Guardrails for avoiding harmful medical product recommendations and off-label promotion in generative AI models

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

Abstract

Generative AI (GenAI) models have demonstrated remarkable capabilities in a wide variety of medical tasks. However, as these models are trained using generalist datasets with very limited human oversight, they can learn uses of medical products that have not been adequately evaluated for safety and efficacy, nor approved by regulatory agencies. Given the scale at which GenAI may reach users, unvetted recommendations pose a public health risk. In this work, we propose an approach to identify potentially harmful product recommendations, and demonstrate it using a recent multimodal large language model.

1. Introduction

Rapid advancements in generative artificial intelligence (GenAI) have led to the development of sophisticated models such as DALL-E [18], GPT-4 [53], PaLM2 [7], Llama 2 [63] and Claude 3 [8, 10]. These have the capability to model complex relationships from massive multimodal datasets, generate new content and perform tasks they were never trained for. Their potential applications in medicine include medical image analysis [3, 33], providing differential diagnoses [34], summarizing charts [67], writing letters to patients [5], providing medical education [6, 41, 61], aiding pharmacy providers (e.g. prescription generation, safety evaluation, decision support) [6, 42, 45, 52], or working as a chatbot to answer questions about patients' specific concerns or medical products [6], among others. Medicine is inherently a multimodal discipline, and new GenAI models such as multimodal large language models (MLLMs) continue to be developed that integrate diverse multimodal data streams [64]. Moreover, general-purpose GenAI models are often used for medical-related tasks, despite not being specifically developed for those.

While these technologies hold great promise, they also present numerous ethical and legal challenges, can pose significant risk to public health, and cause harm to individuals and organizations [2, 4, 13, 32, 36, 48, 69]. Therefore, it is paramount that GenAI models comply with legal or regulatory regimes. Given that existing laws and regulations written to govern the use of AI often struggle to address the amplified challenges associated with GenAI [48, 50], new ones are rapidly being developed [48, 69]. Meanwhile, there is an imperative for model developers to adhere to the existing frameworks and the trust and safety principles that guided them, to mitigate potential harm and maintain public trust in these breakthrough technologies.

In this work, we shed light into an overlooked issue that impacts most GenAI models, that is, the potential to promote unapproved and potentially harmful uses of medical products. Traditionally, specialized ML models have been trained to address a specific task using highly domain and problem-specific training data [20]. However, GenAI models are typically not developed to do particular medical tasks. Moreover, GenAI models are trained on much more broadly available generalist datasets [44] 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 is unsafe and illegal, and ought to be avoided.

To avoid this issue, we propose a method to identify instances of off-label promotion in GenAI outputs (Sec.3). We demonstrate it using a recently introduced MLLM (Sec.4), and surface examples where it is producing potentially harmful responses (Sec.5). Finally, we briefly discuss how guardrails may be introduced so that GenAI models strictly adhere to product labels when producing outputs related to medical products (Sec.6).

2. Background

Here we discuss the regulatory framework governing the promotion of medical products in the USA (Sec.2.1), the

specific concerns impacting GenAI (Sec.2.2), and previous work on detecting off-label promotion (Sec.2.3)

2.1. Regulatory framework

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 [65]. 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. Hence, the specific use is "off-label" (i.e. not approved by the FDA and not listed in FDA-required labeling information). This term can also apply to 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.

Off-label use can be motivated by several factors [57, 68]. 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.

Off-label use is quite common in clinical practice; up to one-fifth prescriptions are off-label [68]. 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 FDA approval. Therefore, drug proprietors may never seek FDA approval for common uses.

Although off-label use is not illegal, off-label marketing is prohibited. Off-label marketing refers to directly promoting or advertising a medical product for any indication that the FDA has not approved. In fact, this is considered to be fraud and is punishable under the False Claims Act (FCA) [14, 17, 66].

2.2. Harms of off-label promotion by GenAI

Social media websites, including online health communities, Twitter, Facebook, Amazon, 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 [23]. 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) [28].

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 GenAI models operate, this can lead to significant public health risk and potential penalties [2, 66].

