

Online Myths on Opioid Use Disorder: A Comparison of Reddit and Large Language Model

Shravika Mittal¹, Hayoung Jung², Mai ElSherief³, Tanushree Mitra², Munmun De Choudhury¹

¹Georgia Institute of Technology

²University of Washington

³Northeastern University

smittal87@gatech.edu, hjung10@uw.edu, m.elsherif@northeastern.edu, tmitra@uw.edu, munmund@gatech.edu

Abstract

Online communities on Reddit are a popular choice among people with opioid use disorder (OUD) to seek information on drug use, withdrawal symptoms, and recovery. LLM-powered chatbots (e.g., ChatGPT) are widely being adopted as question-answer systems for health-related queries. However, such online health information seeking could potentially be hindered by myths and misinformation on OUD, misleading or causing genuine harm to people with OUD. In this work, we examine the prevalence of 5 OUD-related myths, on treatment models and patient characteristics, within human- (taken from Reddit) and LLM-generated responses to queries on OUD. We further explore the framing strategies used within responses (both human- and LLM-generated) promoting and countering the myths. We found that all 5 myths were more widespread within human-generated responses. In addition, myth-promoting responses adopted trustworthy and authoritative framings, compared to knowledge-imparting linguistic cues within those countering the myths. Our work offers recommendations to reduce online OUD misinformation.

Introduction

Opioid overdose continues to be a leading cause of death in the United States. In 2021, there were more than 106,000 opioid drug-involved deaths in the U.S. (NIDA 2023). Central to harm reduction is decreasing barriers to treatment, such as expanding access to clinically-approved medications for addiction treatment (MAT) (National Academies of Sciences, Engineering, and Medicine 2019) and recognizing opioid use disorder (OUD) as a disease (Garett and Young 2022). However, ongoing efforts (CDC 2021) remain ineffective due to myths, misinformation, and stigma around the issue (Corrigan 2017), in both offline and online contexts.

Due to extensive stigma in the offline world (Kruis, McLean, and Perry 2021; Wayne Kepner and Nobles 2022), people with OUD often turn to non-conventional ways to recover, i.e., online platforms. Owing to pseudonymity, Reddit is a popular platform to discuss stigmatized issues (Yao et al. 2020). In particular, people with OUD use communities on Reddit to discuss substance (mis)use (Balsamo et al. 2023), alternate treatments (Chancellor et al. 2019), and recovery attempts (Bunting et al. 2021). More recently,

with democratized access to Large Language Model (LLM)-powered chatbots (e.g., ChatGPT), people are adopting them as question-answer systems, to seek nuanced health information and even self-disclose (Alanezi 2024). In addition, people are increasingly using LLMs to co-write content they publish online (Jakesch et al. 2023). Despite the benefits provided by these online forums, individuals' attempts at seeking information on substance use could potentially be challenged by inaccurate or harmful misinformation.

Scholars have looked at problematic online portrayals of MAT. Dekeseredy et al. (2021) revealed dissemination of conspiracy theories about medications for OUD on Twitter. People described them as being ineffective. An analysis of Reddit posts uncovered *intervention-based stigma* and misinformation (Wayne Kepner and Nobles 2022). ElSherief et al. (2021) performed a cross-platform analysis to study the prevalence of a leading OUD medication myth – *MAT is simply replacing one drug with another*. They found pronounced presence of it on Twitter. Such negative attitudes could serve as barriers to harm reduction and treatment. Online OUD misinformation contributes to increased reluctance to initiate treatment, distrust toward the healthcare system, and lowered self-esteem (Woo et al. 2017).

Although, unpacking online OUD misinformation has received some attention, it has only focused on myths about MAT. A deeper analysis of the varied OUD myths people are exposed to via online technologies, both human- (e.g., Reddit) and LLM-driven, is much needed to inform public health efforts. Moreover, health information seeking has been diversified with LLMs. The veracity of LLMs, in the context of OUD, has not been studied yet. Additionally, *framing* of content can not only inform the public *what* to think about but *how* to think about the underlying issue (Entman 1993). Scholars have not yet explored the underlying framing of online OUD misinformation, which can *drive* and *impact* public engagement (Zhou et al. 2023). Does content promoting OUD myths adopt trustworthy, evidence-based, or persuasive framings? Such an intricate analysis could inform well-guided efforts towards harm reduction. Our study pursues the following two research questions:

RQ1: Do human- and LLM-generated responses (a) promote or (b) counter myths associated with OUD? If so, how prevalent are they?

RQ2: (a) How do responses promoting and countering the

myths differ in framing? (b) Are there differences across human- and LLM-generated responses?

To answer these, we used a Reddit Question-Answer (Reddit-QA) dataset containing posts with a query on OUD and the corresponding top-voted comments: *proxy for human-generated responses*. We then used a carefully-crafted prompt to get LLM-generated responses, via GPT-4, for OUD queries. This was followed by developing few-shot classifiers to identify human and LLM responses that promote and counter 5 OUD myths (RQ1), encompassing a wide array of online misinformation on treatment models and patient characteristics. Our study experimented with 2 prompting strategies: (a) *few-shot*, and (b) *few-shot with chain-of-thought*, and 3 LLMs: (a) LLaMa-2-13B, (b) Mistral-7B, and (c) gpt-3.5-turbo for classification. We employed expert-annotated ground truth examples (ElSherief et al. 2024) as few-shots to the classifiers.

Our analysis revealed that gpt-3.5-turbo, with a few-shot chain-of-thought prompt, achieved the best classification performance. Myth-promoting content was more widespread within the human-generated responses, in comparison to those generated by the LLM (i.e., GPT-4). As per framing (RQ2 (a)), adopting the taxonomy proposed for identifying credible and misinformation (Zhang et al. 2018), responses containing myths used more trustworthy and assertive linguistic cues, while those containing counter-myths had a higher tendency to impart knowledge. Finally, LLM responses were a lot more agreeable and knowledge-based in comparison to human-generated ones (RQ2 (b)).

Overall, this work makes the following contributions. **(1)** We offer a comprehensive evaluation of OUD-related myths within human- (taken from Reddit) and LLM-generated responses to OUD queries. To better understand the landscape, we also identify responses that counter (or challenge) the said myths. Our work looks at 5 leading myths on treatment models and patient characteristics. **(2)** We contribute *few-shot learners*, grounded in representative examples annotated by public health experts, to identify responses that promote and counter the myths. **(3)** We identify the presence of 4 framing strategies – assertive confidence, alternate diverging views, external knowledge, and persuasive appeal – and examine how they differ across (a) myth-promoting and myth-countering responses, and (b) human- and LLM-generated responses. Our findings offer recommendations to reduce online misinformation on OUD such as machine-assisted content moderation and automated rewrites of harmful OUD information.

We caution the readers that some examples, provided to better contextualize the results, can be misleading.

Background and Related Work

Myths, Misinformation, and Stigma Around OUD

OUD is more stigmatized than other health conditions (Corrigan 2017). In a nationally representative web-based survey, respondents expressed that people with OUD are to blame for their own condition, lack self-discipline, and should be socially distanced (Kennedy-Hendricks et al. 2017). Such public stigma propagates myths and misinformation, which

serve as a barrier to harm reduction and treatment. Opioid addiction is *falsely* recognized as a “willful choice”, and not a disease treatable with medicine (Garett and Young 2022). Misinformation surrounds Food and Drug Administration (FDA) approved medications for addiction treatment. People, including first responders, believe that “[MAT] puts more drugs on streets” (Kruis, McLean, and Perry 2021).

To avoid such negative attitudes in the offline world, people with OUD often turn to non-conventional *online recovery pathways*. Online discussions span across different phases of opioid addiction – *using, withdrawing, and recovering*. For instance, using machine learning models, scholars identified individuals at risk of fentanyl misuse on Reddit (Garg, Taylor, and et al. 2021). Fan et al. (2017) built a classifier to detect those suffering from OUD on Twitter. Content analysis of 100 Reddit conversations revealed therapeutic community interactions on OUD (D’Agostino et al. 2017). Factors including instillation of hope, universality, imparting information, and altruism were evident. MacLean et al. (2015) examined recovery trajectories from substance use disorder, revealing positive correlations between online forum use and recovery. People use support communities on Reddit to freely discuss substance use (Balsamo et al. 2023), share recovery experiences (Bunting et al. 2021), or explore alternate treatment models (Chancellor et al. 2019). However, attempts at seeking information on substance use from online forums are challenged due to the presence of inaccurate and harmful health misinformation.

Deductive qualitative analysis of 269 Reddit posts found misinformation on MAT (Wayne Kepner and Nobles 2022). A large percentage of posts did not recognize individuals on MAT as those recovering from opioid use. ElSherief et al. (2021) made the first large-scale attempt to identify the online presence of a leading myth surrounding MAT, i.e., “MAT simply replaces one drug with another.” They found the lowest prevalence of the said myth on web-based health communities such as Reddit and Drugs-Forum, and the highest on Twitter. More recently, ElSherief et al. (2024) proposed a multistage analytic pipeline to uncover OUD-related myths and misinformation on social media. Through this, they contributed an expert-annotated dataset of 303 online posts, spreading 5 myths on OUD treatment models and patient characteristics. Our work makes use of this dataset to contribute computational techniques that can perform a large-scale analysis and inform the prevalence of a wide variety of OUD health misinformation on Reddit, a platform where people suffering from OUD and other socially stigmatized conditions feel safe to seek information (Balsamo et al. 2023), warranted to facilitate mitigation efforts.

