Evaluating Large Language Models: Stigma and Opioid Use Disorder

Shravika Mittal
smittal87@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Mai ElSherief
m.elsherif@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

Tanushree Mitra
tmitra@uw.edu
University of Washington
Seattle, Washington, USA

Munmun De Choudhury
munmund@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

Drug and opioid overdose continues to be a leading cause of death in the United States. Public, social, intervention, and provider-based stigma surrounding Opioid Use Disorder (OUD) or medications for addiction treatment (MAT) cause individuals to turn to online resources or communities to seek knowledge and support. However, these online platforms are known to promote harmful myths related to OUD and MAT. Additionally, with the democratized access to LLM-powered chatbots (e.g., ChatGPT), people are increasingly using them to answer their day-to-day queries. This widespread usage calls for a comprehensive evaluation of LLMs, i.e., to see if and how LLMs propagate myths, stigma, or misconceptions on OUD. In this position paper, we present a list of auditing approaches, open questions, and challenges to initiate a discussion on evaluating LLMs for OUD.

KEYWORDS

Opioid Use Disorder, medication-assisted treatment, myths/stigma, LLM evaluation

1 INTRODUCTION

In 2021, the National Institute for Drug Abuse (NIDA) reported more than 106,000 opioid drug-involved deaths in the United States [21]. For individuals suffering from Opioid Use Disorder (OUD), suggested pathways to recovery include Food and Drug Administration (FDA) approved medications for addiction treatment (MAT) [20], combined with peer support, and counseling/behavioral therapies [30]. However, misperceptions surrounding OUD and MAT pose as barriers to recovery. Stigma about MAT is common within mutual help organizations and psychosocial programs [32]. Intervention stigma, from people who disagree with the use of medications to treat OUD, is widespread [19]. For them, MAT is equivalent to "trading one drug with another". Language used to describe addiction further limits professional healthcare. For example, recognizing opioid addiction as a "willful choice", and not a disease, separates OUD treatment models from the rest of the medical system [10].

Due to the presence of extensive stigma within the offline communities, people with OUD often turn to non-conventional ways to recover – e.g., using online resources or communities to seek information and support. People use Reddit to investigate alternative substances used for opioid use recovery [6], share their experiences [5], and freely discuss substance use [4]. However, individuals' attempts at seeking information on substance (mis)use from online forums are challenged due to the presence of inaccurate and harmful health misinformation. A systematic review of four web-based platforms (Twitter, Reddit, YouTube, and Drugs-Forum) revealed pronounced online prevalence of a myth associated with MAT, i.e., "MAT is simply replacing one drug with another", especially on Twitter [9].

With the recent advancements in Large Language Models (LLMs), easy to access LLM-powered chatbots, such as ChatGPT, have become the "go-to" online resource to seek knowledge, share experiences, and self-disclose [2]. They can quickly provide information, reducing the time and effort required to research a topic manually. In addition, people are increasingly using LLMs to co-write content, e.g., to write opinionated views on controversial topics [14], emails, and social media posts. In the healthcare context, which is of particular interest to us, scholars have identified the potential of LLM-powered chatbots in assisting communication between healthcare workers and patients [1, 3], making diagnosis, and recommending tests or treatments [31]. However, it is warranted to consider potential harms of LLMs in conjunction to their benefits. The models are trained on datasets that are not transparent and may contain biased information or misinformation, leading to LLMs further reproducing those in their generated output [13]. Zack et al. [34] found that GPT-4 produced medical diagnosis that stereotyped certain races, ethnicities, and gender identities. LLMs were also able to generate highly convincing and persuasive health misinformation against precautionary measures to take during COVID-19 [35].

Given the potential harms, it is essential to evaluate LLMs, especially when used in socially stigmatizing, high-stakes contexts such as OUD, mental healthcare, or gender-related issues. Factually inaccurate information on OUD, generated by LLM-powered chatbots, could mislead individuals during their opioid recovery journey. For example, as shown below, GPT-4[1] generates a misinforming response to a question posted on the r/OpiatesRecovery subreddit [2]. It generates a generic response stating that "detoxing from one opioid with another can be a risky strategy" – dismissing

1st HEAL Workshop at CHI Conference on Human Factors in Computing Systems, May 12, 2024, Honolulu, HI, USA
2024.

