collection of timestamped structured (tabular) and unstructured (text) events, which when ordered by time form a patient's longitudinal EHR

timeline. Our EHR data is represented using the OMOP CDM, a standardized schema for exchanging medical data, translated into a single, XML markup document per record (example provided in Appendix) to enable simple data exploration via an XML viewer. Figure 1 outlines the workflow for building MEDALIGN including: (1) pairing clinician-generated instructions with patient EHR markup, and (2) evaluating language model responses against gold responses written by clinicians.

Collection Protocol Reviewing patient medical data requires adhering to strict security protocols to protect patient privacy and prevent protected health information (PHI) leaks. This motivated our 3-stage curation process: (1) online instruction collection from clinicians; (2) instruction-EHR matching; and (3) response generation. Note we deliberately decouple instruction collection from response generation. This enables sampling a larger set of instructions from a more diverse set of clinician specialities while minimizing exposure to patient data. However, this approach requires defining a matching function to pair instructions with relevant patient EHRs, a process which may generate errors due to irrelevant instruction-EHR pairings. We discuss the performance of a retrieval-based matching system below.

Stage 1: Collecting Instructions Clinicians were recruited in our academic medical center via email. Through the use of an online form, clinicians were asked to submit instructions as posed to a hypothetical AI assistant designed to facilitate EHR-based tasks. Participants were instructed to envision a clinical vignette typical of their daily practice and to formulate an instruction that the AI could perform to make their work easier, faster, and less stressful. For each instruction, participants were asked to provide metadata to assist in matching the instruction to a patient, including pertinent clinical characteristics and the clinical context where the instruction could be used, e.g., using contrast in a CT scan. See the Appendix for all collected fields.

Stage 2: Instruction-EHR matching All submitted instructions include metadata information (by request) on their clinical context and target patient population. We used instructions tagged "applicable to patients generally" to maximize their relevance in EHR matching. We evaluated two methods for matching instructions with EHRs: (1) a simple baseline based on uniform random sampling; and (2) a retrieval-based method using BM25Okapi (Trotman, Puurula, and Burgess 2014).

For the retrieval approach, we concatenated every instruction with its corresponding patient characteristics and clinical context to construct a search query. We used this query to retrieve the 5 most relevant EHRs within a randomly selected subsample of patients (77200) from our hospital database. This same subsample was used to match patients for our baseline uniform random sample. After matching, the authors conducted a manual review to assess binary relevance of all generated instruction-EHR pairs.

Stage 3: Instruction Response Generation For this stage, clinicians were tasked with reviewing the instruction and associated EHR data, then writing a response to that instruction. Whenever feasible, instructions were assigned to clinicians within the same specialty as the original submitter but not the original submitter themselves. In cases where this was not possible, the instruction was randomly assigned to a clinician, in any specialty, that did not submit the instruction. Clinicians were asked whether the instruction could be feasibly applied to the patient in the EHR (e.g., not asking about smoking history in an infant) and if the EHR contained all necessary information to answer the instruction. They then manually generated an expert response to the instruction. This response was intended to be brief and clinically relevant, drawing on any information available in the supplied EHR record, as well as any appropriate external references. The most recent timestamp in the EHR was designated as the "time anchor", meaning the response was written as if the instruction had been posed at that point in time.

Category	Example Instruction	Gold	All
Retrieve & Summarize	Summarize the most recent annual physical with the PCP	223	667
Care Planning	Summarize the asthma care plan for this patient including relevant diagnostic testing, exacerbation history, and treatments	22	136
Calculation & Scoring	Identify the risk of stroke in the next 7 days for this TIA patient	13	70
Diagnosis Support	Based on the information I've included under HPI, what is a reasonable differential diagnosis?	4	33
Translation	I have a patient that speaks only French. Please translate these FDG-PET exam preparation instructions for her	0	2
Other	What patients on my service should be prioritized for discharge today?	41	75
Total		303	983

Table 2: MEDALIGN categories and example instructions.