From a regulatory perspective, it is not clear what technical category GenAI will fall into, nor what regulations they will be subjected to. For example, the FDA does not categorically prohibit discussing off-label uses, making a nuanced distinction between communication and promotion. Moreover, based on the differences between GenAI and prior ML methods, new regulatory frameworks may be developed to address these GenAI-specific challenges and risks [48]. Regardless of this, off-label promotion poses potential dangers to public health that ought to be minimized.

2.3. Detecting off-label use with ML

Previous work has focused on applying ML to detecting off-label use in electronic health records [37, 38], online health communities such as MedHelp, WebMD, Drugs.com, and HealthBoards.com [51, 71, 73, 74], and more recently social media sites [23, 35, 46]. Recent work has leveraged transformer-based methodologies (e.g. BERT [31]) to identify these off-label uses. However, to the best of our knowledge, the issue of off-label promotion by GenAI models has not been explored.

3. Methods

Fig.2 contains a diagram outlining the overall approach to off-label promotion detection. It focuses on evaluating the output of a MLLM that consumes image and text and produces text. In this work, we specifically consider the use case where the user provides the MLLM model with an image of a product label, and asks a question about it. The proposed method evaluates the model response, taking into account the user query, to detect instances of off-label promotion.

The evaluation algorithm consists of the following 4 sequential steps: (1) input standardization, (2) named entity recognition, (3) product and indication recognition, and (4) off-label identification.

3.1. Input standardization

This step aims to standardize the language of the user query by correcting irregular spellings and orthographic errors. Specifically, we used a context-sensitive spelling correction model for clinical text [40]. Note that abbreviations may

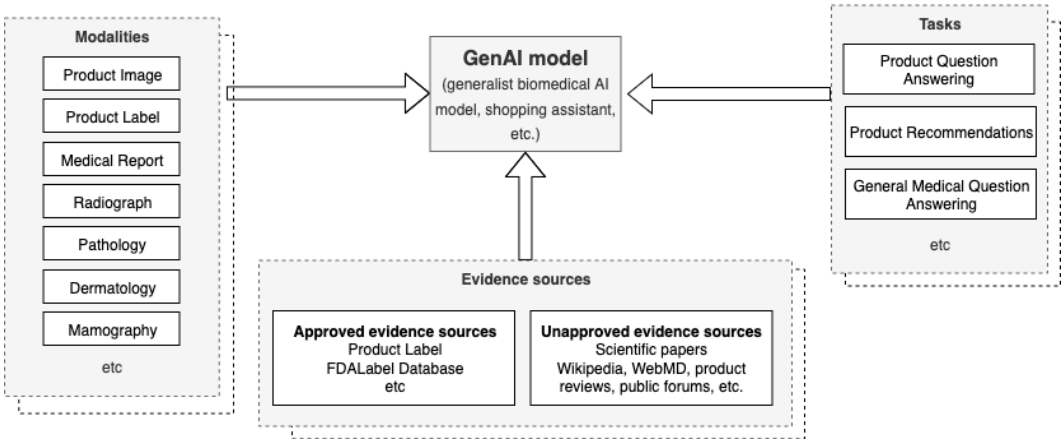

Figure 1. Overview of a generative artificial intelligent system for medical product question answering, product recommendation, and general medical question answering. Such system may handle a handle a diverse range of biomedical data modalities, and use a number of evidence sources, some of which are not approved by the FDA.

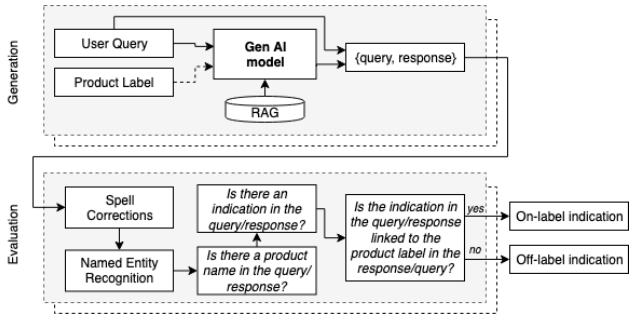

Figure 2. Process flow for GenAI model generation (top) and evaluation of the responses for off-label identification (bottom).

be present in the input query, and although not done in this work, these can also be standardized [59].