[1] Generated using the gpt-4-0613 model, 0.0 temperature, and other parameters as default.

[2] https://www.reddit.com/r/OpiatesRecovery/

the effectiveness of all opioids, including methadone, which is an FDA approved medication to treat OUD [3].

> *Question from r/OpiatesRecovery subreddit:* "I have been hooked on tramadol for about 3 years now [...] Can I detox from one opioid with the help of another (e.g., oxycodone or methadone)?"
> *Response from GPT-4:* "[...] it's crucial to understand that detoxing from one opioid with another can be a risky strategy. While it might seem like you're making progress as you're not experiencing withdrawal symptoms, you're still feeding your body opioids, which can lead to dependence on the new drug."

Incorporating LLMs into a psychotherapy tool (or chatbot) could cause genuine harm to patients with OUD – chatbots may not posses the required empathy, understanding, and contextual knowledge [29] to extend assistance. LLMs could further exacerbate misconceptions and stigmatizing views on people with OUD, MAT, or opioid addiction in general. Dissemination of such stigmatizing views could consequentially impact and shape public knowledge, perceptions, and attitudes toward people with OUD. LLM-informed negative perceptions on OUD could further marginalize the community. Through this discussion, we conclude that it is important to audit LLMs for OUD. We envision that the following approaches could be adopted to audit LLMs in the context of OUD.

- *Experimental approach:* We could recruit people to use LLMs for a reasonable amount of time to gain information related to OUD. This could be followed by a pre-/post-treatment analysis to understand the impact of using LLMs in the context of OUD.
- *Observational approach:* We could recruit people who have previously used LLMs in the context of OUD and ask them to donate their LLM interaction data. This could be followed by an observational analysis of the collected historical data (e.g., LLM-based chat or search user interactions).
- *Simulation-based approach:* We could create "agents" or "prototypical personas" using prompt-engineering techniques to simulate potential user-LLM interactions in the context of OUD (e.g., synthetic agents seeking support during opioid use recovery or relapse). This could be followed by a thorough analysis of the simulated interactions.

2 RELATED WORK

2.1 Stigma, Myths, and Misconceptions around OUD

OUD is a highly stigmatized issue. There is a widespread misconception that opioid addiction is a "willful choice", and not a disease [10]. Similar such notions pose as barriers to harm reduction strategies, e.g., MAT. First responders were found to hold negative attitudes toward the use of MAT [18]; they propagated provider-based stigma by stating that "[MAT] puts more drugs on the streets". A nationally representative web-based survey revealed pronounced presence of public/social stigma around OUD – respondents expressed that people with OUD are to blame for their own condition, lack self-discipline, and should be socially distanced [17]. Scholars have also made efforts to uncover OUD-related stigma in mass and social media. A qualitative study of the Indian online news media revealed the presence of derogatory language (e.g., the usage of "addicts", "jobless", or "junkies") when describing people with substance use disorder [11]. ElSherief et al. [9] identified the online presence of a leading myth surrounding MAT, i.e., "MAT is simply replacing one drug with another". Drawing knowledge from public health expert annotations, authors found the lowest prevalence of the said myth on web-based health communities such as Reddit and Drugs-Forum, and the highest on Twitter. A deductive content analysis of 259 Reddit posts characterized the major sources of OUD-related stigma as public (including family members), provider-based (healthcare professionals), structural (workplace, law enforcement, and self-help groups), and self [33]. Given how people are using LLM-powered chatbots or search engines for their everyday queries, it becomes essential to understand if and how LLMs propagate myths, stigma, or misconceptions related to OUD.