Dataset Description

Instructions Collected A total of 15 clinicians submitted instructions during the data collection process. These medical practitioners represented 7 distinct specialties, which included Internal Medicine (492 instructions submitted), Neurology (320), Radiology (402), Cardiology (71), Oncology (14), Surgery (12), and Primary Care (3). Clinicians provided a varying number of instructions ranging from 1 to 278 with a mean of 87 instructions per clinician (see Appendix). From the 1314 instructions collected, 455 were marked as applicable to patients generally and 859 were relevant only

Model	Context	Correct ↑	WR ↑	Rank ↓
GPT-4 (MR)	32768†	**65.0%**	0.658	2.80
GPT-4	32768	60.1%	**0.676**	**2.75**
GPT-4	2048*	51.8%	0.598	3.11
Vicuña-13B	2048	35.0%	0.401	3.92
Vicuña-7B	2048	33.3%	0.398	3.93
MPT-7B	2048	31.7%	0.269	4.49

Table 3: Human evaluation of LLM responses. Context: The model's context length, using its native tokenizer. Correct: The percentage of model responses deemed correct by clinicians. WR: Average win-rate marginalizing over model pairings. Rank: Empirical mean of human-assigned rankings. †With multi-step refinement the effective context length is infinite, as the model observes the entire EHR albeit in small chunks at a time. *For GPT-4 (2k) we used the GPT-4 32k models from OpenAI but restricted its context length using the Vicuña-native tokenizer for direct comparison.

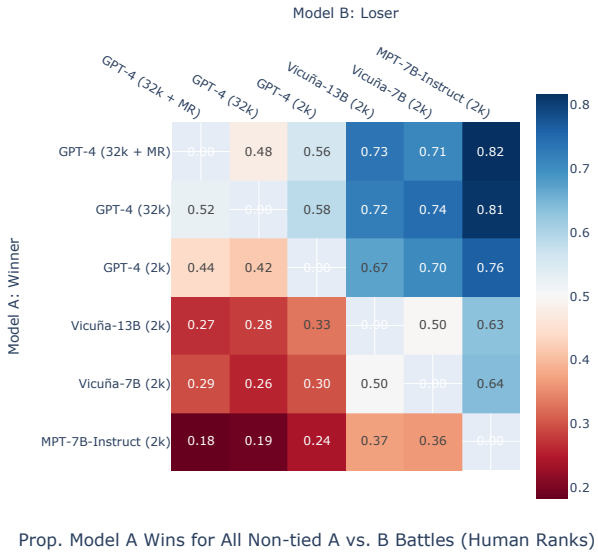

Figure 2: Head-to-head comparison of model performance based on human ranks. The number in row i, column j indicates the proportion of instructions for which the response generated by the model in row i was strictly preferred over the model in row j. Compare to Figure 3 constructs the same matrix but using rankings derived from COMET, an automated metric, rather than clinician-generated rankings.

to patients with specific clinical characteristics. We removed near-identical instructions (defined by a ROUGE-L similarity above 0.7), yielding 983 instructions of which 407 were marked as applicable to patients generally.

Instruction-EHR Matches Based on evaluation by the authors, for 240 (59%) of the instructions applicable to "patients in general" the first record retrieved by BM25 was relevant. For 303 instructions (74%), at least one of the top 5 EHRs returned by BM25 was relevant. In contrast, only 38% of EHRs retrieved via uniform random sampling were deemed relevant.

Instruction Taxonomy To better understand higher-level themes within the instructions submitted, a practicing clinician developed a taxonomy of instructions. This taxonomy, described in detail in the Appendix, includes 6 categories spanning 20 subcategories. We summarize the distribution of instruction categories across the set of all instructions submitted and those that received responses from a clinician in Table 2. We include further analysis in the Appendix.

Benchmarking LLM Performance

LLM Selection We evaluated six distinct LLMs, chosen to capture both state-of-the-art, closed-source LLM capabilities available to consumers via an API as well as smaller, open-source and user-modifiable LLMs with more lenient commercial licensing (e.g., MosaicML's MPT-7B model). Additionally, we designed our experiments to directly evaluate the impact of model parameters and context length.

For a state-of-the-art LLM, we selected GPT-4 (through Microsoft's Azure OpenAI HIPAA compliant gpt-4-32k-0301 API) due to its state-of-the-art performance on various medical tasks, its long 32k context length, and its availability to researchers and clinics. However, despite this context length, it proved insufficient for accommodating full EHRs (more than 80% of EHRs in MEDALIGN contain more than 32k tokens, see Appendix). To address this limitation, we explored a multi-step refinement (MR) approach to maximize effective context length. In this approach, the EHR is divided into "chunks" designed to be as big as possible (30k tokens, without concern for maintaining valid XML structure) while still fitting within the model's context length. A response to the instruction is generated using the chronologically first/earliest EHR "chunk" as context, then the second "chunk" is given to the model and the model is instructed to update its response if appropriate or maintain the same response otherwise, and so on, until the entire EHR has been fed through the model. We acknowledge the potential effectiveness of other methods, such as Retrieval Augmented Generation (RAG), in answering questions regarding long documents. However, our primary interest was in measuring LLM's ability to discern and utilize clinically relevant material when answering questions about the EHR. While methods such as RAG would likely be performant in this area, they would not have enabled us to assess the LLM's ability to sift through irrelevant material by nature of its design.