3.2. Named entity recognition

This second step identifies medical product names and indications (i.e. diseases, conditions) in the GenAI model responses. To do so, we leverage existing named entity recognition (NER) methods for biomedical terms. There exist a large number of such methods [24]. In this work, we specifically used two BioBERT-based [43] biomedical representation language models fine-tuned to perform NER for drug names and diseases respectively. We found that these BERT models perform significantly better than previous rule- and dictionary-based approaches such as cTAKES [62].

3.3. Product and indication recognition

Products and indications identified in the previous step (Sec.3.2) are matched to those in the FDALabel database [25]. This web-based application[1] was developed by the FDA and allows access to the most up-to-date medical labeling data for over 147,000 human prescription and over-the-counter (OTC) drugs and devices. It contains images of the product labels as well as information about approved indications, active ingredients, usage, dosage, contraindications, side effects, etc.

This matching was done by computing word embeddings and using the cosine similarity to match the NER entities to those in FDALabel. We specifically used an embedding developed for medical concepts, BioBERT [43] , but other embeddings may be used [19, 21, 29, 39, 60, 72].

3.4. Off-label identification

The final step identifies any product-indication association between the user query and the GenAI response that is not FDA-approved. For each product identified in the query, we extract the list of FDA-approved indications from the FDALabel dataset. If any disease not listed in the list of FDA-approved indications is identified in the GenAI response (following Sec.3.3), we zero-shot prompt a T5-large model [58] to determine if the association entails a recommendation. If so, we conclude that it constitutes an instance of off-label promotion.

4. Experimental Setup

To narrow down the experimentation, we considered a shopping context where a user interacts with a MLLM model that helps customers find answers to product questions (see Fig.1). The user provides the model with a picture of a product label, and asks a question about it. The model responds with a textual output. Additionally, but not considered here,

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

Product Name	FDA-approved indications	Off-label indications
Lorazepam	Anxiety, status epilepticus, preanesthetic	Insomnia, panic disorder, delirium
Prazosin	Hypertension	PTSD nightmares, prostatic hypertrophy, Raynaud phenomenon
Quetiapine	Schizophrenia, bipolar disorder depression	Anxiety, insomnia
Donepezil	Dementia of the Alzheimer's type	Lewy body dementia, vascular dementia
Citalopram	Depression	Obsessive-compulsive disorder, panic disorder, premature ejaculation
Sildenafil	Erectile dysfunction, pulmonary hypertension	Female sexual arousal disorder, altitude-induced hypoxemia

Table 1. Examples of off-label indications identified in model responses.

the model may also provide images of recommended products (as done in [47]).

4.1. Model

For the GenAI model depicted in Fig.1, we used Anthropic's Claude 3 Sonnet [8–10]. This MLLM 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 [11], reinforcement learning from human feedback [15], constitutional AI [16], red-teaming [26], evaluation with language model-generated tests [55], and self-correction [27], among others.

4.2. Synthetic user query generation

A common approach for evaluating LLMs is through human testers that probe the system to discover failures [12, 30, 70]. However, these are manual, time consuming, costly and tedious processes that are limited in their ability to adversarialy test GenAI models. Synthetic data generation presents a better alternative that enables generating synthetic user queries at scale and amplifiesd the ability to uncover model defects [54–56].

In this work, we implement a red teaming approach based on 100 human-generated templates that are then populated to generate a large number of synthetic queries. These templates specifically populated using indications from the FDALabel database [25] or disease names form ICD-10 [49].

Given the large number of products in the FDALabel database ($\mathcal{O}(100\text{k})$) and disease names in ICD-10 ($\mathcal{O}(10\text{k})$), to make the analysis tractable, we focused our generation on medical products with known off-label uses. Specifically, we manually generated a list of 35 medical products with a total of 143 known off-label uses, leading to $100 \times 143 = 14300$ synthetic queries about these uses. For each product, we downloaded a copy of its label from the FDA site [25], which was provided to the MLLM together with the query.