Health Information Seeking and LLMs

With the release of ChatGPT, an LLM-powered conversational agent, people are increasingly engaging with it to discuss high-stakes issues such as suicidal thoughts, medical diagnosis and treatment (Reardon 2023). This democratized access has motivated scholars across academic, medical, industry, and public policy domains to consider the benefits and harms of LLMs in healthcare (De Choudhury, Pendse, and Kumar 2023). In a cross-sectional study (Ay-

ers et al. 2023), researchers asked licensed healthcare professionals to assess ChatGPT responses for randomly sampled questions on the r/AskDocs subreddit. ChatGPT responses were rated higher for both quality and empathy compared to physician responses. Conversely, Zhou et al. (2023) found that GPT-3 generated persuasive COVID-19 health misinformation with enhanced details. A careful evaluation of 4 LLMs – Bard, Claude, ChatGPT, and GPT-4 – surfaced the perpetuation of debunked and race-based medicine (Omiye et al. 2023). When asked to generate instructions to help a non-breathing victim, Google Bard and Bing Chat gave plausible sounding deceptive guidance (Birkun and Gautam 2023). Xiang et al. (2023) found that LLMs produced misinformation on maternity and infant care when answering expert-verified questions in Chinese.

Recently, Tate et al. (2023) positioned the role of Generative AI in addiction medicine research. The authors argue that though LLM-powered chatbots increase access to counseling services (by potentially serving as *online therapists*), there is a need to evaluate if they can cause genuine harm to people with OUD. This study makes the first attempt to evaluate whether or not LLM responses, generated by GPT-4, to online queries on Reddit, propagate myths on OUD.

Data

Owing to pseudonymity, Reddit enables candid discussions on OUD – people with OUD often share original and unfiltered first-hand experiences (Chancellor et al. 2019). Reddit’s popularity among people with OUD (Balsamo et al. 2023) and the availability of long form content (Smith et al. 2021) motivated us to look at (a) *online queries* on OUD and (b) the corresponding *human-generated responses* posted on Reddit. We used a Reddit Question-Answer dataset (Laud et al. 2025), hereafter referred to as *Reddit-QA*.

Reddit-QA consists of 150,436 original posts containing an OUD-related query and the associated comments as a proxy for human-generated responses. On average, each post has 12 comments. Posts containing a query were identified using a BERT-based question classifier (Turc et al. 2019), trained on a Quora Question-Answer dataset and manually validated. The dataset was collected using the Pushshift Reddit API from January 2018 to September 2021. *Reddit-QA* belongs to a period before the availability of democratized LLMs, e.g., ChatGPT (released in 2022), making it difficult for lay individuals to use LLMs to write posts/comments on Reddit. This ensures that the comments within *Reddit-QA* are majorly human-generated. The posts belong to 19 OUD-specific subreddits, e.g., r/Methadone, r/OpiatesRecovery, and r/Heroin, including those explored by prior work (Chancellor et al. 2019; Balsamo et al. 2023) and having a large number of members. They contain queries on diverse OUD topics, e.g., drug use, lifestyle, treatment, and withdrawal.

Filtering and Data Cleaning

Within *Reddit-QA*, we removed posts that received no comments. This reduced the posts to 139,013. There is evidence that users’ engagement on Reddit via upvoting and

Dataset	Mean (#words)	Median	Std dev
Reddit-QA-posts	585.61	318.00	964.11
Reddit-QA-comments	300.09	176.00	400.32
GPT-4-responses	383.13	116.00	388.21

Table 1: Statistics of *Reddit-QA* and GPT-4 responses.

downvoting is reflective of their interest and conformity to community norms (D’Agostino et al. 2017). Highly upvoted posts and comments are also more likely to be viewed by people (Salihefendic 2015). We filtered out posts that did not receive a high score (score = #upvotes - #downvotes). On average, posts within *Reddit-QA* received a score of 3. We removed posts that got a less than average score, finally leaving us with 70,149 posts. We refer to these as *Reddit-QA-posts*. This filtering helped us explore higher engagement, more likely to be seen posts, which may receive myth-promoting responses.

For each post, we only used the top-most voted comment as the representative human-generated response to it. We refer to these comments as *Reddit-QA-comments* throughout the paper. Again, this decision was based on how Reddit drives user experiences (Salihefendic 2015). In addition, considering a single comment in contrast to all allowed us to conduct a fair comparison with LLMs, which in the default case generate a single response. Table 1 provides length of posts and comments within the final filtered dataset. Table A3 lists paraphrased exemplars within *Reddit-QA-posts* and *Reddit-QA-comments*.

LLM-generated Responses to OUD Queries

We used GPT-4 (OpenAI et al. 2024), with a manually designed prompt, to gather LLM-generated responses to OUD queries. GPT-4 is known to outperform previous LLMs and state-of-the-art systems on a suite of natural language generation tasks (OpenAI et al. 2024). It also has substantially improved ability to follow user intent (Ouyang et al. 2022). Specifically, we used `gpt-4-0613`, the most capable and stable text generation model available at the time of this work. We chose a sampling temperature of 0.7. This temperature value was chosen, following prior work (Jakesch et al. 2023), to generate varied and creative responses that could potentially be posted on Reddit.

Prompt creation: We used prompt design to hand-craft a text guiding GPT-4 to generate responses to OUD queries. Specifically, we referred to well-adopted prompt engineering guidelines (OpenAI 2023b). These guidelines recommend to include (a) simple task-specific instructions, (b) context, (c) question of interest, and (d) output format.

To conduct a fair comparison of human-generated responses from Reddit (*Reddit-QA-comments*) and LLM-generated responses, we assigned GPT-4 the persona of an active Reddit user who gathers and provides information on OUD. We then asked it to generate a comment answering a Reddit post containing an OUD-related query. These Reddit posts were taken as is from the earlier described *Reddit-QA-posts* dataset. To assist generation, we provided the post’s subreddit name and description. We

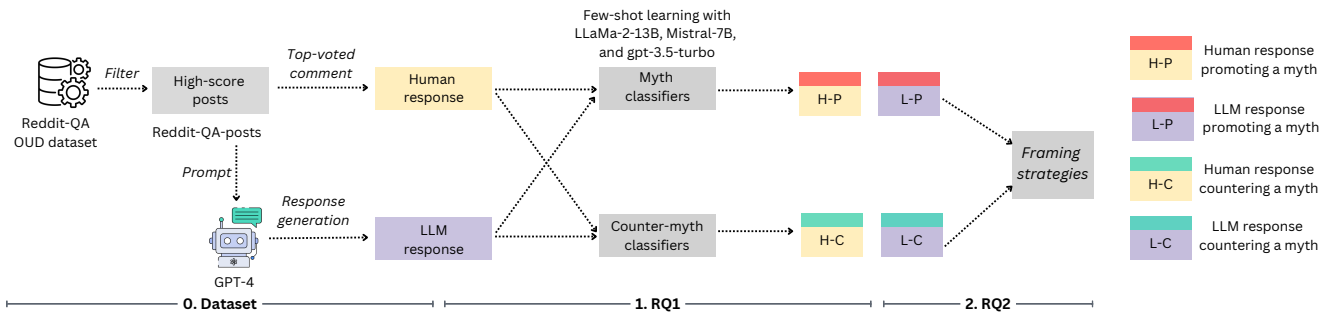


Figure 1: Overview of our method: (0) dataset; both human- and LLM-generated, (1) myth and counter-myth classification per RQ1, and (2) identifying framing strategies per RQ2.

explored multiple prompts to finalize these choices, as detailed in Section “Prompt Exploration” in the Appendix. The final prompt is detailed in Table A1. We refer to the generated dataset as `GPT-4-responses`. Table 1 provides statistics on the length of `GPT-4` generated responses.

Data validation: We took multiple steps to validate the `GPT-4-responses` dataset, making sure that it was comparable to `Reddit-QA-comments`. To begin, Table 1 shows that human- and LLM-generated responses were of similar length. The first-author, having prior experience in social computing research and in consultation with two coauthors who have conducted research in the context of OUD, read through 50 samples randomly drawn from the `Reddit-QA-posts` dataset and the corresponding (a) human-generated responses taken from `Reddit-QA-comments` and (b) LLM-generated responses taken from `GPT-4-responses`. This analysis revealed that the human- and LLM-generated responses were semantically similar. Similar to Reddit, `GPT-4` responses contained personal narratives, information on opioids, and peer support indicators. Table A3 lists some comparative human- and LLM-generated responses as examples.

Methods

Myths on OUD: Training Data

For RQ1, we used the dataset curated by ElSherief et al. (2024) to train our myth and counter-myth detection classifiers. It consists of social media posts (from Twitter, YouTube, Reddit, and Drugs-Forum) annotated by public health experts with binary labels for whether or not a post (a) promotes and (b) counters the myth. Table 2 lists OUD myths for which annotations are available. *Our work focuses on these five myths.* Given our focus on Reddit, we limit the training data to only contain Reddit posts. Table 2 presents the number of Reddit posts that promote and counter the myths within this ground truth data.

