Trustworthiness in Medical Product Question Answering by Large Language Models

Daniel Lopez-Martinez
kdlopezm@amazon.com
Amazon
Santa Clara, California, USA

ABSTRACT

Large language models (LLMs) have achieved remarkable progress in recent years. These models have the capability to answer complex questions about medical disorders, their pathophysiology, etiology and corresponding interventions. However, when providing information about medical products and treatments, it is important to ensure that models respond reliably with factually correct information that adheres to product labels, and do not produce factual errors in which a claim contradicts established ground-truth knowledge. To this end, in this paper we propose an evaluation method to determine whether claims in LLM responses to questions about medical products are supported by FDA-approved product information.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Large Language Model, Evaluation, Medical, Trustworthiness.

ACM Reference Format:
Daniel Lopez-Martinez. 2024. Trustworthiness in Medical Product Question Answering by Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Recent advancements in generative artificial intelligence have led to the development of large language models (LLMs) capable of processing and generating text with human-like performance. These include open-source models such as Llama [80, 81], and commercial models such as OpenAI's ChatGPT [62] or Anthropic's Claude [5, 7]. These generalist models have achieved impressive results on varied benchmarks including question answering [10, 43, 49, 50, 63, 87]. In healthcare, they have the demonstrated remarkable capabilities in a number of complex expert tasks (e.g. providing differential diagnoses [39], summarizing charts [85], medical image analysis [2, 38], etc.) and shown potential to democratize medical knowledge and facilitate access to healthcare [25]. To this end, progress towards specialized medical LLMs advances rapidly [23, 70, 82, 86]. Fueled by their vast promise, both generalist and specialized LLMs are starting to be adopted in the real-world clinical setting to streamline clinical and administrative tasks [19, 56, 64, 72, 100], and a growing number of clinicians report using LLMs in their clinical practice or education [17, 54, 78, 79].

While LLMs hold vast promise and their capabilities are evolving at a breathtaking speed, their rapid adoption also introduces concerns about their trustworthiness [76]. Specifically, they can hallucinate and produce factual errors in which a claim in the response contradicts established ground-truth knowledge [3, 41, 76], despite appearing plausible and confident. In the medical context, incorrect information can pose significant risk to public health, and cause harm to individuals and organizations [1, 4, 9, 37, 44, 57, 90]. This issue is exacerbated by the potential for patients to use LLMs as a source of medical information, as they may rely wholly on an them for prognosis and treatment, thereby reducing or eliminating reliance on appropriate professional medical judgement and support [90]. Since incorrect information may be indistinguishable from factually accurate responses, patients may be provided with incorrect information. Hence, LLMs have the potential to result in patient harm and lead to severe health consequences, if not adequately deployed with robust guardrails and quality controls.

Given the real-life risks to public health of incorrect health-related information, it is paramount that LLMs are evaluated thoroughly prior to deployment. Even if model developers issue warnings of the potential limitations of LLMs, their misuse can still pose risks [90]. Hence, the development of methods to evaluate the answers of LLMs to medical questions is not just of academic interest but of great practical importance.

While previous works have evaluated the quality of LLMs responses to biomedical and clinical knowledge questions [12, 16, 22, 64, 73], in this work we focus into an overlooked issue that impacts most LLMs, that is, the potential to provide potentially harmful information about medical products, specifically drugs. Traditionally, specialized ML models have been trained to address a specific task using highly domain and problem-specific training data [21]. However, LLM models are trained on much more broadly available generalist datasets [51] with less hands-on human oversight in their development. Therefore, they can learn complex unvetted relationships from the training data and produce outputs about medical products that do not strictly adhere to the approved product labels. Promoting a medical product for anything other than its approved use, often denoted *off-label promotion*, can be unsafe if not done with adequate professional supervision [83]. Therefore, it is preferred that LLMs provide information that adhere to the approved labeling documents [11, 45, 48, 53, 75, 83, 84].

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN XXXXXXX.XXXXXXX
https://doi.org/XXXXXXX.XXXXXXX

To avoid this issue, building upon previous work on factuality evaluation of LLM responses [47, 58, 88], we propose a method to evaluate if LLM responses strictly adhere to FDA product labels (Sec.3). Our method uses a language model to first decompose a long-form response into individual claim. Then, each claim is evaluated to determine if it relates to one of the standardized FDA labeling document sections. Claims that relate to the labeling documents are fact-checked by comparing it against the corresponding labeling document to determine whether the claim is supported. We demonstrate this methodology using synthetic user questions and LLM responses from Claude 3 [5–7]. While many prior works have evaluated the factiality of LLM responses, to our knowledge this is the first work focused on evaluating medical product question answers and ensuring adherence to the information in the labeling documents.

2 BACKGROUND

Here we discuss medical product labeling in the USA and within-label and off-label promotion (Sec.2.1), the specific concerns impacting LLMs (Sec.2.2), previous work on detecting off-label promotion (Sec.2.3) and recent advances in medical question answering evaluation (Sec.2.4).