For smaller, open-source models we evaluated Vicuña-7B and Vicuña-13B (Chiang et al. 2023) as well as MPT-7B-Instruct (MosaicML 2023). These models are widely available and user-modifiable with favorable licensing agreements, but they have considerably smaller context lengths (2048 tokens) compared to GPT-4. To enable more direct comparisons, we assessed GPT-4 under a restricted context length designed to exactly match the context length of the Vicuña model.

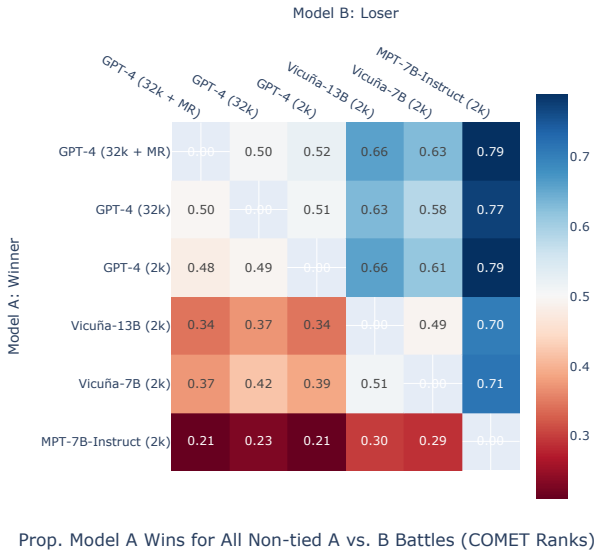

Figure 3: Head-to-head evaluation of model performance using COMET Ranks (compare to Figure 2). Model win rates using COMET follow a similar pattern as to model win rates using human rankings.

Generating LLM Responses to EHR-based Questions and Instructions Using a standard prompt template (see Appendix), each model was tasked to fulfill the given instruction grounded on its corresponding EHR pair. Due to current models' context length restrictions, EHRs needed to be truncated. This truncation process involved taking each model's maximum context length (in number of tokens under that model's specific tokenizer), reserving 256 tokens for generation, and subtracting any tokens used for the corresponding structured prompt and instruction. This truncation was performed by counting tokens from the end of the record, ensuring that as much recent information as possible was retained.

Clinician Evaluation of LLM Responses Nine clinicians were asked to evaluate and rank the responses generated by 6 separate LLMs. Clinicians did not evaluate their own responses or responses to instructions that they submitted. When feasible, clinicians evaluated responses to instructions that were written by a clinician in their same specialty. The instructions and EHRs reviewed by the clinicians were exactly the same in structure and content as those provided to the LLMs. Clinicians recorded a binary evaluation of whether the response was correct or incorrect, with "incorrect" defined as meeting at least one of the following criteria:

1. Response is not clinically appropriate based on the available EHR information;
2. Response includes errors that, if corrected, would change the clinical interpretation;
3. Response does not address the instruction.

Responses *not* marked as "incorrect" were deemed to be "correct". Clinicians then ranked the quality of the LLM responses based on which provided the most clinically relevant and appropriate response. Ties were permitted. The clinicians were blinded to which LLM generated each output, and the order of LLM output was reshuffled for each instruction. Each clinician reviewed 49 instruction-patient pairs on average, yielding 303 pairs reviewed overall with 50 instruction-EHR pairs being reviewed by three clinicians.

Overall, we found that more than half of the responses generated by the GPT-4 variants we tested were deemed correct by clinicians (65% for GPT-4 (32k + MR), 60.1% for GPT-4 (32k), 51.8% for GPT-4 (2k)). By contrast, only about one in three responses generated by the Vicuña and MPT-7B models were considered correct (35% for Vicuña-13B, 33.3% for Vicuña-7B, 31.7% for MPT-7B-Instruct; see Table 3). In head-to-head comparisons, GPT-4 without context length restriction was preferred over the Vicuña-13B model in 72% of instances, and preferred over MPT-7B-Instruct 81% of the time (see Figure 2). The GPT-4 model with 32k context length and no multi-step refinement had the highest overall average win-rate against all other models (0.676).