Why these myths? In ElSherief et al. (2024), three public health experts – two clinicians and one doctoral-level epidemiologist – used their domain expertise, performed a systematic review of the 4 online platforms, and referred to substance use literature (Heimer, Hawk, and Vermund 2019; Wakeman and Barnett 2018) to identify the five myths.

These myths are also recognized by health organizations such as the Substance Abuse and Mental Health Services Administration (SAMHSA), Johns Hopkins Medicine, and Yale Medicine (Chen 2017; Johns Hopkins Medicine 2020).

Myth and Counter-myth Classification

To identify human- and LLM-generated responses that (a) promote and (b) counter the OUD-related myths listed in Table 2, we built LLM-based classifiers using (a) *few-shot* (Brown et al. 2020) and (b) *few-shot with chain-of-thought* (Wei et al. 2022) prompting strategies. We used a few-shot learning approach due to the limited availability of high-quality ground truth annotations, e.g., only 7 samples in our ground truth dataset promote M5 (Table 2). Few-shot prompts that condition LLMs on a few task-specific examples have been shown to outperform supervised training paradigms, e.g., RoBERTa, particularly to detect misinformation, stereotypes, and hate speech (Zhang et al. 2023). Furthermore, a step-by-step chain-of-thought reasoning with few-shot examples, indicating why a sample belongs to a class, facilitates LLM-based classification (Madaan, Hermann, and Yazdanbakhsh 2023). We created a separate prompt for the two tasks, i.e., to identify responses that (a) *promote* and (b) *counter* OUD-related myths. Within the two tasks, we had a different prompt (thereby, a different classifier) for each of the five myths. The few-shot prompts contained five ground truth examples (3 positive; 2 negative) from the training data, which is the most common setting (Parnami and Lee 2022). In addition to the few-shot examples, the chain-of-thought prompts contained step-by-step reasoning explaining why the examples promote or counter a myth. We referred to prompt engineering guidelines for creating few-shot and chain-of-thought prompts (OpenAI 2023b; DAIR.AI 2023). Refer to Tables A4 and A5 for the prompts used. Based on their widespread usage, we used two open-source LLMs, `LLaMa-2-13B` (Touvron et al. 2023) and `Mistral-7B` (Jiang et al. 2023), and the proprietary LLM, `gpt-3.5-turbo`. For `gpt-3.5-turbo`, we used a sampling temperature of 0 as the underlying task, i.e., classification, is more well-defined and coherent, not requiring creative or diverse outputs (OpenAI 2023a). We did not include `GPT-4` as a classifier because we used it to obtain

ODD-related myth	N_P	N_C
M1: Agonist therapy or MAT is merely replacing one drug with another.	22	63
M2: People with OUD are not suffering from a disease treatable with medication but from a self-imposed condition maintained through the lack of moral fiber.	10	6
M3: The ultimate goal of treatment for OUD is abstinence from any opioid use.	16	16
M4: Only patients with certain characteristics are vulnerable to addiction.	11	8
M5: Physical dependence or tolerance is the same as addiction.	7	14

Table 2: OUD-related myths for which annotations are available. N_P (N_C) represents the number of Reddit posts promoting (countering) the myth in the ground truth training data (ElSherief et al. 2024).

LLM-generated responses to OUD queries and wanted as little overlap as possible across classification and generation. Appendix Section “Rationale for Classifier Choices” further expands on our rationale for the classifiers.

Framing of Myths and Counter-myths

Frames in a communicating text can not only inform the public *what* to think about, but *how* to think about the underlying issue (Entman 1993). Zhang et al. (2018) proposed a domain-agnostic taxonomy of framing strategies used to convey credible and misinformation in mass and social media. To answer RQ2, we use 4 of these – (1) assertive confidence, *expressions demonstrating trust and authority*; (2) alternate diverging views, *reasoning other possible standpoints*; (3) external knowledge, *exchange of expert know-how*; and (4) persuasive appeal, *descriptions of lived experiences* (Table A6) – to study how responses that promote and counter myths differ in framing. From the proposed taxonomy, we only considered strategies that are relevant to this work’s context, i.e., Reddit – dropping irrelevant ones such as presence of citations, advertisement, or clickbait titles.

We used the following computational tools to perform a large-scale analysis of the 4 framing strategies:

1. *Social dimension classifiers*: Choi et al. (2020) provided 10 social dimension classifiers to model interactions on online communities. The classifiers are fine-tuned BERT (Devlin et al. 2019) models, trained on a dataset of 7,855 Reddit posts. They generate a score in [0, 1] to indicate how well an input text aligns with the social dimension. Prior work has used these to study peer support (Balsamo et al. 2023) and the role of conversations in meeting success (Zhou et al. 2021). Amongst the 10, we only used the relevant 4 to identify 3 of the framing strategies – (a) Trust, Power for assertive confidence, (b) Conflict for alternate diverging views, and (c) Knowledge for external knowledge. We selected these classifiers by matching their descriptions to those of the strategies (refer to Section “Description of Framing Strategies and Computational Methods” in the Appendix).
2. *StorySeeker*: Antoniuk et al. (2024) developed a fine-tuned RoBERTa classifier to get persuasive storytelling narratives in text. The classifier was trained on a binary labeled dataset, indicating presence of a storytelling narrative, of 502 texts taken from Reddit. We adopt this tool to identify persuasive appeal framing strategy.

Results

Evaluation of Classifiers

Per RQ1, we first evaluated the 3 LLM-based classifiers, LLaMa-2-13B, Mistral-7B, and gpt-3.5-turbo, on the task of identifying human- and LLM-generated responses that *promote* an OUD myth. As described earlier, we explored two prompting techniques: (a) *few-shot* and (b) *few-shot with chain-of-thought* and assessed the classifiers on both. Given the limited availability of ground truth (refer to Table 2), we manually annotated a random sample of the Reddit-QA-comments and GPT-4-responses datasets, against which we then validated our classifiers.

To validate the classifier for each myth, we randomly sampled 50 responses each from the Reddit-QA-comments and GPT-4-responses datasets and labeled them manually: ‘0’ to indicate that the response does not promote the myth and ‘1’ to indicate it does. In total, we annotated 5 random samples, of 100 responses each, corresponding to the 5 myths. The first two authors initially labeled 20 responses referring to the expert-annotated ground truth examples (described in Section “Myths on OUD: Training Data”) and substance use literature (Heimer, Hawk, and Vermund 2019). As an indicator of inter-rater reliability, the annotators received a Cohen’s Kappa of 0.83 (M1), 0.77 (M2), 0.73 (M3), 0.77 (M4), and 0.71 (M5). After reaching 100% agreement and consensus, the first author then annotated the remaining 80 responses. As a demonstration of annotator credibility, both authors have prior experience in social computing, particularly working with Reddit data, and healthcare research. We then compared the binary labels assigned by the LLM-based myth classifiers against our hand-annotated ground truth labels. Table 3 summarizes classification performance. LLaMa-2-13B consistently achieved the lowest classification performance, i.e., the lowest F1-score. Few-shot with chain-of-thought prompt led to a minor improvement in classification across all 3 classifiers. To identify responses promoting M1, F1-score for gpt-3.5-turbo increased by 2.44% (from 0.82 to 0.84) using the few-shot with chain-of-thought prompt. Finally, gpt-3.5-turbo, with a few-shot chain-of-thought prompt, performed the best across all classifiers in identifying all five myths. It achieved an F1-score as high as 0.84 to detect responses promoting M1. The gpt-3.5-turbo classifier was the least competent in identifying M2 (*people with OUD are not suffering from a disease treatable with medication*) and M5 (*physical dependence or tolerance is*

Myth	LLaMa-2-13B			Mistral-7B			gpt-3.5-turbo			LLaMa-2-13B			Mistral-7B			gpt-3.5-turbo			McNemar's	
	Few-shot prompt									Few-shot with Chain-of-Thought prompt									χ^2	p
	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1		
M1	0.68	0.63	0.65	0.72	0.68	0.70	0.84	0.79	0.82	0.70	0.64	0.67	0.74	0.68	0.71	0.86	0.83	0.84	3.15	*
M2	0.64	0.61	0.63	0.68	0.66	0.67	0.80	0.75	0.77	0.69	0.61	0.65	0.72	0.69	0.70	0.82	0.76	0.79	5.09	**
M3	0.68	0.67	0.67	0.68	0.64	0.66	0.82	0.79	0.80	0.71	0.69	0.70	0.74	0.70	0.72	0.84	0.80	0.82	2.41	**
M4	0.69	0.65	0.67	0.72	0.69	0.70	0.82	0.79	0.81	0.70	0.66	0.68	0.72	0.70	0.71	0.85	0.81	0.83	4.39	*
M5	0.63	0.59	0.61	0.70	0.62	0.66	0.78	0.74	0.76	0.64	0.61	0.62	0.72	0.65	0.68	0.80	0.77	0.79	4.97	**

Table 3: Evaluation of LLaMa-2-13B, Mistral-7B, and gpt-3.5-turbo as classifiers to identify whether a human- or LLM-generated response promotes an OUD-related myth. Performance summarized for two prompting strategies: (a) *few-shot* and (b) *few-shot with chain-of-thought* via 3 metrics: precision (Pr.), recall (Re.), and F1-score (F1). McNemar’s tests ($p < 0.1$: ‘*’, $p < 0.05$: ‘**’, $p < 0.01$: ‘***’) were performed to compare gpt-3.5-turbo’s performance for few-shot and few-shot with chain-of-thought prompting strategies.

the same as addiction) – 0.79 F1-score using the few-shot with chain-of-thought prompt.