2.2 Evaluating LLMs for Healthcare

With the release of ChatGPT [23], researchers have begun to discuss the potential of and evaluate LLMs for different healthcare applications. By comparing human- and LLM-generated COVID-19 health misinformation, Zhou et al. [35] found LLM-generated misinformation to contain a persuasive tone and exaggerated details. In a careful evaluation of 4 LLMs – Bard, Claude, ChatGPT, and GPT4 – Omiye et al. [22] found the perpetuation of debunked and race-based medicine. On similar lines, Zack et al. [34] evaluated GPT-4 produced medical diagnosis using deductive content analysis. This uncovered medical stereotypes based on race, ethnicity, and gender. De Choudhury et al. [8] adopted an ecological framework to assess the opportunities and risks posed by LLMs in the context of digital mental health. In a cross-sectional study [3], researchers asked licensed healthcare professionals to evaluate ChatGPT responses for randomly sampled questions on the subreddit r/AskDocs. ChatGPT responses were rated higher for both quality and empathy compared to physician responses. Scholars have also stressed towards the creation of specialized audit models for LLMs in healthcare [26]. One such discussed framework is "The Governance Model for AI in Healthcare (GMAIH)", which consists of four components: fairness, transparency, trustworthiness, and accountability [27]. Singhal et al. [28] evaluated LLMs for clinical knowledge. The authors introduced HealthSearchQA, a dataset consisting of commonly searched consumer medical questions. Using this dataset, along with others, they then introduced a framework for physician and lay user evaluation to assess LLM-generated responses across multiple axes, e.g., reading comprehension, recall of relevant knowledge, manipulation of knowledge, relevance, and helpfulness. Recently, Jin et al. [16] introduced XlingEval for assessing LLM responses to human-authored health-related questions via three criteria: correctness, consistency, and verifiability. They found pronounced disparity in LLM-generated responses across four major global languages: English, Spanish, Chinese, and Hindi. Despite these efforts, there is still a need to evaluate LLMs specifically for OUD. The nuanced and

[3] https://www.samhsa.gov/medications-substance-use-disorders/medications-counseling-related-conditions/methadone

sensitive nature of the topic warrants dedicated efforts to assess the applicability of LLMs.

3 APPROACHES TO AUDIT LLMS

Prior works often adopt an *observational approach* to audit LLMs [12, 16, 28], which involves conducting a post-hoc evaluation of LLM interaction data, e.g., historical LLM-powered chatbot interactions, donated by users or carefully curated by scholars. The LLM-generated data is then evaluated using metrics, such as truthfulness, information quality, or logical cohesion [12], depending on the context. For example, Jin et al. [16] used three metrics – correctness, consistency, and verifiability – to evaluate LLM-generated responses to human-authored health-related questions. On similar lines, we envision an observational audit to assess LLMs for OUD via metrics including, but not limited to, those listed in Section 4, e.g., presence of stigma, information credibility, and empathy. Though observational audits work with high-quality first-person interaction data, the historical interactions could themselves be influenced by external factors – leading to biases in our findings. Additionally, researchers make use of an *experimental approach* to audit LLMs. In this, participants are recruited to engage with LLMs for a prolonged duration. This is followed by pre-/post-treatment analysis, user interviews, or participatory engagement sessions. Jakesch et al. [15] conducted a large-scale experimental study to evaluate how LLM-powered writing assistants impact people's perception on contested issues. More recently, Rakova [25] evaluated "Zuzi", an LLM-based chatbot that provides legal assistance and social support to survivors of gender-based violence (GBV) in South Africa. Following a scenario-based approach, participants identified strengths and weaknesses of the chatbot. Though such experimental audits provide high-quality data by collecting first-person interaction experiences, they can potentially risk exposure of harmful information to an already vulnerable population. Lastly, existing work has used a *simulation-based approach* to evaluate LLMs' capabilities [7, 24, 36]. This involves the creation of "agents" or "synthetic personas", using prompt-engineering techniques, to simulate or mimic real-world user-LLM interactions. The synthetic interactions are then carefully examined to assess LLMs. For example, Zhou et al. [36] created `SOTOPIA`, a simulation-based interactive environment to evaluate LLMs' social intelligence. Though this approach does not pose any direct threat to people, it surfaces concerns on how close the simulated interactions are to *actual* real-world interactions.