2.1 Medical product labeling

In the US, under the Federal Food, Drug and Cosmetic Act (FDCA), regulated by the Food and Drug Administration (FDA), medical products such as pharmaceutics, biologics or medical devices, must be approved, authorized, or otherwise cleared for each intended use by the FDA before a company can market it [83]. Off-label use refers to using or prescribing marketed medical products for indications (e.g. a disease or symptom) that are not included in their FDA-approved labeling information, as well as the use of a marketed product in a patient population (e.g. pediatric, pregnant, etc.), dosage, or dosage form that does not have FDA approval. Hence, the specific use is "off-label" (i.e. not approved by the FDA and not listed in FDA-required labeling information).

Off-label use can be motivated by several factors [67, 89]. For example, a product may be used for a specific population for which it has not been approved. Also, if a medication has been approved to treat a specific condition, medications from the same class of drugs may also be used to treat that condition. Finally, if the features of two medical conditions are similar, a physician may use a medication approved for one of these conditions to treat both. However, many other factors may motivate off-label use as well [67, 89].

Off-label use is quite common in clinical practice; up to one-fifth prescriptions are off-label [89]. There are many reasons why it remains common. For example, adding additional indications for an already approved medication can be costly and time-consuming, and revenues for the new indication may not offset the expense and effort of obtaining approval. Moreover, generic medications may not have the requisite funding foundations needed to pursue additional FDA approvals. Therefore, drug proprietors may never seek FDA approval for common uses.

Although off-label use is not illegal, following off-label use recommendations without adequate medical supervision is not recommended as it may inadvertently lead to harm. Off-label promotion refers to directly promoting a medical product for any indication that has not been listed in the product label, as well as providing information (e.g. usage information) that does not adhere to the FDA-approved labeling document [11, 15, 84]. Without medical supervision or adequate warnings, off-label promotion should be avoided.

2.2 Harms of off-label promotion by LLMs

Social media websites, including online health communities, Twitter, Facebook, and others, as well as scientific articles in academic journals, are potentially the largest source of data related to off-label use of medical products [27]. Because LLMs are trained on massive datasets, they can learn these off-label uses and remain in parametric memory, or alternatively be surfaced via retrieved augmented generation (RAG) [34].

This poses potential dangers to public health. For example, a user may be misled to believe that an off-label use of a prescription drug or medical product is safe or effective, exposing them to the potential adverse side effects of a product that has not been adequately tested for safety and effectiveness in treatment of a particular condition. They may also be recommended treatments that are ineffective, or even nonsensical treatments, or be recommended more expensive, yet inadequately tested products. Given the massive scale at which LLM models operate, this can lead to significant public health risk [1, 84].

2.3 Detecting off-label use with ML

Previous work has focused on applying ML to detecting off-label use in electronic health records [45, 46], online health communities such as MedHelp, WebMD, Drugs.com, and HealthBoards.com [59, 93, 98, 99], and more recently social media sites [27, 40, 55]. Recent work has leveraged transformer-based methodologies (e.g. BERT [36]) to identify these off-label uses. However, to the best of our knowledge, the issue of off-label promotion by LLM models has not been explored.

Moreover, these previous works have focused on one form of off-label use (the use of products to treat unapproved indications) and did not study the detection of off-label use with respect to populations (e.g. age, gender), dosage, contraindications, or any other component of the labeling document information.

2.4 Medical question answering evaluation

Recent works on medical evaluation of LLMs for uses in healthcare can be classified into the following categories, among others [42, 69]: evaluations of knowledge and capability, trustworthiness, transparency and fairness. These evaluations are typically use-specific, such as evaluating LLMs for EHR answering [49, 65, 74, 95] or summarization [77]. In the context of medical question answering, previous works have evaluated the quality of LLMs responses to biomedical and clinical knowledge questions [12, 22, 63, 73], showing remarkable capabilities in a number of medical knowledge benchmarks such as the popular MedQA (USMLE) benchmark [63, 70], which consists of a multi-choice dataset for medical domain

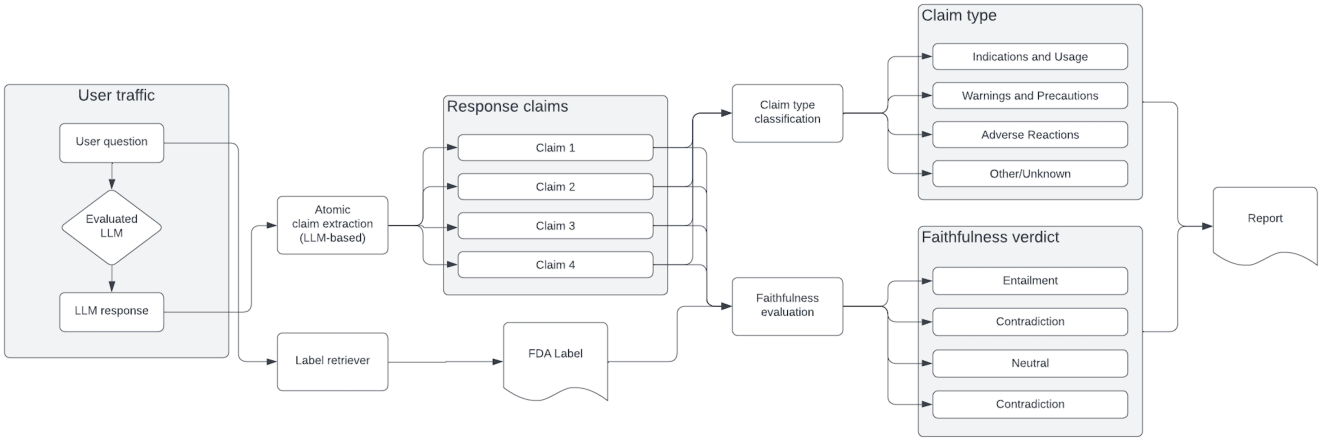

Figure 1: Response evaluation framework.

question answering. However, LLMs have been shows to provide responses that are not supported by the sources they provide [91], raising concerns about both trustworthiness and transparency. In addition to this, previous works evaluating LLMs in medicine have focused on fairness and bias detection [20], revealing race- and gender-based stereotypes [61, 94, 96, 97].