We adopted a similar methodology to evaluate LLM-based classifiers for identifying responses that counter OUD myths. The performance is summarized in Table A8. Again, gpt-3.5-turbo with a few-shot chain-of-thought prompt performed the best. Refer to Section “Evaluation of Counter-myth Classifiers” in the Appendix for more details.

We discuss some of the typical types of misclassification (false positives or false negatives) in an error analysis given in Appendix Section “Error Analysis of Classification”.

RQ1: Prevalence of Myths and Counter-myths

To study the prevalence of myths and counter-myths in human- and LLM-generated responses (RQ1), we used the best-performing classifier, gpt-3.5-turbo with a few-shot chain-of-thought prompt, to classify responses within the `Reddit-QA-comments` (human-generated) and `GPT-4-responses` (LLM-generated) datasets.

Human-generated responses: Table 4 summarizes raw statistics of human-generated responses that promote and counter the 5 OUD-related myths as identified by the classifier. It can be observed that M4, i.e., *only patients with certain characteristics are vulnerable to addiction*, is the most prevalent myth within the `Reddit-QA-comments` dataset (1592 or 2.27% of the responses promote this myth). To make things worse, it is the least countered myth within the human-generated responses. Only 891 or 1.27% of the responses do so. The ratio between responses that promote and counter M4 is greater than 1 (1.787), indicating that the responses promoting M4 surpass those countering it. On the other hand, M3 (*the ultimate goal of treatment for OUD is abstinence from any opioid use*) is the least commonly occurring myth within the human-generated responses – 154 or 0.22% of the responses promote this myth. It is also countered by a fair number of responses (1992); the ratio (0.077) between the number of responses that promote and counter M3 is close to 0. The most countered myth is M2, i.e., *people with OUD are not suffering from a disease treatable with medication* – 2118 or 3.02% of responses counter M2.

Consider the following exemplars, taken from the `Reddit-QA-comments` dataset, that promote or counter

some of the OUD-related myths. The paraphrased excerpt below associates certain derogatory characteristics (“if you’re weak”) with opioid use relapse (“you will fall back”), thereby promoting M4. Contrastively, the comment following it speaks against such negative attributions.

“You will fall back just by going to the clinic. If you’re weak like me you will. There is dope everywhere.” (*Promotes M4*)

“[...] This is almost unheard of in Europe, Canada. They do not see them as weak or jobless.” (*Counters M4*)

LLM-generated responses: Next, as identified by the gpt-3.5-turbo classifier, M1 (*agonist therapy or MAT is merely replacing one drug with another*) is promoted the highest number of times within the `GPT-4-responses` dataset – 182 or 0.26% of the LLM-generated responses promote M1. It is also the second least countered myth within our dataset; 2224 or 3.17% of the responses counter M1. Similar to the human-generated responses, M2 is the most countered myth within `GPT-4` generated responses – 11601 or 16.54% of the responses counter M2. In addition, it is promoted minimal number of times. Only 77 or 0.11% of the LLM-generated responses promote M2. Consequently, M2 is the least widespread – the ratio between the number of responses that promote and counter M2 is the lowest (0.007). Again, identical to human-generated responses M3 is the least commonly occurring myth here. Only 63 or 0.09% of the responses promote M3.

We provide some paraphrased excerpts below to show how responses generated by `GPT-4` promote or counter the OUD-related myths. The following response promotes M2 by equating opioid addiction to a self-imposed condition (“not fully committed”). On the contrary, the response below it supports the disease model of addiction (“addiction is a disease, not a moral failing”) and recognizes that addiction is an involved process, which extends beyond the lack of willpower (“not about wanting to stop”).

“[...] using opioids indicates that you’re not fully committed to recovery. Recovery isn’t about stopping the use, it’s about changing your mindset.” (*Promotes M2*)

“It’s crucial to understand that addiction is a disease, not a moral failing. [...] It’s not about wanting to stop, it’s about needing help to stop.” (*Counters M2*)

Comparing across human- and LLM-generated responses, it can be observed that myths are more widespread within

Myth	Human-generated responses			LLM-generated responses		
	Count (P; C)	% (P; C)	Ratio (P/C)	Count (P; C)	% (P; C)	Ratio (P/C)
M1	1011; 1108	1.44; 1.58	0.912	182; 2224	0.26; 3.17	0.082
M2	870; 2118	1.24; 3.02	0.411	77; 11602	0.11; 16.54	0.007
M3	154; 1992	0.22; 2.84	0.077	63; 3346	0.09; 4.77	0.019
M4	1592; 891	2.27; 1.27	1.787	168; 1599	0.24; 2.28	0.105
M5	923; 1915	1.32; 2.73	0.482	70; 5886	0.10; 8.39	0.012

Table 4: Raw statistics of human- (from `Reddit-QA-comments`) and LLM-generated (from `GPT-4-responses`) responses that promote (*P*) and counter (*C*) the OUD-related myths. *Count* and *%* represent the raw count and percentage of responses respectively. *Ratio* indicates the ratio between the number of responses that promote and counter the myths.

the `Reddit-QA-comments` dataset. Ratio of number of responses that promote and counter the myths (refer to *Ratio (P/C)* in Table 4) for LLM-generated responses is less than that for human-generated responses across all myths (e.g., M1: 0.912 for human and 0.082 for LLM responses).

RQ2: Framing Strategies

For **RQ2 (a)**, we explored how framing strategies (described in Section “Framing of Myths and Counter-myths”) differed across responses that promote and counter OUD-related myths, as classified by the `gpt-3.5-turbo` classifier. After identifying responses in the `Reddit-QA-comments` and `GPT-4-responses` datasets that promote and counter the myths, we studied (1) their alignment to the four social dimensions (trust, power, conflict, and knowledge) and (2) the presence of persuasive story-telling narratives to understand the underlying framing strategies.

Looking at score distributions for how aligned the human- (Figures 2a, 2b, and 2c) and LLM-generated (Figures 2e, 2f, and 2g) responses are to the social dimensions, it can be observed that those promoting M1¹ contain more trustworthy, authoritative, and diverging linguistic cues compared to ones countering it – distribution of myth-promoting responses skews towards the right for trust, power, and conflict dimensions. This finding is echoed in the aggregates presented in Tables 5 and 6. For example, the average alignment, to trust, power, and conflict dimensions, of human-generated responses promoting M1 (m_P) is greater than that of those countering it (m_C) – (0.031 vs. 0.028) for trust, (0.011 vs. 0.009) for power, and (0.467 vs. 0.313) for conflict (Table 5). This observation is also consistent across all five myths. In summary, both human- and LLM-generated responses had a higher tendency to contain the *assertive confidence* (more trust, power) and the *alternate diverging views* (more conflict) framing strategies when promoting myths.

As an example, the following paraphrased Reddit response, which received high scores for trust (0.394) and power (0.011) dimensions, uses *assertive confidence* while promoting M1. It uses deterministic (“you’ll be addicted”) and trust-provoking phrases (“trust me”) to imply that MAT, in this case Methadone, replaces one drug with another.

¹Due to space constraints we only included plots for M1, the most widely studied myth in literature. Others present in Appendix (refer to Figures A1, A2, A3, and A4).

“So you want to take Methadone? You’ll be addicted. Trust me, stay away pal!” (*Promotes M1: trust, power*)

Similarly, the `GPT-4` generated response below, which scored high on the trust (0.341) and power (0.011) dimensions, uses an assertive tone (“you are still feeding”, “this will lead to dependence”) to discourage the use of Methadone, a medication-assisted treatment for OUD.

“You are still feeding your body opioids [Methadone]. [...] will lead to dependence.” (*Promotes M1: trust, power*)

On the other hand, human- (Figure 2d) and LLM-generated (Figure 2h) responses that countered M1 contain more knowledge-based framings in comparison to those that promoted it. Per Tables 5 and 6, average alignment to knowledge is higher for responses countering M1 (0.140 vs. 0.376 for human- and 0.306 vs. 0.564 for LLM-generated responses). Again, this finding is consistent across all five myths. Therefore, both human- and LLM-generated responses used the *external knowledge* framing strategy more frequently while countering a myth.

Consider the following paraphrased Reddit exemplar, which had a strong alignment to the knowledge dimension (0.888). The poster advocates for MAT (“the best thing for you is to get on MAT”) and refers to an online resource to promote awareness on Suboxone.

“You’ve been through a lot! I think the best thing for you is to get on MAT. There is a website: [website] where you can know about and get Suboxone.” (*Counters M1: knowledge*)

On similar lines, in the following LLM-generated response, which strongly aligned to the knowledge dimension (0.892), `GPT-4` refers to services and helplines to assist with MAT treatment (“nearest methadone clinic”).

“You can contact the Oklahoma Department of Mental Health and Substance Abuse Services. They have a helpline [helpline number]. They can guide you to the nearest methadone clinic.” (*Counters M1: knowledge*)

Effect size measurement revealed meaningful differences, ranging from small to large effect, in score distributions of responses that promoted and countered the myths against all 4 social dimensions (refer to $|Cohen’s\ d|$ statistic in Tables 5 and 6). Differences were also statistically significant under Mann-Whitney U-tests (refer to p in Tables 5 and 6).