Automated Evaluation of LLM Responses

With the aim to to find an automated proxy for clinician-in-the-loop evaluation, we analyzed the correlation between a suite of automated metrics and human preference rankings using the Kendall's Rank Correlation. We also calculated the inter-rater correlation between human rankers, yielding a mean Kendall's Tau coefficient of 0.44. The average correlations between metrics and human rankings is shown in Table 4. As noted by previous studies (Nimah et al. 2023), the majority of these metrics have shown moderate correlation with human preference and are widely reported in NLG tasks.

We evaluated each model output using both source-free (SF) and source-augmented (SA) automated metrics. Source-free metrics compare a model's output to a gold standard reference answer (in our case generated by a clinician) without the use of any additional context or sources (i.e., without any information from the EHR). We selected BERTScore (Zhang et al. 2020), METEOR (Banerjee and Lavie 2005), chrF++ (Popović 2017), GoogleBLEU (Wu et al. 2016), and ROUGE-L (Lin 2004) due to their availability and wide use. Source-augmented metrics consider source (e.g., the EHR) in addition to a gold reference and model output. The SA metrics we considered (and the LMs they use) include UniEval (T5 -large) (Zhong et al. 2022) and COMET (XLM-RoBERTa) (Rei et al. 2020). As these models have limited context length we used the BM25Okapi algorithm to retrieve relevant snippets from within the patient's EHR using the instruction as a search query.

Overall, COMET (Rei et al. 2020) exhibited the strongest correlation with clinician preference rankings, approaching the level of human inter-reviewer reliability (0.37 vs. 0.44). As seen in Figures 2 and 3, the overall trends of head-to-head comparisons were preserved when using COMET as the source of model output rankings vs. clinician-generated rankings. Specifically, the GPT-4 were

Automated Metric	Source Augmented	Avg. Corr.	95% CI
COMET	✓	0.37	0.33-0.41
BERTScore		0.34	0.30-0.38
METEOR		0.32	0.28-0.36
chrF++		0.29	0.25-0.33
GoogleBLEU		0.29	0.25-0.33
ROUGE-L		0.27	0.23-0.31
BLEURT		0.25	0.21-0.30
LENS		0.18	0.14-0.22
UniEval Relevance	✓	0.27	0.23-0.32
UniEval Fluency	✓	0.11	0.06-0.15
UniEval Coherence	✓	0.09	0.04-0.13
UniEval Consistency	✓	0.09	0.04-0.13
UniEval Overall	✓	0.20	0.15-0.24
Inter-Rater Reliability		0.44	0.34-0.53

Table 4: Correlation (mean Kendall's Tau) between automated metrics' ranking and human ranking of LLM outputs. Mean Kendall's Tau between human reviewers was 0.43.

consistently preferred over the Vicuña and MPT-7B models by both COMET and clinicians, and the Vicuña models were consistently preferred over the MPT-7B model. Within the GPT-4 variants and between the two Vicuña models considered, win-rate preferences were not necessarily preserved, suggesting utility of COMET as a reasonable but perhaps coarse measure of model performance in this setting. The next most correlated metric with human rankings after COMET was BERTScore, a source-free metric, with an average correlation coefficient of 0.34.

Using our best performing automated metrics, COMET and BERTScore, we evaluated four recently released instruction-tuned medical LLMs (all based on Llama-2 (Touvron et al. 2023)): AlpaCare (Zhang et al. 2023), ClinicalCamel (Toma et al. 2023) and Med42 (Christophe et al. 2023). Figure 4 shows that current medical instruction tuning largely causes worse performance in MEDALIGN vs. the base Llama-2 model.

Discussion and Conclusion

Readily available datasets and benchmarks for easy-to-evaluate tasks like closed-form question answering have helped to measure the remarkable progress of LLMs, even in medical domains (Kung et al. 2023). However, logistical difficulties and significant labeling costs have hindered progress towards establishing a shared dataset and benchmark for tasks amenable to LLMs and which truly represent clinician needs. We share such a benchmark dataset with the research community, which takes a novel approach towards instruction gathering by modularizing and isolating the process of solicitation and EHR pairing. To the best of our knowledge, our dataset is the first to evaluate LLM performance on clinician-generated questions and instructions using comprehensive, longitudinal EHRs. This affords several new insights.