Despite these recent works, to our knowledge, the evaluation of LLMs for medical product question answering and adherence to labeling documents remains unexplored.

3 METHODOLOGY

In this section, we describe the methodology that we applied in this work. This is illustrated in Fig.1, which provides a high-level overview of the framework for LLM response evaluation outlining all key components.

First, in Sec.3.1 we describe the language models used in this work. Sec.3.2 described the dataset used as knowledge source for factual verification. Then, Sec.3.3 describes the methods used to extract atomic claims from LLM responses. Sec.3.4 proceeds to outline an LLM-based multi-class model for classifying the type of claims. Finally, Sec.3.5 describes the approach to evaluate whether a given claim is supported by the FDA labeling document information.

3.1 Models

In this work, we used Anthropic's Claude 3 Sonnet [5–7]. This LLM was released in 2024 and is available via a website (https://claude.ai/) and as an API. While few details are available about the model's development, several aspects of its training and evaluation have been documented in Anthropic's research papers. These include preference modeling [8], reinforcement learning from human feedback [13], constitutional AI [14], red-teaming [33], evaluation with language model-generated tests [66], and self-correction [32], among others.

In addition to this, for the claim type classifier introduced in Sec.3.4, we used a text-to-text encoder-decoder model, Flan-T5 [24], which is an instruction fine-tuned version of T5 [68]. We also considered encoder-only models, specifically BERT [26], DestilBERT [71], and RoBERTa [52].

3.2 FDALabel Database

The FDALabel database [28] is an FDA web-based application[1] used to perform customizable searches of over 147,000 human over-the-counter (OTC) and prescription medical products. It contains up-to-date medical labeling data, including product label images, as well as information about approved indications, active ingredients, usage, dosage, contraindications and side effects, among other information.

We use the FDALabel database as our knowledge source for factual medical product information.

3.3 Claim extraction

LLM responses typically consist of a large number of pieces of information that may be a mixture of defective and non-defective, hence making a binary judgement inadequate. Therefore, following previous work that also sought to evaluate the factuality of LLM responses [58, 88], we first use a *claim extractor* to break the LLM response into atomic claims.

Atomic claims are short sentences containing one piece of information each [58], and are different from normal sentences as the latter may contain multiple pieces of information each.

Our claim extractor first breaks out the LLM response automatically by splitting it into individual sentences. Specifically, we used NLTK's Punkt sentence tokenizer [18, 60] which divides a text into a list of sentences by using an unsupervised algorithm. Following this, as in FactScore [58], SAFE [88] and Fables [47], each sentence is sent to a commercial LLM with a series of instructions to further break it down to a series of atomic facts. In this work, we leveraged Claude 3, described in Sec.3.1, as it has recently shown great performance in extracting claims from long-form text [47].

[1] https://nctr-crs.fda.gov/fdalabel/

3.4 Claim type classification

The goal of the *claim type classifier* is to classify each of the claims in the LLM response, extracted following Sec.3.3, into one of several claim types corresponding to the different sections of the product labels. Specifically, in the USA the FDA requires human prescription drugs and biological product labels to follow the Physician Label Rule (PLR) [30]. The PLR contains a set of requirements for the content and format of the labels. Among other requirements, the PLR dictates that FDA labeling documents must contain the following sections: (1) *Indications and usage*, (2) *Warnings and precautions*, and (3) *Adverse Reactions* [31].

The *Indications and usage* section includes a concise statement of each of the product's indications, briefly noting any major limitations of use [31]. The *Warnings and precautions* section includes a concise summary of the most clinically significant safety concerns that affect decisions about whether to prescribe the drug, recommendations for patient monitoring to ensure safe use of the drug, and measures that can be taken to prevent or mitigate harm [31]. The *Adverse Reactions* section includes a listing of the most frequently occurring adverse reactions and the criteria used to determine inclusion (e.g., frequency cutoff rate) [31].

To evaluate the type of each claim in the LLM response, following prior work on drug labeling text classification [35], we developed a multiclass classification model that assigns each claim to one of the aforementioned key PLR sections. In addition to this, a fourth class was added for claims that do not belong to neither of the three classes: (4) *Other/Unknown*. This forth class may contain claims that are not a good fit for any of aforementioned three classes but may still be present in other sections of the PLR, or claims that are unrelated to the labeling document.