Lastly, to understand prevalence of the *persuasive appeal* framing strategy, we used *StorySeeker* to identify responses that contained a persuasive storytelling narrative.

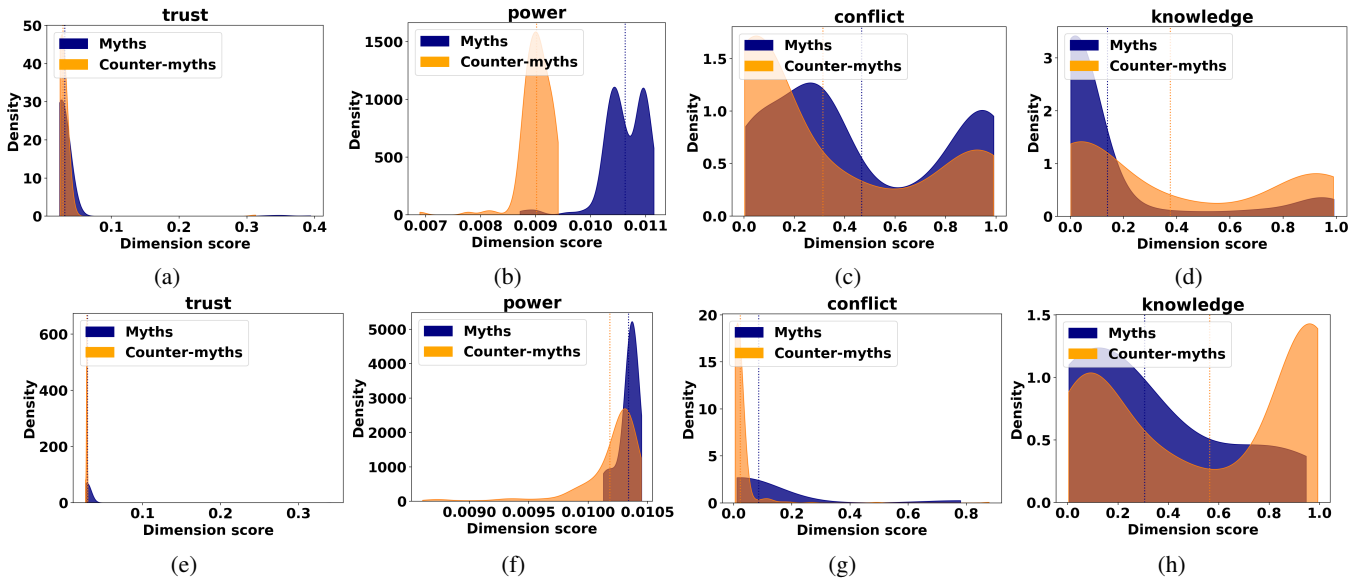


Figure 2: Density plot distributions for scores indicating the alignment of (a-d) human- and (e-h) LLM-generated responses, promoting or countering M1, to the four social dimensions: trust, power, conflict, and knowledge. (All) Dotted lines represent the average alignment score of the responses to the social dimensions.

	Trust				Power				Conflict				Knowledge			
	m_P	m_C	p	d	m_P	m_C	p	d	m_P	m_C	p	d	m_P	m_C	p	d
M1	0.031	0.028	***	0.25	0.011	0.009	***	0.32	0.467	0.313	***	0.53	0.140	0.376	***	0.67
M2	0.033	0.026	***	0.33	0.015	0.010	**	0.36	0.737	0.372	***	0.97	0.102	0.247	**	0.57
M3	0.031	0.027	**	0.26	0.021	0.011	*	0.34	0.521	0.356	*	0.54	0.145	0.310	**	0.80
M4	0.033	0.029	*	0.20	0.011	0.010	***	0.26	0.620	0.401	***	0.55	0.199	0.442	***	0.63
M5	0.027	0.018	*	0.14	0.011	0.011	*	0.17	0.353	0.352	*	0.21	0.143	0.422	***	0.78

Table 5: Statistical analysis of score distributions indicating the alignment of responses within the `Reddit-QA-comments` dataset, to the four social dimensions, that promote (P) and counter (C) a myth. m_P (m_C) represents the average alignment score of responses that promote (counter) the myth. Mann-Whitney U-tests were performed to compare the score distributions for responses that promote and counter the myths ($p < 0.1$: *, $p < 0.05$: **, $p < 0.01$: ***). d represents |Cohen’s d | or the effect size measurement between the two distributions.

	Trust				Power				Conflict				Knowledge			
	m_P	m_C	p	d	m_P	m_C	p	d	m_P	m_C	p	d	m_P	m_C	p	d
M1	0.030	0.028	*	0.29	0.011	0.010	*	0.73	0.086	0.023	**	0.49	0.306	0.564	**	0.70
M2	0.033	0.028	**	0.29	0.011	0.010	*	0.42	0.040	0.026	*	0.38	0.106	0.379	**	0.61
M3	0.036	0.024	**	0.29	0.021	0.009	**	0.50	0.071	0.029	**	0.42	0.217	0.501	***	0.72
M4	0.038	0.026	*	0.26	0.025	0.010	***	0.90	0.043	0.018	*	0.33	0.427	0.759	***	0.93
M5	0.029	0.025	*	0.18	0.010	0.010	**	0.58	0.027	0.016	*	0.29	0.567	0.678	**	0.48

Table 6: Statistical analysis of score distributions indicating alignment of responses within the `GPT-4-responses` dataset, to the 4 social dimensions, that promote (P) and counter (C) a myth. Refer to Table 5 caption for descriptions of p and d .

Table 7 summarizes this analysis. In general, a higher percentage of responses countering the myths contained a storytelling narrative in comparison to those promoting them. Thus, counter-myth responses tend to use a *persuasive appeal* framing strategy more frequently than myth responses in our human- and LLM-generated datasets. χ^2 tests re-

vealed significant differences between the distributions of persuasive storytelling narratives within responses that promote and counter the myths, across both human- and LLM-generated responses (refer to χ^2 statistic in Table 7).

In the following paraphrased Reddit response, which counters M1, the author narrates their lived experience in

Myth	Human-generated				LLM-generated			
	% P	% C	χ^2	p	% P	% C	χ^2	p
M1	8.30	12.66	20.93	*	4.30	8.42	12.11	*
M2	8.06	16.89	27.56	**	5.44	6.84	21.26	**
M3	9.09	12.77	29.43	**	8.75	11.90	19.32	*
M4	29.96	7.87	43.13	**	7.11	7.39	22.53	**
M5	4.88	7.69	19.84	*	4.10	6.89	22.93	**

Table 7: Percentage of human- and LLM-generated responses, which promote (% P) or counter (% C) a myth, containing a persuasive storytelling narrative. χ^2 tests were performed ($p < 0.1$: “*”, $p < 0.05$: “**”, $p < 0.01$: “***”).

a compelling manner (“helped me save my life”) to recommend MAT as a treatment.

“Methadone helped me save my life. [...] Looking back I wish I would have gotten on methadone sooner but the stigma kept me away. [...]” (*Counters M1: storytelling*)

Next, consider the GPT-4 generated response below. It uses a convincing framing to counter M4 (“addiction doesn’t discriminate”, “can affect anyone”) by highlighting prior involvement with the issue (“being involved”).

“Take it from me, after being involved in discussions on opioid use disorder, addiction doesn’t discriminate and it can affect anyone.” (*Counters M4: storytelling*)

Only for M4, human-generated responses promoting the myth contained the *persuasive appeal* framing strategy more often. This usually happened when Reddit users described their interactions with people with OUD. As an example, the author of the response below promotes M4 while recounting their experience with a co-worker – associating derogatory characteristics to their opioid use (“good liars”).

“[...] addicts sure learn quickly to be good liars! My co-worker in recovery calls in sick every other day. I know they are lying [...]” (*Promotes M4: storytelling*)

For RQ2 (b), we analyzed how framing strategies differed across human- and LLM-generated responses. Here we only discuss strategies for which differences were significant.

Alternate diverging views strategy. Consider the average alignment of human- and LLM-generated responses, promoting (Figure 3a) and countering (Figure 3b) the myths, to the conflict dimension. Human-generated responses have a higher alignment to conflict compared to those generated by GPT-4, in turn having a higher tendency to use the *alternate diverging views* framing strategy.

In the paraphrased Reddit excerpt below, which received a high score for conflict (0.891), the author reasons for complete abstinence (“get off completely”), therefore promoting M3, by strongly expressing a diverging viewpoint (“I can’t stand someone like that”). Contrastively, though the GPT-4 response also promotes M3 by advocating for a life free of opioids (including Methadone), it does so using comparatively lesser disagreement (scored 0.174 against conflict).

“Old justification. Methadone users are like, well at least I am not using heroin. I can’t stand someone like that. Get off completely!” (*paraphrased Reddit response: conflict*)

“[...] aim to get off opioids, including Methadone, completely.” (*GPT-4 response: conflict*)

Healthy usage of the *alternate diverging views* framing strategy could potentially help counter OUD-related myths. For example, the following paraphrased Reddit response, which scored 0.991 against the conflict dimension, challenges the existing stigma and false claims around MAT via tense framings (“anyone who disagrees feel free to argue”) to counter M1 and M3. The GPT-4 generated response also counters M1 and M3, without explicitly questioning intervention stigma around MAT (scored 0.204 against conflict).