In addition to this, to enable the identification method described in Sec.3.3 which requires a product name, these labels were processed by running them through an optical character recognition software for product labels [1].

5. Experimental Results

A total of 14300 synthetic customer queries were generated for 35 pharmaceuticals. After the model responses were processed using our proposed off-label detection method, a total of 15.4% responses were identified to contain off-label indications. We were able to identify off-label indications for 33 products out of the 35 products considered in this work. A small example of off-label indications observed for our selected product list is shown in Table 1.

Using human annotations on a 2000 random sample, we evaluated the performance of our off-label detection method, and concluded it achieved a precision, recall and F1 score of 85.75%, 80.47% and 83.02% respectively.

6. Conclusion

The primary objective of this study was to investigate a key shortcoming of generalist GenAI in medical uses, that is, the off-label promotion of medical products, and highlight the importance by model developers to adhere to existing regulations. Using Claude 3 as an example, we demonstrated that models trained on a vast corpus of internet data with limited filtering can learn unvetted product-indication uses, and consequently promote products for uses for which safety and efficacy has not been adequately evaluated.

In addition to this, we demonstrated a proof-of-principle method for the detection of off-label medical product promotion in MLLM responses. Using our algorithm, we identified instances of off-label promotion for a selection of 35 pharmaceuticals. This method may be used to introduce post-hoc guardrails that monitor and filter the MLLM responses before presenting them to the user, and adapt them to make them harmless [22, 54].

Limitations and Future Work. This is a proof-of-concept work that aimed to highlight potential GenAI harms and regulatory breaches. While we have relied on Claude 3, we have observed similar behavior in other MLLMs, and a more comprehensive evaluation will be needed before conclusions can be made about the prevalence of off-label recommendations. Also, our work focused on one form of off-

label use (the use of products to treat unapproved indications) and did not detect off-label use with respect to age, gender, dosage and contraindications.

References

[1] Tesseract-opencv-OCR-for-product-labels. https://github.com/a1xg/OCR-pipeline-for-product-labels, 2021. Accessed: 2024-3-7. 4

[2] David Adam. Medical AI could be 'dangerous' for poorer nations, WHO warns. *Nature*, 2024. 1, 2

[3] Lisa C Adams, Felix Busch, Daniel Truhn, Marcus R Makowski, Hugo J W L Aerts, and Keno K Bressem. What does DALL-E 2 know about radiology? *J. Med. Internet Res.*, 25:e43110, 2023. 1

[4] Hazrat Ali, Junaid Qadir, Tanvir Alam, Mowafa Househ, and Zubair Shah. ChatGPT and large language models in healthcare: Opportunities and risks. In *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, pages 1–4. IEEE, 2023. 1

[5] Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, and Iain S Whitaker. Using ChatGPT to write patient clinic letters. *Lancet Digit Health*, 5(4):e179–e181, 2023. 1

[6] Mirana Angel, Haiyi Xing, Anuj Patel, Amal Alachkar, and Pierre Baldi. Performance of large language models on pharmacy exam: A comparative assessment using the NAPLEX. 2023. 1

[7] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, Yaguang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 technical report. 2023. 1

[8] Anthropic. Model card and evaluations for claude models. Technical report, 2023. 1, 4

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

[10] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, 2024. 1, 4

[11] 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. A general language assistant as a laboratory for alignment. 2021. 4

[12] Joshua Attenberg, Panos Ipeirotis, and Foster Provost. Beat the machine. *ACM J. Data Inf. Qual.*, 6(1):1–17, 2015. 4

[13] Joshua Au Yeung, Zeljko Kraljevic, Akish Luintel, Alfred Balston, Esther Idowu, Richard J Dobson, and James T Teo. AI chatbots not yet ready for clinical use. *Front Digit Health*, 5:1161098, 2023. 1

[14] Richard C Ausness. 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):1253–1326, 2008. 2

[15] 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. Training a helpful and harmless assistant with reinforcement learning from human feedback. 2022. 4

[16] 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. Constitutional AI: Harmlessness from AI feedback. 2022. 4