To implement the claim type classifier, we evaluated several modeling architectures, introduced in Sec.3.1. The first one consisted of encoder-only transformers. Specifically, we evaluated BERT [26], DestilBERT [71], and RoBERTa [52]. In addition to this, we evaluated encoder-decoder models. Specifically, we focused on Flan-T5 [24] which transforms the classification task into a text-to-text task, such that the output of the model is the tokens denoting the class assignment. Among the available encoder-decoder LLMs, we chose FLAN-T5 because the quality of its generalized representation of natural language, the possibility of easily adapting the model to a downstream task with little fine-tuning without adjusting its architecture, and its availability in different model size configurations. Specifically, several variants of this LLM are available, ranging from 77M parameters for `flan-t5-small` to 11.3B parameters for `flan-t5-xxl`. This allows us to investigate the tradeoff between model performance and computational load. Finally, we evaluated zero-shot prompting of Claude 3.

3.5 Faithfulness evaluator

The *faithfulness evaluator* evaluates each atomic claim in the LLM response extracted by the claim extractor (Sec.3.3) and determined to belong to classes 1-3 (that is, not class (4) *Unknown/other*) by the *claim type classifier*. Specifically, it evaluates whether the claim is faithful to the corresponding knowledge source, that is, the FDA labeling document. This approach is based on prior work on LLM agents as factuality autoraters that compared model responses to a preset reference answer or knowledge sources, such as FactScore [58], SAFE [88] and Fables [47]. In our work, for the evaluator LLM agent, we used Claude 3 Sonnet, introduced in Sec.3.1. Compared to Flan-T5, which supports up to 512 input tokens, Claude's context window accepts up to 200,000 tokens (roughly 150,000 words, or over 500 pages of material). This enables each atomic claim to be evaluated against the entirety of the FDA labeling document.

3.6 Report generation

The final stage outlined in Fig.1 merges the outputs of the claim type evaluator (Sec.3.4) and compliance evaluator (Sec.3.5) into a single report. If any claim classified to be types 1-3, that is, not "Other/Unknown", is determined to contradict the product label, then the entire response is deemed defective.

4 EXPERIMENTS AND RESULTS

To narrow down the experimentation, we considered a medical product question answering context where a user interacts with an LLM-based AI assistant that helps customers find answers to medical product questions. Specifically, we focused on prescription drugs. However, the methodology described in Sec.3 also applies to other types of medical products, including over-the-counter drugs, as well as both prescription and over-the-counter biologics and medical devices.

4.1 Question and response generation

We implement a template-based method to synthetically generate medical-related user prompts. Specifically, 20 human generated prompt templates were generated. These templates represented questions about indications and usage (10 templates), warnings and precautions (7 templates) and adverse reactions (3 templates), corresponding to the PLR labeling document sections discussed in Sec.3.4. These were questions that a patient with no domain knowledge may ask about a prescription drug, e.g. *"I am considering taking* `{DRUG_NAME}`*. Are there any adverse reactions associated with the use of this medication?"*.

Using this template-based prompt generation method , we generated a total of 2000 synthetic user prompts for a total of 100 human prescription drugs randomly selected from the FDALabel database [28], out of the 57,293 present in the database[2]. Using Claude 3 Sonnet, we generated the corresponding LLM responses.

4.2 Atomic claim extraction

Following Sec.3.3, we extracted claims for each of the 2000 responses we generated for the 100 prescription drugs using the 20 templates. Table 1 shows the statistics of the claim extraction results. The average number of claims per response was 27.69 (σ = 8.00). This was more than the average number of sentences per response, which was 22.23 (σ = 7.05). Each claim contained an average of 10.67 (σ = 1.88) words. In total, we extracted 55,388 claims from the 2,000 responses.

Human validation of a random sample of 100 extracted atomic claims demonstrated 100% precision, that is, each claim can be

[2]As of May 7, 2024.

Question type	# templates	# responses	# sentences / response	# claims / response	# words / claim
Indications and Usage	10	1000	22.57 (4.72)	29.31 (5.10)	10.74 (1.85)
Warnings and Precautions	7	700	16. 89 (3.57)	20.43 (3.98)	10.58 (1.92)
Adverse reactions	3	300	33.56 (5.75)	39.26 (6.15)	10.70 (1.88)
Total	20	2000	22.23 (7.05)	27.69 (8.00)	10.67 (1.88)

Table 1: Statistics of the atomic claim extraction results.

Claim type	Development set	Test set
Indications and Usage	453 (45.3%)	223 (44.6%)
Warnings and Precautions	321 (32.1%)	161 (32.2%)
Adverse reaction	148 (14.8%)	69 (13.8%)
Other/Unknown	78 (7.8%)	47 (9.4%)
Total claims	1000 (100%)	500 (100%)

Table 2: Composition of development and test sets, used to train and evaluate the claim type classifier described in Sec.3.4

Label	Precision	Recall	F1	Support
Supported	0.87	0.93	0.90	418
Not supported	0.95	0.81	0.87	371
Irrelevant	0.68	0.84	0.75	117
Overall	0.88	0.87	0.87	906

Table 4: Performance of Claude 3 in determining whether an atomic claim extracted from an LLM response is supported, not supported or irrelevant based on the corresponding FDA-approved labeling document.

traced to the original LLM response without any extra or incorrect information.