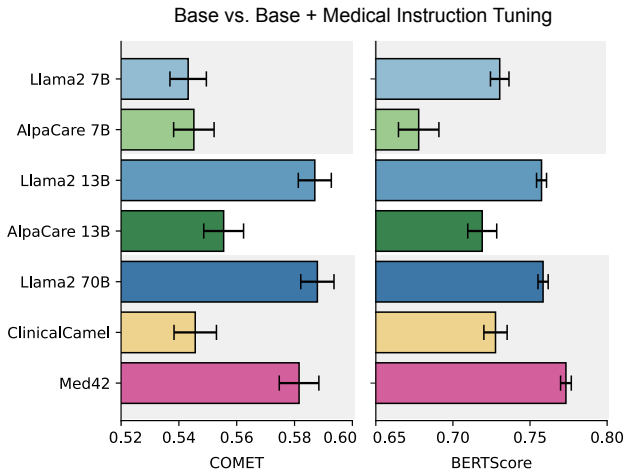

Figure 4: Automated evaluation of medical instruction-tuned LLMs vs. general instruction-tuned counterparts using the best-performing metrics (COMET and BERTScore).

The Importance of Context Length. While GPT-4 with a restricted context length of 2048 tokens achieved a correctness rate of 51.8%, the exact same GPT-4 model given 32000 tokens of context from the EHR achieved a correctness rate of 60.1%. Thus the additional context length yielded an additional 8.3% in the proportion of correct responses. Given the sheer quantity of tokens and concepts contained within comprehensive EHRs, including in MEDALIGN (see Appendix), it is perhaps not surprising that instruction following performance was poor with a limited context length. Indeed, not a single EHR in MEDALIGN can fit entirely within the Vicuña or MPT-7B's 2048 context length, and only 19.6% of these records can entirely fit within the 32k context length afforded by GPT-4. This highlights the importance of context length in applying LLMs to EHR-based tasks and motivates efforts to increase context lengths via e.g., methods that do so implicitly via position interpolation (Chen et al. 2023) or approaches that explicitly improve the training efficiency of mathematical operations (Dao et al. 2022).

Misalignment with Current Benchmarks Medical instruction tuning in academic models currently favors shorter contexts, optimizing for tasks like MedQA and MMLU. MedQA, consisting of USMLE-style questions covering diagnosis support and care planning, is a popular choice for assessing the medical skills of an LLM. However, USMLE-style questions only comprise 17% of clinician-submitted instructions in MEDALIGN while 68% involve retrieving and summarizing data from the EHR. Our results highlight that current medical instruction tuning degrades performance in longer context tasks, with base Llama-2 models outperforming medical instruction-tuned LLMs in most cases. Given the importance of long contexts and summarization skills in addressing clinician information needs, our work underscores the need to evaluate instruction tuning tasks beyond MedQA and similar narrow benchmarks.

Limitations. Our approach of first soliciting instructions and *then* pairing these instructions to EHRs can increase the scale and diversity of instructions collected, but at a cost. Despite yielding almost twice as many relevant pairings as simply randomly selecting an EHR for each instruction, our BM25 approach did not yield a relevant match for approximately 30% of instructions. In other words, while an instruction submitted by a clinician was of course relevant to the *hypothetical* patient they had in mind at the time of submission, it frequently ended up not being relevant to an *actual* patient EHR. There are potential ways to improve this matching process e.g., by using vector databases powered by BERT-style models which could better capture semantic alignment between queries and EHRs relative to BM25 (Wei et al. 2022b). Additionally, while we solicited instructions from a large number of clinicians at our academic medical center with diverse specialties and backgrounds, the clinicians who submitted data to MEDALIGN represent only a small fraction of the overall clinician workforce.

Conclusion. This work establishes, for the first time, the performance of some of the most capable LLMs available — GPT-4, LLaMA, and MPT-7B-Instruct — on EHR-based instruction-following tasks. We find that approximately one-third of the best-performing LLM's responses are incorrect. The benchmark dataset we share, MEDALIGN enables researchers to measure what matters and focus on tasks that are clinically relevant with significant potential positive impact. In addition, our findings establishing significant correlation between human preference and existing automated metrics provide a path for researchers to make technical progress without requiring the organizational infrastructure for clinical labeling. Finally, our novel approach towards soliciting clinician instructions paves the way for even larger-scale data collection efforts, both for training and evaluation purposes.