4 QUESTIONS TO ANSWER WHILE AUDITING LLMS FOR OUD

Based on our discussion in Sections 1 and 2, following are some potential open questions that could be answered while auditing LLMs for OUD:

- *Stigma:* Do LLMs exacerbate or abate stigma related to OUD? Can LLM-generated responses surface new unidentified stigmatized perceptions related to OUD?
- *Credibility:* Do LLMs spread misinformation or misconceptions related to OUD? If yes, do LLMs use persuasive framing, or provide reasoning to justify such misinformation? Contrastively, do LLMs counter or correct misinformation related to OUD? Or, how can we teach LLMs to self-correct in case they generate misinformative responses?
- *Empathy:* How capable are LLMs in providing genuine support to people with OUD? Do they generate trustworthy, empathetic responses containing shared-lived experiences?
- *OUD Recovery support:* What are the knowledge-sharing capabilities of LLMs in terms of recovery or treatment support provision for OUD? Do LLMs provide recommendations situated in medically-approved treatments for OUD?

5 BOTTLENECKS WHILE AUDITING LLMS FOR OUD

Here, we list some challenges researchers may face when evaluating LLMs for OUD:

- Given that OUD is a high-stakes sensitive issue, it is essential to gain the trust of people with OUD to recruit them to conduct experimental or observational audits for LLMs.
- The nuanced context of OUD warrants human-centered knowledge from domain experts to build effective LLM evaluation frameworks. We would have to come up with OUD-centric evaluation metrics to assess the quality of LLM-generated responses. For example, dedicated efforts would be needed to evaluate the level of OUD-related stigma in LLM-generated responses.
- The LLM landscape is rapidly evolving. For instance, the response generated by ChatGPT today may not be reproducible in the future. Consequently, frameworks adopted to audit LLMs should be flexible enough to adapt to the changes and not assume that the underlying system would remain static.

REFERENCES

[1] 2023. Will ChatGPT transform healthcare? *Nature Medicine* 29, 3 (01 Mar 2023), 505–506.

[2] Fahad Alanezi. 2024. Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *Journal of Multidisciplinary Healthcare* 17 (2024), 461–471.

[3] 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, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

[4] Duilio Balsamo, Paolo Bajardi, Gianmarco De Francisci Morales, Corrado Monti, and Rossano Schifanella. 2023. The Pursuit of Peer Support for Opioid Use Recovery on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 12–23.

[5] Amanda M Bunting, David Frank, Joshua Arshonsky, Marie A Bragg, Samuel R Friedman, and Noa Krawczyk. 2021. Socially-supportive norms and mutual aid of people who use opioids: An analysis of Reddit during the initial COVID-19 pandemic. *Drug and alcohol dependence* 222 (2021), 108672.

[6] Stevie Chancellor, George Nitzburg, Andrea Hu, Francisco Zampieri, and Munmun De Choudhury. 2019. Discovering alternative treatments for opioid use recovery using social media. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[7] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. arXiv:2311.09618 [physics.soc-ph]

[8] Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and Harms of Large Language Models in Digital Mental Health. https://doi.org/10.31234/osf.io/y8ax9

[9] Mai ElSherief, Steven A Sumner, Christopher M Jones, Royal K Law, Akadia Kacha-Ochana, Lyna Shieber, LeShaundra Cordier, Kelly Holton, and Munmun De Choudhury. 2021. Characterizing and identifying the prevalence of web-based

misinformation relating to medication for opioid use disorder: Machine learning approach. *Journal of medical Internet research* 23, 12 (2021), e30753.

[10] Renee Garett and Sean D. Young. 2022. The Role of Misinformation and Stigma in Opioid Use Disorder Treatment Uptake. *Substance Use & Misuse* 57, 8 (2022), 1332–1336.