[17] James Beck. Off-Label use in the Twenty-First century: Most myths and misconceptions mitigated. *UIC J. Marshall Law Review*, 54(1), 2021. 2

[18] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023. 1

[19] Xiangrui Cai, Jinyang Gao, K Ngiam, B Ooi, Ying Zhang, and Xiaojie Yuan. Medical concept embedding with time-aware attention. *Int Jt Conf Artif Intell*, pages 3984–3990, 2018. 3

[20] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nat. Mater.*, 18(5):410–414, 2019. 1

[21] E Choi, Cao Xiao, W Stewart, and Jimeng Sun. MiME: Multilevel medical embedding of electronic health records for predictive healthcare. *Adv. Neural Inf. Process. Syst.*, abs/1810.09593, 2018. 3

[22] Yi Dong, Ronghui Mu, Gaojie Jin, Yi Qi, Jinwei Hu, Xingyu Zhao, Jie Meng, Wenjie Ruan, and Xiaowei Huang. Building guardrails for large language models. 2024. 4

[23] Brian Dreyfus, Anuj Chaudhary, Parth Bhardwaj, and V Karthikhaa Shree. Application of natural language processing techniques to identify off-label drug usage from various online health communities. *J. Am. Med. Inform. Assoc.*, 28(10):2147–2154, 2021. 2

[24] María C Durango, Ever A Torres-Silva, and Andrés Orozco-Duque. Named entity recognition in electronic health records: A methodological review. *Healthc. Inform. Res.*, 29(4):286–300, 2023. 3

[25] Hong Fang, Stephen C Harris, Zhichao Liu, Guangxu Zhou, Guoping Zhang, Joshua Xu, Lilliam Rosario, Paul C Howard, and Weida Tong. FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. *Drug Discov. Today*, 21(10):1566–1570, 2016. 3, 4

[26] 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. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. 2022. 4

[27] 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. The capacity for moral Self-Correction in large language models. 2023. 4

[28] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-Augmented generation for large language models: A survey. 2023. 2

[29] Emily Getzen, Yucheng Ruan, Lyle Ungar, and Qi Long. Mining for health: A comparison of word embedding methods for analysis of EHRs data. 2022. 3

[30] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improv-

ing alignment of dialogue agents via targeted human judgements. 2022. 4

[31] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. 2022. 2

[32] Stefan Harrer. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine*, 90: 104512, 2023. 1

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

[34] Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. 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), 2023. 1

[35] Yining Hua, Hang Jiang, Shixu Lin, Jie Yang, Joseph M Plasek, David W Bates, and Li Zhou. 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): 1668–1678, 2022. 2

[36] Jenelle A Jindal, Matthew P Lungren, and Nigam H Shah. Ensuring useful adoption of generative artificial intelligence in healthcare. *J. Am. Med. Inform. Assoc.*, 2024. 1

[37] Kenneth Jung, Paea Lependu, and Nigam Shah. Automated detection of systematic off-label drug use in free text of electronic medical records. *AMIA Jt Summits Transl Sci Proc*, 2013:94–98, 2013. 2

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

[39] Faiza Khan Khattak, Serena Jeblee, Chloé Pou-Prom, Mohamed Abdalla, Christopher Meaney, and Frank Rudzicz. A survey of word embeddings for clinical text. *J. Biomed. Inform.*, 100S:100057, 2019. 3

[40] Juyong Kim, Jeremy C Weiss, and Pradeep Ravikumar. Context-sensitive spelling correction of clinical text via conditional independence. *CHIL*, pages 234–247, 2022. 2

[41] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, 2023. 1

[42] Yuki Kunitsu. The potential of GPT-4 as a support tool for pharmacists: Analytical study using the japanese national examination for pharmacists. *JMIR Med Educ*, 9:e48452, 2023. 1

[43] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 3

[44] Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. Datasets for large language models: A comprehensive survey. 2024. 1

[45] Zhengliang Liu, Zihao Wu, Mengxuan Hu, Bokai Zhao, Lin Zhao, Tianyi Zhang, Haixing Dai, Xianyan Chen, Ye Shen, Sheng Li, Brian Murray, Tianming Liu, and Andrea Sikora. PharmacyGPT: The AI pharmacist. 2023. 1