Ethics Statement

Security and Compliance. A university institutional review board granted approval for this study (reference number 57916). All authors handling data individually completed institutional HIPAA and data privacy training prior to engagement with the data. All models exposed to data were deployed within HIPAA-compliant compute infrastructure.

Privacy and Data Deidentification All data were de-identified using a "hiding in plain sight" protocol wherein protected health information (PHI) is replaced by coherent synthetic alternatives (Carrell et al. 2013), e.g., tagging all person names and replacing them with a randomly generated name. For the research release of the MEDALIGN dataset, all documents will undergo human review to minimize risk of inadvertently exposing PHI. The dataset will be hosted in an university-approved, secure data portal and will require user credentialing to access, i.e., completing CITI ethnics training and agreeing to the terms of our data use agreement.

Patient Consent Every patient at our medical center has provided their signature on a privacy notice, which explains that their medical records could be utilized for research. This data, once de-identified, is accessible to researchers under a comprehensive IRB protocol of the university.

Societal impact. LLMs could streamline clinician workflows within the EHR by replacing clunky point-and-click interfaces with natural language interactions, improving clinician efficiency. Muhiyaddin et al. (2021) found EHR-related documentation tasks to be a leading cause of physician burnout, resulting in low-quality care, costly turnover, and a decline in patient safety. By easing documentation burden, LLMs could thus increase care quality, decrease clinician turnover, and improve patient safety. MEDALIGN provides a way to assess whether LLMs are safe and ready for the deployments necessary to realize these potential benefits.

Introducing LLMs into the clinic also poses potential risks. Even the best-performing model of those we assessed (GPT-4) produced incorrect responses for more than 33% of the clinician-generated instructions. These errors could *decrease* patient safety by leading to poor clinical decision making. More insidiously, a recent study by Omiye et al. (2023) noted that commercial LLMs propagate harmful race-based stereotypes in medicine. We analyzed LLM performance differences across race in MEDALIGN (see Appendix) and found minimal disparities, but more work is needed. Additionally, we did not examine specific failure modes like hallucination and leave this for future work.

References

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Carrell, D.; Malin, B.; Aberdeen, J.; Bayer, S.; Clark, C.; Wellner, B.; and Hirschman, L. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2): 342–348.

Chen, S.; Wong, S.; Chen, L.; and Tian, Y. 2023. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Christophe, C.; Gupta, A.; Hayat, N.; Kanithi, P.; Al-Mahrooqi, A.; Munjal, P.; Pimentel, M.; Raha, T.; Rajan, R.; and Khan, S. 2023. Med42 - A Clinical Large Language Model.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Culbertson, N. 2021. The skyrocketing volume of healthcare data makes privacy imperative. *Forbes Technology Council Post*.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention

with io-awareness. *Advances in Neural Information Processing Systems*, 35: 16344–16359.

Fan, J. 2019. Annotating and characterizing clinical sentences with explicit why-QA cues. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 101–106.

Freitag, M.; Rei, R.; Mathur, N.; Lo, C.-k.; Stewart, C.; Avramidis, E.; Kocmi, T.; Foster, G.; Lavie, A.; and Martins, A. F. 2022. Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68.

Gaffney, A.; Woolhandler, S.; Cai, C.; Bor, D.; Himmelstein, J.; McCormick, D.; and Himmelstein, D. U. 2022. Medical documentation burden among US office-based physicians in 2019: a national study. *JAMA Internal Medicine*, 182(5): 564–566.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; and Uzuner, O. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1): 3–12.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.

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. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2): e0000198.

Lehman, E.; Lialin, V.; Legaspi, K. E.; Sy, A. J.; Pile, P. T.; Alberto, N. R.; Ragasa, R. R.; Puyat, C. V.; Taliño, M. K.; Alberto, I. R.; Alfonso, P. G.; Moukheiber, D.; Wallace, B.; Rumshisky, A.; Liang, J.; Raghavan, P.; Celi, L. A.; and Szolovits, P. 2022. Learning to Ask Like a Physician. In Naumann, T.; Bethard, S.; Roberts, K.; and Rumshisky, A., eds., *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 74–86. Seattle, WA: Association for Computational Linguistics.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Melnick, E. R.; Dyrbye, L. N.; Sinsky, C. A.; Trockel, M.; West, C. P.; Nedelec, L.; Tutty, M. A.; and Shanafelt, T. 2020. The association between perceived electronic health record usability and professional burnout among US physicians. In *Mayo Clinic Proceedings*, volume 95, 476–487. Elsevier.