4.3 Claim classification

Model	Precision	Recall	F1 score
BERT	0.69	0.61	0.65
DistilBERT	0.70	0.63	0.66
RoBERTa	0.70	0.61	0.65
Flan-T5-small	0.75	0.69	0.72
Flan-T5-base	0.82	0.75	0.78
Flan-T5-large	0.85	0.77	0.91
Claude 3 Sonnet	0.82	0.74	0.78

Table 3: Claim classification performance showing macro averages of precision, recall and F1 score., for the classification of each claim into the corresponding PLR section.

Using the claim type classifier introduced in Sec.3.4, each of the extracted atomic claims was assigned to one of the following 4 labels: (1) *Indications and usage*, (2) *Warnings and precautions*, (3) *Adverse Reactions*, and (4) *Other/Unknown*.

To train and evaluate the models, we used the data described in Table 2, which was obtained by annotating a random selection of 1,500 atomic claims from the 55,388 claims extracted in Sec.4.2.

The performance on the test set of the different models evaluated in shown in Table 3. The best performance was obtained by Flan-T5-large, which was fine-tuned using the development set. This surpassed the performance of Claude 3 Sonnet, which was not fine-tuned and used zero-shot prompting.

4.4 Claim support evaluation

Finally, we evaluated the performance of Claude 3 in determining whether a claim was supported, not supported, or irrelevant given the corresponding product label from the FDALabel database. Given that we expected most claims in the LLM responses from Sec.4.1 to be factually correct based on the overall performance of Claude 3 in providing high quality responses, we synthetically built an evaluation dataset. To do so, we used the 453 claims not labeled *other/unknown* described in Sec.4.3 and shown in Table 2. Each claim was associated with the corresponding FDA labeling document. We duplicated these claims, associating them with a different labeling document randomly selected from 57,293 prescription drug labels present in the FDALabel database. In total, we had 906 claims and corresponding product labels. These were manually annotated by a human annotator using the following 3 classes: (1) supported, (2) not supported, (3) irrelevant. The distribution of annotations in shown in Table 4.

Using Claude 3 Soonet and zero-shot promoting, each claim was automatically assigned to one of the three aforementioned classes. Statistics of the data and model performance results are reported in Table 4.

5 CONCLUSION AND FUTURE WORK

While LLMs have shown impressive reasoning and question answering capabilities, they can produce false outputs and innacurate answers [29, 92]. Therefore, in this work, we aimed to investigate factuality and adherence to the information provided in the relevant labeling documents in medical product question answering by LLMs.

Using a synthetically generated user question and the FDALabel database, we demonstrated a methodology for response evaluation that breaks down a response into a series of atomic claims. Each claim is then evaluated to determine if it is associated to one of the several PLR sections in the FDA labeling documents. If so, the claim is evaluated against the corresponding labeling document to determine if it is supported, not supported, or irrelevant based on the the information contained in the label. Claims that are not supported are considered to contain off-label information, as the claim cannot be supported by the labeling document alone. The proposed methodology builds upon prior work on factuality evaluation [47, 58, 88], and uses LLMs as evaluators.

REFERENCES

[1] David Adam. 2024. Medical AI could be 'dangerous' for poorer nations, WHO warns. *Nature* (Jan. 2024).

[2] Lisa C Adams, Felix Busch, Daniel Truhn, Marcus R Makowski, Hugo J W L Aerts, and Keno K Bressem. 2023. What Does DALL-E 2 Know About Radiology? *J. Med. Internet Res.* 25 (March 2023), e43110.

[3] Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. (Sept. 2023). arXiv:2311.01463 [cs.CL]

[4] Hazrat Ali, Junaid Qadir, Tanvir Alam, Mowafa Househ, and Zubair Shah. 2023. ChatGPT and Large Language Models in Healthcare: Opportunities and Risks. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*. IEEE, 1–4.

[5] Anthropic. 2023. Claude 2. https://www.anthropic.com/news/claude-2. Accessed: 2024-1-28.

[6] Anthropic. 2023. *Model Card and Evaluations for Claude Models.* Technical Report.

[7] Anthropic. 2024. *The Claude 3 Model Family: Opus, Sonnet, Haiku.* Technical Report.

[8] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. (Dec. 2021). arXiv:2112.00861 [cs.CL]

[9] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. 2023. AI chatbots not yet ready for clinical use. *Front Digit Health* 5 (April 2023), 1161098.

[10] Sören Auer, Dante A C Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge. *Sci. Rep.* 13, 1 (May 2023), 7240.

[11] Richard C Ausness. 2008. There's Danger Here, Cherie!: Liability for the Promotion and Marketing of Drugs and Medical Devices for Off-Label Uses. *Brooklyn Law Review* 73, 4 (2008), 1253–1326.

[12] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern. Med.* 183, 6 (June 2023), 589–596.

[13] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. (April 2022). arXiv:2204.05862 [cs.CL]

[14] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. (Dec. 2022). arXiv:2212.08073 [cs.CL]

[15] James Beck. 2021. Off-Label Use in the Twenty-First Century: Most Myths and Misconceptions Mitigated. *UIC J. Marshall Law Review* 54, 1 (2021).

[16] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R Chaurasia, Nirav R Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A Pfeffer, and Nigam H Shah. 2024. A systematic review of testing and evaluation of healthcare applications of large language models (LLMs). (April 2024).