[46] Tim Ken Mackey, Jiawei Li, Vidya Purushothaman, Matthew Nali, Neal Shah, Cortni Bardier, Mingxiang Cai, and Bryan Liang. 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):e20794, 2020. 2

[47] Rajiv Mehta. Amazon announces rufus, a new generative AI-powered conversational shopping experience. https://www.aboutamazon.com/news/retail/amazon-rufus, 2024. Accessed: 2024-2-1. 4

[48] Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med*, 6(1):120, 2023. 1, 2

[49] Harris Meyer. Coding complexity: US health care gets ready for the coming of ICD-10. *Health Aff.*, 30(5): 968–974, 2011. 4

[50] Timo Minssen, Effy Vayena, and I Glenn Cohen. The challenges for regulating medical use of ChatGPT and other large language models. *JAMA*, 330(4):315–316, 2023. 1

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

[52] J Ong, Liyuan Jin, K Elangovan, Gilbert Yong San Lim, D Lim, G Sng, Yuhe Ke, Joshua Yi Min Tung, Ryan Jian Zhong, Christopher Ming Yao Koh, Keane Zhi Hao Lee, Xiang Chen, J Chng, A Than, Ken Junyang Goh, and Daniel Shu Wei Ting. Development and testing of a novel large language model-based clinical decision support systems for medication safety in 12 clinical specialties. *ArXiv*, abs/2402.01741, 2024. 1

[53] 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. GPT-4 technical report. 2023. 1

[54] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. 2022. 4

[55] 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. Discovering language model behaviors with Model-Written evaluations. 2022. 4

[56] Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. AART: AI-Assisted Red-Teaming with diverse data generation for new LLM-powered applications. 2023. 4

[57] David C Radley, Stan N Finkelstein, and Randall S Stafford. Off-label prescribing among office-based physicians. *Arch. Intern. Med.*, 166(9):1021–1026, 2006. 2

[58] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019. 3

[59] Alvin Rajkomar, Eric Loreaux, Yuchen Liu, Jonas Kemp, Benny Li, Ming-Jun Chen, Yi Zhang, Afroz Mohiuddin, and Juraj Gottweis. Deciphering clinical abbreviations with a privacy protecting machine learning system. *Nat. Commun.*, 13(1):7456, 2022. 3

[60] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digit Med*, 4(1): 86, 2021. 3

[61] Malik Sallam, Nesreen A Salim, Muna Barakat, and Ala'a B Al-Tammemi. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. *Narra J*, 3(1):e103, 2023. 1

[62] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.*, 17(5):507–513, 2010. 3

[63] 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. Llama 2: Open foundation and Fine-Tuned chat models. 2023. 1

[64] 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. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024. 1

[65] Gail A Van Norman. Drugs, devices, and the FDA: Part 1: An overview of approval processes for drugs. *JACC Basic Transl Sci*, 1(3):170–179, 2016. 2

[66] Gail A Van Norman. 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):359–370, 2023. 2

[67] 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. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.*, 2024. 1

[68] Christopher M Wittich, Christopher M Burkle, and William L Lanier. Ten common questions (and their answers) about off-label drug use. *Mayo Clin. Proc.*, 87(10):982–990, 2012. 2

[69] World Health Organization. *Ethics and governance of artificial intelligence for health: guidance on large multi-modal models*. World Health Organization, 2024. 1

[70] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Recipes for safety in open-domain chatbots. 2020. 4

[71] Christopher C Yang and Mengnan Zhao. Determining associations with word embedding in heterogeneous network for detecting Off-Label drug uses. In *2017*

IEEE International Conference on Healthcare Informatics (ICHI), pages 496–501. IEEE, 2017. 2

[72] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Sci Data*, 6(1):52, 2019. 3

[73] Mengnan Zhao and Christopher C Yang. 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*, pages 449–454, New York, NY, USA, 2017. Association for Computing Machinery. 2

[74] Mengnan Zhao and Christopher C Yang. Exploiting OHC data with tensor decomposition for Off-Label drug use detection. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 22–28. IEEE, 2018. 2