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 pharmaceuticals, 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 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 application¹ 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,
### Table 1. Examples of off-label indications identified in model responses.

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

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 adversarially test GenAI models. Synthetic data generation presents a better alternative that enables generating synthetic user queries at scale and amplifies 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 ($\mathcal{O}(100k)$) and disease names in ICD-10 ($\mathcal{O}(10k)$), 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


[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu,


[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


[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


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


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


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


[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