[17] Trista M Benítez, Yueyuan Xu, J Donald Boudreau, Alfred Wei Chieh Kow, Fernando Bello, Le Van Phuoc, Xiaofei Wang, Xiaodong Sun, Gilberto Ka-Kit Leung, Yanyan Lan, Yaxing Wang, Davy Cheng, Yih-Chung Tham, Tien Yin Wong, and Kevin C Chung. 2024. Harnessing the potential of large language models in medical education: promise and pitfalls. *J. Am. Med. Inform. Assoc.* 31, 3 (Feb. 2024), 776–783.

[18] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* "O'Reilly Media, Inc.".

[19] Marco Cascella, Federico Semeraro, Jonathan Montomoli, Valentina Bellini, Ornella Piazza, and Elena Bignami. 2024. The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. *J. Med. Syst.* 48, 1 (Feb. 2024), 22.

[20] Crystal T Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H Jacobson, Binglan Li, Celeste H Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A Fries, Jiapeng Xu, Joe Jamison, John S Tamaresis, Jonathan H Chen, Joshua Lazaro, Juan M Banda, Julie J Lee, Karen Ebert Matthys, Kirsten R Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T Truong, Scott L Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y Eulalio, Wenyuan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam H Shah, and Roxana Daneshjou. 2024. Red Teaming Large Language Models in Medicine: Real-World Insights on Model Behavior. *medRxiv* (2024).

[21] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. 2019. How to develop machine learning models for healthcare. *Nat. Mater.* 18, 5 (May 2019), 410–414.

[22] Shan Chen, Benjamin H Kann, Michael B Foote, Hugo J W L Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol* 9, 10 (Oct. 2023), 1459–1462.

[23] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. (Nov. 2023). arXiv:2311.16079 [cs.CL]

[24] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. (Oct. 2022). arXiv:2210.11416 [cs.LG]

[25] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, Sophia J Wagner, and Jakob Nikolas Kather. 2023. The future landscape of large language models in medicine. *Commun. Med.* 3, 1 (Oct. 2023), 141.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (Oct. 2018). arXiv:1810.04805 [cs.CL]

[27] Brian Dreyfus, Anuj Chaudhary, Parth Bhardwaj, and V Karthikhaa Shree. 2021. Application of natural language processing techniques to identify off-label drug usage from various online health communities. *J. Am. Med. Inform. Assoc.* 28, 10 (Sept. 2021), 2147–2154.

[28] Hong Fang, Stephen C Harris, Zhichao Liu, Guangxu Zhou, Guoping Zhang, Joshua Xu, Lilliam Rosario, Paul C Howard, and Weida Tong. 2016. FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. *Drug Discov. Today* 21, 10 (Oct. 2016), 1566–1570.

[29] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (June 2024), 625–630.

[30] Food and Drug Administration (FDA). 2005. Providing Regulatory Submissions in Electronic Format — Content of Labeling. (April 2005).

[31] Food and Drug Administration (FDA). 2013. Labeling for Human Prescription Drug and Biological Products – Implementing the PLR Content and Format Requirements. (Feb. 2013).

[32] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jackson Kernion, Jamie Kerr, Jared Mueller, Joshua Landau, Kamal Ndousse, Karina Nguyen, Liane Lovitt, Michael Sellitto, Nelson Elhage, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robert Lasenby, Robin Larson, Sam Ringer, Sandipan Kundu, Saurav Kadavath, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Henighan, Tristan Hume, Yuntao

Bai, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, Christopher Olah, Jack Clark, Samuel R Bowman, and Jared Kaplan. 2023. The Capacity for Moral Self-Correction in Large Language Models. (Feb. 2023). arXiv:2302.07459 [cs.CL]

[33] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. (Aug. 2022). arXiv:2209.07858 [cs.CL]

[34] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for large Language Models: A survey. (Dec. 2023). arXiv:2312.10997 [cs.CL]

[35] Magnus Gray, Joshua Xu, Weida Tong, and Leihong Wu. 2023. Classifying Free Texts Into Predefined Sections Using AI in Regulatory Documents: A Case Study with Drug Labeling Documents. *Chem. Res. Toxicol.* 36, 8 (Aug. 2023), 1290–1299.

[36] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (March 2022). arXiv:2203.05794 [cs.CL]

[37] Stefan Harrer. 2023. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* 90 (April 2023), 104512.

[38] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, and Dinggang Shen. 2023. Transformers in medical image analysis. *Intelligent Medicine* 3, 1 (Feb. 2023), 59–78.

[39] Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. 2023. Diagnostic Accuracy of Differential-Diagnosis Lists Generated by Generative Pretrained Transformer 3 Chatbot for Clinical Vignettes with Common Chief Complaints: A Pilot Study. *Int. J. Environ. Res. Public Health* 20, 4 (Feb. 2023).

[40] Yining Hua, Hang Jiang, Shixu Lin, Jie Yang, Joseph M Plasek, David W Bates, and Li Zhou. 2022. Using Twitter data to understand public perceptions of approved versus off-label use for COVID-19-related medications. *J. Am. Med. Inform. Assoc.* 29, 10 (Sept. 2022), 1668–1678.