[11] Abhishek Ghosh, Chandrima Naskar, Nidhi Sharma, Shinjini Choudhury, Aniruddha Basu, Renjith R Pillai, Debasish Basu, and SK Mattoo. 2022. Does online newsmedia portrayal of substance use and persons with substance misuse endorse stigma? A qualitative study from India. *International Journal of Mental Health and Addiction* 20, 6 (2022), 3460–3478.

[12] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736* (2023).

[13] Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias assessment and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345* (2023).

[14] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[15] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[16] Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. *ArXiv* abs/2310.13132 (2023). https://api.semanticscholar.org/CorpusID:264405758

[17] Alene Kennedy-Hendricks, Colleen L Barry, Sarah E Gollust, Margaret E Ensminger, Margaret S Chisolm, and Emma E McGinty. 2017. Social stigma toward persons with prescription opioid use disorder: associations with public support for punitive and public health–oriented policies. *Psychiatric services* 68, 5 (2017), 462–469.

[18] Nathan E. Kruis, Katherine McLean, and Payton Perry. 2021. Exploring first responders' perceptions of medication for addiction treatment: Does stigma influence attitudes? *Journal of Substance Abuse Treatment* 131 (2021), 108485.

[19] Erin Fanning Madden. 2019. Intervention stigma: How medication-assisted treatment marginalizes patients and providers. *Social Science Medicine* 232 (2019), 324–331. https://doi.org/10.1016/j.socscimed.2019.05.027

[20] National Academies of Sciences, Engineering, and Medicine. 2019. Medications for Opioid Use Disorder Save Lives. https://nap.nationalacademies.org/catalog/25310/medications-for-opioid-use-disorder-save-lives.

[21] NIDA. 2023. Drug Overdose Death Rates. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates.

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

[23] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt.

[24] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[25] Bogdana Rakova. 2024. Evaluating LLMs Through a Federated, Scenario-Writing Approach. https://foundation.mozilla.org/en/blog/evaluating-llms-through-a-federated-scenario-writing-approach/.

[26] Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked* 41 (2023), 101304. https://doi.org/10.1016/j.imu.2023.101304

[27] Sandeep Reddy, Sonia Allan, Simon Coghlan, and Paul Cooper. 2020. A governance model for the application of AI in health care. *Journal of the American Medical Informatics Association* 27, 3 (2020), 491–497.

[28] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[29] Steven Tate, Sajjad Fouladvand, Jonathan H. Chen, and Chwen-Yuen Angie Chen. 2023. The ChatGPT therapist will see you now: Navigating generative artificial intelligence's potential in addiction medicine research and patient care. *Addiction* 118, 12 (2023), 2249–2251.

[30] Kathlene Tracy and Samantha P Wallace. 2016. Benefits of peer support groups in the treatment of addiction. *Substance abuse and rehabilitation* (2016), 143–154.

[31] Francisco Tustumi, Nelson Adami Andreollo, and José Eduardo de Aguilar-Nascimento. 2023. Future of the language models in healthcare: the role of chatGPT. *ABCD. Arquivos Brasileiros de Cirurgia Digestiva (São Paulo)* 36 (2023).

[32] Sarah E. Wakeman and Josiah D. Rich. 2018. Barriers to Medications for Addiction Treatment: How Stigma Kills. *Substance Use & Misuse* 53, 2 (2018), 330–333.

[33] Meredith C. Meacham Wayne Kepner and Alicia L. Nobles. 2022. Types and Sources of Stigma on Opioid Use Treatment and Recovery Communities on Reddit. *Substance Use & Misuse* 57, 10 (2022), 1511–1522.

[34] Travis Zack, Eric Lehman, Mirac Suzgun, Jorge Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David Bates, Raja-Elie Abdulnour, Atul Butte, and Emily Alsentzer. 2023. Coding Inequity: Assessing GPT-4's Potential for Perpetuating Racial and Gender Biases in Healthcare.

[35] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. https://doi.org/10.1145/3544548.3581318

[36] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *ICLR*. https://openreview.net/forum?id=mM7VurbA4r