“Don’t just stop using, do methadone maintenance. Step away from the false claims, anyone who disagrees with this feel free to argue so perhaps I can educate your hard head!” (*paraphrased Reddit response: conflict*)

“[...] Methadone can help reduce cravings and the desire to use opioids.” (*GPT-4 response: conflict*)

External knowledge strategy. Next, we looked at the average alignment of human and LLM responses, promoting (Figure 3c) and countering (Figure 3d) the five myths, to the knowledge social dimension. It can be observed that human-generated responses have a lesser alignment compared to LLM responses. Therefore, LLM responses had a higher proclivity to use the *external knowledge* framing strategy. Mann-Whitney U-tests revealed significant differences across human and LLM responses.

Consider the following GPT-4 generated response that received a high score (0.715) against the knowledge dimension. It promotes M1, discouraging the use of MAT (that are also opioids) for OUD treatment, by referring to external sources of knowledge (“evidence”). Such insight-heavy myth-promoting responses could be detrimental to harm reduction and OUD treatment. On the other hand, the paraphrased Reddit response (scored 0.421 against knowledge) below though uses technical know-how (“works in your brain like other opioids”) to promote M1, the resource or reference is not as explicit as that of GPT-4.

“There is evidence that detoxing from one opioid with another can be a risky strategy. [describes a case study]” (*GPT-4 response: knowledge*)

“Methadone is not a miracle, it works in your brain like other opioids and is highly addictive. [...]” (*paraphrased Reddit response: knowledge*)

Discussion

Recommendations to Reduce Online OUD Harm

Online myths and misinformation on OUD can lead to grave consequences, including alienation from 12-Step recovery programs and overdose deaths (Heimer, Hawk, and Vermund 2019). To this end, our work contributes few-shot LLM-based classifiers to identify the prevalence of 5 OUD-related myths within human- (taken from Reddit) and LLM- (via GPT-4) generated responses. Our experimentation revealed that the best performing classifier, gpt-3.5-turbo, achieved reasonable performance (Table 3). Therefore, as also supported by prior work (Ziems et al. 2024), there is promise in using few-shot LLM-based

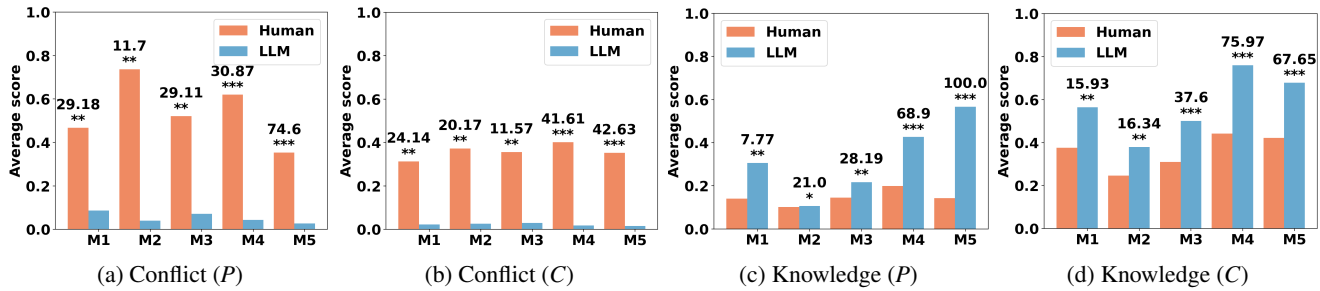


Figure 3: Averaged alignment of human and LLM responses (a, c) promoting (*P*) and (b, d) countering (*C*) myths to the conflict and knowledge dimensions. Mann-Whitney U-tests performed to compare score distributions ($p < 0.1$: ‘*’, $p < 0.05$: ‘**’, $p < 0.01$: ‘***’). Test statistic and p-values are present on the bars.

classifiers to detect OUD myths, as they can reduce the reliance on difficult-to-curate large OUD misinformation datasets. Notwithstanding some expected misclassifications, a cautious automatic approach like ours could help identify online OUD myths on a large scale and inform public health responses toward harm reduction. The classifiers could potentially act as a mitigation tool against harmful OUD practices promoted online by detecting the myths. For instance, they could be integrated into communities on Reddit, for **machine-assisted content moderation**, with human supervision, facilitating harm reduction. Our few-shot learning approach, requiring limited training data, offers promise to be easily and inexpensively adaptable to practical public health use scenarios, where prevention measures necessitate the quick **discovery of novel health misinformation**.

Specific to RQ1, human responses promoted all 5 myths much more than the LLM responses. Similar to prior work on re-writing (Sharma et al. 2021), before a comment is published on Reddit, an **automated re-authoring feature** could provide editing suggestions in case the to-be-published comment attempts to promote OUD-related myths (potentially identified by our classifiers). Moreover, RQ1 revealed a good proportion of Reddit responses countering the myths; except for M4, percentage of responses countering was more than those promoting the myths. Therefore, platforms could invest in affordances to increase the visibility of responses countering OUD-related myths. An automated system could provide **nudges** to the readers to prioritize responses that either counter or do not promote OUD myths (again, potentially identified by our classifiers). Responses generated by GPT-4 consistently promoted the explored myths less frequently (Table 4). As a result, they could be integrated into online communities to **generate credible responses with moderator verification**, supplementing existing human-generated responses for queries on OUD. Nevertheless, LLM-generated responses did promote OUD myths to some extent. Targeted expert evaluation, by fact-checking LLM-generated responses, could be performed to counter OUD misinformation. For instance, substance use experts could harness guidelines (SAMHSA 2024) adopted by public health organizations, such as the SAMHSA, to review myths within LLM responses. Insights from such a detailed expert review could then be used to fine-tune *safe* LLMs

for OUD. These guidelines could be adapted periodically as LLM capabilities change rapidly, and as new threats of OUD misinformation emerge in AI-generated content.

Next, in RQ2 (a), we found that the 5 myths and counter-myths are framed differently. Both human- and LLM-generated responses promoting OUD myths employed significantly more trustworthy, authoritative, and assertive framings than those countering them. It is known that many people place a great deal of unwarranted trust in AI systems (Rainie et al. 2022); hence, such framings could potentially mislead individuals recovering from OUD into adopting harmful beliefs, while also misleading human content moderators not to flag relevant misinformation in recovery communities. These findings highlight the need for ongoing and persistent efforts to mitigate online OUD misinformation, which can also **consider the underlying linguistic framings while detecting misinformation**.

Lastly, per RQ2 (b) we observed that the LLM-generated responses were more agreeable: they utilized the *alternate diverging views* strategy significantly less than the human-generated responses, resulting in content overbearing with external knowledge strategies. Contrarily, our findings showed that the same framing strategy could help counter OUD myths. We suggest that LLMs could be trained to **adopt more disagreeable framings to counter OUD myths** or correct misperceptions. This suggestion forms a parallel with existing work that enlists benefits of AI stepping away from the sycophantic paradigm into the antagonistic space (Cai, Arawjo, and Glassman 2024).

Limitations and Future Work

We note some limitations which provide excellent directions for future research. First, though we explored 5 myths on OUD treatment and patient characteristics, there can potentially be others within Reddit and LLM responses. Scholars could adopt our few-shot classification pipeline to explore other potential online myths on OUD. We considered high-scoring posts and corresponding responses to account for Reddit’s platform affordance. In the future, scholars should explore content that may receive varying levels of engagement, from varying demographics, especially when studying platforms that are designed differently than Reddit. For some of the myths, there was limited linguistic variation

in the ground truth data used as few-shot examples. For example, for M5, the 7 ground truth examples promoting the myth were linguistically similar to the original statement. We hypothesize this and the myth's intricate nature to be possible reasons for a lower classification performance when detecting M5, compared to others (Table 3). Future researchers could collaborate with substance use experts to gather a few varied examples. Though carefully curated, our findings could be biased based on the prompt used to generate the LLM response or the subreddits covered within the Reddit-QA dataset.

Conclusion

We sought to understand the prevalence of OUD myths and counter-myths within human- and LLM-generated responses. Using LLM-based few-shot learners, with gpt-3.5-turbo achieving the best classification performance, we observed that the explored 5 OUD myths were more widespread in the human responses. Next, myth-promoting responses contained trustworthy, assertive, and diverging framings. Those countering the myths referenced external knowledge. Lastly, LLM responses were more agreeable and knowledge-imparting compared to human-generated responses. We offer recommendations to reduce online OUD harm such as machine-assisted content moderation and automated re-authoring. All code is available at <https://github.com/mittalshravika/OUD-myths>.

Acknowledgements

Mittal and De Choudhury were partly supported through NSF grant #2230692 and a contract from the CDC. Along with Mitra, they were also partly supported through NIDA grant DA056725-01A1. This research project has benefited from the Microsoft Accelerating Foundation Models Research (AFMR) grant program.

References

Agency for Healthcare Research and Quality. 2000. Orientation to Opioid Use Disorder Treatment. <https://integrationacademy.ahrq.gov/products/playbooks/opioid-use-disorder/obtain-training-and-support-providers-and-staff/clinicwide-orientation-oud-treatment>. Accessed: 2025-04-16.

Alanezi, F. 2024. Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study. *Journal of Multidisciplinary Healthcare*.