[41] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. (Nov. 2023). arXiv:2311.05232 [cs.CL]

[42] Yining Huang, Keke Tang, and Meilian Chen. 2024. A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Industry. (April 2024). arXiv:2404.15777 [cs.CL]

[43] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A New Benchmark for Financial Question Answering. (Nov. 2023). arXiv:2311.11944 [cs.CL]

[44] Jenelle A Jindal, Matthew P Lungren, and Nigam H Shah. 2024. Ensuring useful adoption of generative artificial intelligence in healthcare. *J. Am. Med. Inform. Assoc.* (March 2024).

[45] Kenneth Jung, Paea LePendu, William S Chen, Srinivasan V Iyer, Ben Readhead, Joel T Dudley, and Nigam H Shah. 2014. Automated detection of off-label drug use. *PLoS One* 9, 2 (Feb. 2014), e89324.

[46] Kenneth Jung, Paea Lependu, and Nigam Shah. 2013. Automated Detection of Systematic Off-label Drug Use in Free Text of Electronic Medical Records. *AMIA Jt Summits Transl Sci Proc* 2013 (March 2013), 94–98.

[47] Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization. (April 2024). arXiv:2404.01261 [cs.CL]

[48] Joan H Krause. 2015. Off-label drug promotion and the ephemeral line between marketing and education. *J Law Biosci* 2, 3 (Nov. 2015), 705–711.

[49] Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Hyun Kim, Seunghyun Won, and Edward Choi. 2024. EHRNoteQA: A patient-specific question answering benchmark for evaluating Large Language Models in clinical settings.

[50] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (Aug. 2019), 453–466.

[51] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for Large Language Models: A Comprehensive Survey. (Feb. 2024). arXiv:2402.18041 [cs.CL]

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. (July 2019). arXiv:1907.11692 [cs.CL]

[53] Dechert LLPa. 2021. FDA Issues Final Rule Clarifying Evidence of Off-Label Marketing. https://www.jdsupra.com/legalnews/fda-issues-final-rule-clarifying-3589840/. Accessed: 2024-2-8.

[54] Harrison C Lucas, Jeffrey S Upperman, and Jamie R Robinson. 2024. A systematic review of large language models and their implications in medical education. *Med. Educ.* (April 2024).

[55] Tim Ken Mackey, Jiawei Li, Vidya Purushothaman, Matthew Nali, Neal Shah, Cortni Bardier, Mingxiang Cai, and Bryan Liang. 2020. Big Data, Natural Language Processing, and Deep Learning to Detect and Characterize Illicit COVID-19 Product Sales: Infoveillance Study on Twitter and Instagram. *JMIR Public Health Surveill* 6, 3 (Aug. 2020), e20794.

[56] Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang, Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia-Ming Ji, Zifeng Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, Zexuan Guo, Youlan Lei, Chunli Shao, Wenyao Wang, Haojun Fan, and Yi-Da Tang. 2024. The application of large language models in medicine: A scoping review. *iScience* 27, 5 (May 2024), 109713.

[57] Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 6, 1 (July 2023), 120.

[58] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-Tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. (May 2023). arXiv:2305.14251 [cs.CL]

[59] Azadeh Nikfarjam, Julia D Ransohoff, Alison Callahan, Vladimir Polony, and Nigam H Shah. 2019. Profiling off-label prescriptions in cancer treatment using social health networks. *JAMIA Open* 2, 3 (Oct. 2019), 301–305.

[60] NLTK Project. 2023. nltk.tokenize.punkt module. https://www.nltk.org/api/nltk.tokenize.punkt.html. Accessed: 2024-6-13.

[61] Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digit Med* 6, 1 (Oct. 2023), 195.

[62] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler,

Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C J Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. (March 2023). arXiv:2303.08774 [cs.CL]

[63] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. (March 2022). arXiv:2203.14371 [cs.CL]

[64] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical Domain Hallucination Test for Large Language Models. (July 2023). arXiv:2307.15343 [cs.CL]

[65] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. (Sept. 2018). arXiv:1809.00732 [cs.CL]

[66] Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. (Dec. 2022). arXiv:2212.09251 [cs.CL]

[67] David C Radley, Stan N Finkelstein, and Randall S Stafford. 2006. Off-label prescribing among office-based physicians. *Arch. Intern. Med.* 166, 9 (May 2006), 1021–1026.

[68] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. (Oct. 2019). arXiv:1910.10683 [cs.LG]

[69] Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* 41 (Jan. 2023), 101304.

[70] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G T Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-Baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Siwai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S M Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of Gemini Models in Medicine. (April 2024). arXiv:2404.18416 [cs.AI]

[71] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (Oct. 2019). arXiv:1910.01108 [cs.CL]

[72] Nigam H Shah, David Entwistle, and Michael A Pfeffer. 2023. Creation and Adoption of Large Language Models in Medicine. *JAMA* 330, 9 (Sept. 2023), 866–869.

[73] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera Y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (Aug. 2023), 172–180.

[74] Sarvesh Soni, Meghana Gudala, A Pajouhi, and Kirk Roberts. 2022. RadQA: A question answering dataset to improve comprehension of radiology reports. *LREC* (2022), 6250–6259.

[75] Randall S Stafford. 2008. Regulating off-label drug use–rethinking the role of the FDA. *N. Engl. J. Med.* 358, 14 (April 2008), 1427–1429.