MosaicML, N. T. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. Accessed: 2023-05-05.

Muhiyaddin, R.; ElFadl, A. H.; Mohamed, E.; Shah, Z.; Alam, T.; Abd-alrazaq, A. A.; and Househ, M. S. 2021. Electronic Health Records and Physician Burnout: A Scoping Review. *ICIMTH*, 289: 481–484.

Nimah, I.; Fang, M.; Menkovski, V.; and Pechenizkiy, M. 2023. NLG Evaluation Metrics Beyond Correlation Analysis: An Empirical Metric Preference Checklist. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1240–1266. Toronto, Canada: Association for Computational Linguistics.

Nori, H.; King, N.; McKinney, S. M.; Carignan, D.; and Horvitz, E. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Oliveira, L. E. S. e.; Schneider, E. T. R.; Gumiel, Y. B.; Luz, M. A. P. d.; Paraiso, E. C.; and Moro, C. 2021. Experiments on Portuguese clinical question answering. In *Brazilian Conference on Intelligent Systems*, 133–145. Springer.

Omiye, J. A.; Lester, J. C.; Spichak, S.; Rotemberg, V.; and Daneshjou, R. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1): 195.

Pampari, A.; Raghavan, P.; Liang, J.; and Peng, J. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368. Brussels, Belgium: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Popović, M. 2017. chrF++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, 612–618.

Raghavan, P.; Patwardhan, S.; Liang, J. J.; and Devarakonda, M. V. 2018. Annotating electronic medical records for question answering. *arXiv preprint arXiv:1805.06816*.

Rei, R.; Stewart, C.; Farinha, A. C.; and Lavie, A. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Online: Association for Computational Linguistics.

Shah, N. H.; Entwistle, D.; and Pfeffer, M. A. 2023. Creation and Adoption of Large Language Models in Medicine. *JAMA*.

Siems, A.; Banks, R.; Holubkov, R.; Meert, K. L.; Bauerfeld, C.; Beyda, D.; Berg, R. A.; Bulut, Y.; Burd, R. S.; Carcillo, J.; et al. 2020. Structured chart review: Assessment of a structured chart review methodology. *Hospital pediatrics*, 10(1): 61–69.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 1–9.

Soni, S.; Gudala, M.; Pajouhi, A.; and Roberts, K. 2022. RadQA: A Question Answering Dataset to Improve Comprehension of Radiology Reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6250–6259.

Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; and Ting, D. S. W. 2023. Large language models in medicine. *Nature medicine*, 29(8): 1930–1940.

Toma, A.; Lawler, P. R.; Ba, J.; Krishnan, R. G.; Rubin, B. B.; and Wang, B. 2023. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv:2305.12031.

Toscano, F.; O'Donnell, E.; Broderick, J. E.; May, M.; Tucker, P.; Unruh, M. A.; Messina, G.; and Casalino, L. P. 2020. How physicians spend their work time: an ecological momentary assessment. *Journal of General Internal Medicine*, 35: 3166–3172.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Trotman, A.; Puurula, A.; and Burgess, B. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, 58–65.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022a. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Wei, Z.; Xu, X.; Wang, C.; Liu, Z.; Xin, P.; and Zhang, W. 2022b. An Index Construction and Similarity Retrieval Method Based on Sentence-Bert. In *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, 934–938. IEEE.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yue, X.; Gutierrez, B. J.; and Sun, H. 2020. Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4474–4486. Online: Association for Computational Linguistics.

Yue, X.; Zhang, X. F.; Yao, Z.; Lin, S.; and Sun, H. 2021. Cliniqg4qa: Generating diverse questions for domain adaptation of clinical question answering. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 580–587. IEEE.

Zhang, T.; Koshre, V.; Wu, F.; Weinberger, K.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, X.; Tian, C.; Yang, X.; Chen, L.; Li, Z.; and Petzold, L. R. 2023. AlpaCare:Instruction-tuned Large Language Models for Medical Application. arXiv:2310.14558.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhong, M.; Liu, Y.; Yin, D.; Mao, Y.; Jiao, Y.; Liu, P.; Zhu, C.; Ji, H.; and Han, J. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2023–2038. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.