Antoniak, M.; Mire, J.; Sap, M.; Ash, E.; and Piper, A. 2024. Where Do People Tell Stories Online? Story Detection Across Online Communities. arXiv:2311.09675.

Ayers, J. W.; Poliak, A.; Dredze, M.; Leas, E. C.; et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*.

Balsamo, D.; Bajardi, P.; De Francisci Morales, G.; Monti, C.; and Schifanella, R. 2023. The Pursuit of Peer Support for Opioid Use Recovery on Reddit. *Proc. ICWSM*.

Birkun, A. A.; and Gautam, A. 2023. Large Language Model (LLM)-Powered Chatbots Fail to Generate Guideline-Consistent Content on Resuscitation and May Provide Potentially Harmful Advice. *Prehospital and Disaster Medicine*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Bunting, A. M.; Frank, D.; Arshonsky, J.; Bragg, M. A.; Friedman, S. R.; and Krawczyk, N. 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*.

Cai, A.; Arawjo, I.; and Glassman, E. L. 2024. Antagonistic AI. arXiv:2402.07350.

CDC. 2021. CDC's Role in the Opioid Overdose Epidemic. <https://www.cdc.gov/opioids/overdoseprevention/cdc-role.html>. Accessed: 2025-04-16.

Chancellor, S.; Baumer, E. P.; and De Choudhury, M. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proc. CSCW*.

Chancellor, S.; Nitzburg, G.; Hu, A.; Zampieri, F.; and De Choudhury, M. 2019. Discovering alternative treatments for opioid use recovery using social media. In *Proc. CHI*.

Chen, J. 2017. 3 Major Myths About Opioid Addiction. <https://www.yalemedicine.org/news/opioid-addiction-myths>. Accessed: 2025-04-16.

Choi, M.; Aiello, L. M.; Varga, K. Z.; and Quercia, D. 2020. Ten Social Dimensions of Conversations and Relationships. In *Proc. WWW*.

Corrigan, P. W. 2017. Making the most of contact to erase stigma. D'Agostino, A. R.; Optican, A. R.; Sowles, S. J.; Krauss, M. J.; Escobar Lee, K.; and Cavazos-Rehg, P. A. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and alcohol dependence*.

DAIR.AI. 2023. Chain-of-Thought Prompting: Prompt Engineering Guide. <https://www.promptingguide.ai/techniques/cot>. Accessed: 2025-04-16.

De Choudhury, M.; Pendse, S. R.; and Kumar, N. 2023. Benefits and Harms of Large Language Models in Digital Mental Health.

Dekeseredy, P.; Sedney, C. L.; Razaq, B.; Haggerty, T.; and Brownstein, H. H. 2021. Tweeting Stigma: An Exploration of Twitter Discourse Regarding Medications Used for Both Opioid Use Disorder and Chronic Pain.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Donofrio, J. 2023. Opioids: Understanding Addiction Versus Dependence. https://www.hss.edu/conditions_understanding-addiction-versus-dependence.asp. Accessed: 2025-04-16.

D'Agostino, A. R.; Optican, A. R.; Sowles, S. J.; Krauss, M. J.; Escobar Lee, K.; and Cavazos-Rehg, P. A. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and Alcohol Dependence*.

ElSherief, M.; Sumner, S.; Krishnasamy, V.; Jones, C.; Law, R.; Kacha-Ochana, A.; Schieber, L.; and De Choudhury, M. 2024. Identification of Myths and Misinformation About Treatment for Opioid Use Disorder on Social Media. *JMIR Form Res*.

ElSherief, M.; Sumner, S. A.; Jones, C. M.; Law, R. K.; Kacha-Ochana, A.; Shieber, L.; Cordier, L.; Holton, K.; and De Choudhury, M. 2021. Characterizing and identifying the prevalence of web-based misinformation relating to medication for opioid use disorder: Machine learning approach. *JMIR*.

Entman, R. M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*.

Fan, Y.; Zhang, Y.; Ye, Y.; Li, X.; and Zheng, W. 2017. Social Media for Opioid Addiction Epidemiology: Automatic Detection of Opioid Addicts from Twitter and Case Studies. In *Proc. CIKM*.

FORCE11. 2020. The FAIR Data principles. <https://force11.org/info/the-fair-data-principles/>. Accessed: 2025-04-16.

Garett, R.; and Young, S. D. 2022. The Role of Misinformation and Stigma in Opioid Use Disorder Treatment Uptake. *Substance Use & Misuse*.

Garg, S.; Taylor, J.; and et al. 2021. Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on Reddit. *Internet Interventions*.

- Geburu, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*.
- Heimer, R.; Hawk, K.; and Vermund, S. H. 2019. Prevalent Misconceptions About Opioid Use Disorders in the United States Produce Failed Policy and Public Health Responses. *Clinical infectious diseases*.
- Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proc. CHI*.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; et al. 2023. Mistral 7B. arXiv:2310.06825.
- Johns Hopkins Medicine. 2020. Building Healthier Communities: Opioid Use Disorder. <https://www.hopkinsmedicine.org/-/media/johns-hopkins-bayview/myths-about-opioid-use-disorder.pdf>. Accessed: 2025-04-16.
- Kennedy-Hendricks, A.; Barry, C. L.; Gollust, S. E.; Ensminger, M. E.; Chisolm, M. S.; and McGinty, E. E. 2017. Social stigma toward persons with prescription opioid use disorder: associations with public support for punitive and public health-oriented policies. *Psychiatric services*.
- Kruis, N. E.; McLean, K.; and Perry, P. 2021. Exploring first responders' perceptions of medication for addiction treatment: Does stigma influence attitudes? *Journal of Substance Abuse Treatment*.
- Laud, T.; Kacha-Ochana, A.; Sumner, S.; Krishnasamy, V.; Law, R.; Schieber, L.; De Choudhury, M.; and ElSherief, M. 2025. Large-Scale Analysis of Online Questions Related to Opioid Use Disorder on Reddit. *Proc. ICWSM*.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.
- MacLean, D.; Gupta, S.; Lembke, A.; Manning, C.; and Heer, J. 2015. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In *Proc. CSCW*.
- Madaan, A.; Hermann, K.; and Yazdanbakhsh, A. 2023. What Makes Chain-of-Thought Prompting Effective? A Counterfactual Study. In *Findings of EMNLP*.
- 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>. Accessed: 2025-04-16.
- NIDA. 2023. Drug Overdose Death Rates. <https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates>. Accessed: 2025-04-16.
- Omiye, J. A.; Lester, J. C.; Spichak, S.; Rotemberg, V.; and Daneshjou, R. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*.
- OpenAI. 2023a. OpenAI documentation. <https://platform.openai.com/docs/guides/text-generation/faq>. Accessed: 2025-04-16.
- OpenAI. 2023b. Prompt engineering. <https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results>. Accessed: 2025-04-16.
- OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.
- Parnami, A.; and Lee, M. 2022. Learning from Few Examples: A Summary of Approaches to Few-Shot Learning. *ArXiv*, abs/2203.04291.
- Rainie, L.; Funk, C.; Anderson, M.; and Tyson, A. 2022. How Americans Think About Artificial Intelligence. Pew Research Center, March 17, 2022.
- Reardon, S. 2023. AI Chatbots Could Help Provide Therapy, but Caution Is Needed. <https://www.scientificamerican.com/article/ai-chatbots-could-help-provide-therapy-but-caution-is-needed/>. Accessed: 2025-04-16.
- Salihefendic, A. 2015. How Reddit ranking algorithms work. <https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef11e33d0d9>. Accessed: 2025-04-16.
- SAMHSA. 2024. Statutes, Regulations, and Guidelines. <https://www.samhsa.gov/medications-substance-use-disorders/statutes-regulations-guidelines>. Accessed: 2025-04-16.
- Sharma, A.; Lin, I. W.; Miner, A. S.; Atkins, D. C.; and Althoff, T. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *Proc. WWW*.
- Smith, K. E.; Rogers, J. M.; Schriefer, D.; and Grundmann, O. 2021. Therapeutic benefit with caveats?: Analyzing social media data to understand the complexities of kratom use. *Drug and alcohol dependence*.
- Tate, S.; Fouladvand, S.; Chen, J. H.; and Chen, C.-Y. A. 2023. The ChatGPT therapist will see you now: Navigating generative artificial intelligence's potential in addiction medicine research and patient care. *Addiction*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.
- Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv:1908.08962.
- Wakeman, S. E.; and Barnett, M. L. 2018. Primary Care and the Opioid-Overdose Crisis: Buprenorphine Myths and Realities. *New England Journal of Medicine*.
- Wayne Kepner, M. C. M.; and Nobles, A. L. 2022. Types and Sources of Stigma on Opioid Use Treatment and Recovery Communities on Reddit. *Substance Use & Misuse*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; et al. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.
- WHO. 2023. Opioid Overdose. <https://www.who.int/news-room/fact-sheets/detail/opioid-overdose>. Accessed: 2025-04-16.
- Woo, J.; Bhalerao, A.; Bawor, M.; Bhatt, M.; Dennis, B.; Mouravska, N.; Zielinski, L.; and Samaan, Z. 2017. "Don't Judge a Book by Its Cover": A Qualitative Study of Methadone Patients' Experiences of Stigma. *Substance Abuse: Research and Treatment*.
- Xiang, T.; Li, L.; Li, W.; Bai, M.; Wei, L.; Wang, B.; and Garcia, N. 2023. CARE-MI: Chinese Benchmark for Misinformation Evaluation in Maternity and Infant Care. In *NeurIPS*.
- Yale Medicine. 2020. Opioid Use Disorder. <https://www.yalemedicine.org/conditions/opioid-use-disorder>. Accessed: 2025-04-16.
- Yao, H.; Rashidian, S.; Dong, X.; Duanmu, H.; Rosenthal, R. N.; and Wang, F. 2020. Detection of Suicidality Among Opioid Users on Reddit: Machine Learning Based Approach. *JMIR*.
- Zhang, A. X.; Ranganathan, A.; Metz, S. E.; Appling, S.; et al. 2018. A Structured Response to Misinformation: Defining and Annotating Credibility Indicators in News Articles. In *Proc. WWW*.
- Zhang, T.; Luo, H.; Chuang, Y.-S.; Fang, W.; Gaitskell, L.; Hartvigsen, T.; Wu, X.; Fox, D.; Meng, H.; and Glass, J. 2023. Interpretable Unified Language Checking. arXiv:2304.03728.
- Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proc. CHI*.
- Zhou, K.; Constantinides, M.; Aiello, L. M.; Joglekar, S.; and Quercia, D. 2021. The Role of Different Types of Conversations for Meeting Success. *IEEE Pervasive Computing*.
- Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can large language models transform computational social science? *Computational Linguistics*.