[76] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. TrustLLM: Trustworthiness in Large Language Models. (Jan. 2024). arXiv:2401.05561 [cs.CL]

[77] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating Large Language Models on Medical Evidence Summarization. *medRxiv* (April 2023).

[78] Pasin Tangadulrat, Supinya Sono, and Boonsin Tangtrakulwanich. 2023. Using ChatGPT for Clinical Practice and Medical Education: Cross-Sectional Survey of Medical Students' and Physicians' Perceptions. *JMIR Med Educ* 9 (Dec. 2023), e50658.

[79] Mohamad-Hani Temsah, Fadi Aljamaan, Khalid H Malki, Khalid Alhasan, Ibraheem Altamimi, Razan Aljarbou, Faisal Bazuhair, Abdulmajeed Alsubaihin, Naif Abdulmajeed, Fatimah S Alshahrani, Reem Temsah, Turki Alshahrani, Lama Al-Eyadhy, Serin Mohammed Alkhateeb, Basema Saddik, Rabih Halwani, Amr Jamal, Jaffar A Al-Tawfiq, and Ayman Al-Eyadhy. 2023. ChatGPT and the Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. *Healthcare (Basel)* 11, 13 (June 2023).

[80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *ArXiv* abs/2302.13971 (Feb. 2023).

[81] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. (July 2023). arXiv:2307.09288 [cs.CL]

[82] Tu Tao, Azizi Shekoofeh, Driess Danny, Schaekermann Mike, Amin Mohamed, Chang Pi-Chuan, Carroll Andrew, Lau Charles, Tanno Ryutaro, Ktena Ira, Palepu Anil, Mustafa Basil, Chowdhery Aakanksha, Liu Yun, Kornblith Simon, Fleet David, Mansfield Philip, Prakash Sushant, Wong Renee, Virmani Sunny, Semturs Christopher, Mahdavi S. Sara, Green Bradley, Dominowska Ewa, Arcas Blaise Aguera y, Barral Joelle, Webster Dale, Corrado Greg S., Matias Yossi, Singhal Karan, Florence Pete, Karthikesalingam Alan, and Natarajan Vivek. 2024. Towards Generalist Biomedical AI. *NEJM AI* 1, 3 (Feb. 2024), AIoa2300138.

[83] Gail A Van Norman. 2016. Drugs, Devices, and the FDA: Part 1: An Overview of Approval Processes for Drugs. *JACC Basic Transl Sci* 1, 3 (April 2016), 170–179.

[84] Gail A Van Norman. 2023. Off-Label Use vs Off-Label Marketing: Part 2: Off-Label Marketing—Consequences for Patients, Clinicians, and Researchers. *JACC: Basic to Translational Science* 8, 3 (March 2023), 359–370.

[85] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* (Feb. 2024).

[86] Xidong Wang, Nuo Chen, Junying Chen, Yan Hu, Yidong Wang Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024. Apollo: An Lightweight Multilingual Medical LLM towards Democratizing Medical AI to 6B People. (March 2024). arXiv:2403.03640

[87] Zhen Wang. 2022. Modern Question Answering Datasets and Benchmarks: A Survey. (June 2022). arXiv:2206.15030 [cs.CL]
[88] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. (March 2024). arXiv:2403.18802 [cs.CL]
[89] Christopher M Wittich, Christopher M Burkle, and William L Lanier. 2012. Ten common questions (and their answers) about off-label drug use. *Mayo Clin. Proc.* 87, 10 (Oct. 2012), 982–990.
[90] World Health Organization. 2024. *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models*. World Health Organization.
[91] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. (Feb. 2024). arXiv:2402.02008 [cs.CL]
[92] Yijun Xiao and William Yang Wang. 2021. On Hallucination and Predictive Uncertainty in Conditional Language Generation. (March 2021). arXiv:2103.15025 [cs.CL]
[93] Christopher C Yang and Mengnan Zhao. 2017. Determining Associations with Word Embedding in Heterogeneous Network for Detecting Off-Label Drug Uses. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 496–501.
[94] Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and Quantifying Racial Bias of Large Language Models in Medical Report Generation. *ArXiv* (Jan. 2024).
[95] Xiang Yue, Xinliang Frederick Zhang, Ziyu Yao, Simon Lin, and Huan Sun. 2020. CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering. (Oct. 2020). arXiv:2010.16021 [cs.CL]
[96] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, and Emily Alsentzer. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 6, 1 (Jan. 2024), e12–e22.
[97] Angela Zhang, Mert Yüksekgönül, Joshua Guild, James Y Zou, and Joseph C Wu. 2023. ChatGPT exhibits gender and racial biases in acute coronary syndrome management. *ArXiv* abs/2311.14703 (Nov. 2023).
[98] Mengnan Zhao and Christopher C Yang. 2017. Automated Off-label Drug Use Detection from User Generated Content. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics* (Boston, Massachusetts, USA) *(ACM-BCB '17)*. Association for Computing Machinery, New York, NY, USA, 449–454.
[99] Mengnan Zhao and Christopher C Yang. 2018. Exploiting OHC Data with Tensor Decomposition for Off-Label Drug Use Detection. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 22–28.
[100] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A Clifton. 2023. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. (Nov. 2023). arXiv:2311.05112 [cs.CL]

Received 10 June 2024