Paper Checklist

1. For most authors...
 - (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, as discussed in Sections “Introduction” and “Discussion,” our work advances the space of computational social science in general and online harm reduction in particular, in the context of OUD. We offer recommendations to reduce online misinformation on OUD such as, machine-assisted content moderation and automated reauthoring of harmful content.**
 - (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes. We have carefully reviewed that the claims made in abstract and Introduction accurately reflect our contributions.**
 - (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, please refer to Section “Methods.” We justify the appropriateness of our methods by referring to prior works that use them for similar tasks/objectives. We also validate the LLM-based classifiers to identify myth-promoting and myth-countering responses (refer to Sections “Evaluation of Classifiers” within “Results” and “Error Analysis of Classification” in the Appendix). Further, we contextualize our findings in the existing public health and health misinformation literature.**
 - (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes. Please refer to Section “Data” and Table 1 where we describe the characteristics of our data such as, number of Reddit posts/comments, length of posts/comments, data timeline, and the prompt to get LLM-generated response from GPT-4.**
 - (e) Did you describe the limitations of your work? **Yes, we discuss limitations of our work in Section “Limitations and Future Work.”**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes. Refer to Section “Discussion” and “Ethics Statement” where we discuss potential negative societal impacts of our work and ways to mitigate.**
 - (g) Did you discuss any potential misuse of your work? **Yes. Refer to Section “Discussion” and “Ethics Statement” where we discuss potential misuse of our work and careful considerations around those.**
 - (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? **Yes, refer to Section “Ethics Statement.” We followed best practices in social computing research and worked with deidentified data. We also**
provide carefully paraphrased excerpts in the paper to reduce traceability.
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes, we carefully read the ethics guidelines and ensured that the paper conforms to them. For instance, in Section “Ethics Statement,” we elaborate on the societal impact of our findings, data anonymization, and responsible release of paraphrased excerpts.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes. We provide the URL to an anonymized repository containing all our code and instructions to reproduce the results.**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes. In Sections “Myths on OUD: Training Data” and “Myth and Counter-myth Classification” we specify training details such as (a) source for ground truth few shot examples, (b) LLMs used as classifiers, (c) sampling temperature, and (d) prompting strategies for classification. We also provide the few-shot and few-shot with chain-of-thought prompts used in the Appendix (refer to Tables A4 and A5).**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **No, because our work was**

not compute intensive and we did not use any external storage or compute resources.

- (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes.** Please refer to Section “Methods” where we relate our proposed evaluation procedures to the research questions. Also, refer to Section “Discussion” in which we justify that the evaluation/findings are appropriate with respect to our claims, research goals, and prior work.
- (f) Do you discuss what is “the cost” of misclassification and fault (in)tolerance? **Yes.** We evaluate the LLM-based classifiers, as elaborated in Section “Evaluation of Classifiers”. The evaluation is two-fold: (1) we summarize classification performance via evaluation metrics (refer to Tables 3 and A8) and (2) we perform error analysis to elaborate on the best performing classifier’s (gpt-3.5-turbo with a few-shot chain-of-thought prompt) false negative and false positive instances, in addition to the implications of such misclassifications (refer to Section “Error Analysis of Classification”) in the Appendix.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity...**
- (a) If your work uses existing assets, did you cite the creators? **Yes,** refer to Section “Data” where we cite the creators of the `Reddit-QA` dataset.
- (b) Did you mention the license of the assets? **NA.** The `Reddit` dataset used in this work is provided by the authors who curated it (Laud et al. 2025).
- (c) Did you include any new assets in the supplemental material or as a URL? **NA**
- (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **No,** because we utilized a publicly accessible retrospective `Reddit` dataset. The `Reddit` dataset used in this work is provided by the authors who curated it (Laud et al. 2025). Analyzing this retrospective data did not constitute human subjects research and thus informed consent was not required from the authors of the news articles. However, we followed best practices, worked with deidentified data and refrained from revealing any identifiable data to avoid potential harm to those who authored or were referenced in the comments (see Section “Ethics Statement”).
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes,** we discuss best practices followed in our work in Section “Ethics Statement”. These include working with deidentified data and paraphrasing quotes throughout the paper to reduce traceability. We also provide a content warning in the “Introduction”.
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? **NA**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity...**
- (a) Did you include the full text of instructions given to participants and screenshots? **NA**
- (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **No.** We utilize publicly accessible `Reddit` dataset (refer to Section “Data”) without any interaction with the authors of posts or comments and the people referenced in them. As an observational study of retrospectively gathered data, our research did not qualify as human subjects research, per our Institutional Review Board guidelines. Thus, IRB approval was not required. However, ethical considerations extend beyond IRBs, when it comes to web research. Therefore, we used deidentified data and provided paraphrased quotes in the paper to avoid traceability and potential harm to those who authored the comments (refer to Section “Broader Perspectives, Ethics, and Competing Interests”).
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
- (d) Did you discuss how data is stored, shared, and deidentified? **Yes.** As mentioned in Section “Ethics Statement”, we follow best practices proposed by Chancellor, Baumer, and De Choudhury (2019) to work with deidentified data, to use secure machines for our analyses with only authorized access to the paper’s authors granted through the principle of least privilege, and to avoid sharing personally identifiable data in any form.

Ethics Statement

Following best practices (Chancellor, Baumer, and De Choudhury 2019), we worked with deidentified publicly accessible data, and refrained from sharing raw and personally identifiable data in any form. All the quotes, taken from `Reddit`, referenced in this work were paraphrased to reduce traceability and potential harm to those who authored or were referred to in the comments. Given that we found prevalence of myths in the LLM-generated responses, public health considerations should be made before deploying them at scale in this context. Further, given the classifiers’ misclassifications, they should be carefully integrated into `Reddit` (e.g., to facilitate content moderation) and other social media platforms. We make note that conducting research on such a socially stigmatizing issue can have significant emotional labor and strongly encourage future researchers in this area to prioritize their well-being. Additional ethical considerations are provided in the Paper Checklist. The authors declare no competing interests.

You are a Reddit user. You actively browse through different subreddits, which are online communities on Reddit, to gather health-related information. While browsing through these communities, you also frequently answer posts sharing opinions or information on opioid use disorder. Write a comment, which could be posted on the subreddit *[subreddit name]*, answering the following question taken from the same subreddit. For context, the *[subreddit name]* subreddit is described as *[subreddit description]*.

Question: *[post from Reddit-QA-posts]*

Table A1: Prompt to generate LLM responses to queries on OUD using GPT-4.

Appendix

Prompt to get LLM-generated Responses

Table A1 details the hand-crafted prompt used in this work to get LLM-generated responses to queries on OUD via GPT-4.

Prompt Exploration

We explored multiple prompts, with varied rationales, to gather LLM-generated responses. Table A2 lists these prompts. For the purpose of demonstration, we present GPT-4 generated responses to the following paraphrased Reddit post: “Once I am clean, how should I go about telling people about my past addiction to heroin?”

By providing the question as is, GPT-4 generated a bulleted list, which is unlikely to be posted as a comment on Reddit. We referred to prompt engineering guidelines (OpenAI 2023b) that recommend to include (a) simple task-specific instructions, (b) context, (c) question of interest, and (d) output format. Following this, we crafted the second prompt, assigning GPT-4 the persona of a social media user, and elaborating on the task. The generalized role assignment resulted in comments suitable for Twitter – short responses with hashtags. We then explicitly assigned it the role of a Reddit user and highlighted that the generated comment could be posted on a subreddit. This did result in responses suitable for Reddit, however, the comments were not personalized, candid, or actionable – characteristics of content on Reddit (Chancellor et al. 2019). Therefore, in our final prompt, we provided the subreddit’s name and description for additional context. The prompt resulted in responses containing first-person language, personalized support, and community engagement – semantically similar to comments within the *Reddit-QA-comments* dataset.

Examples within the Dataset

Table A3 lists a few paraphrased queries, i.e., Reddit posts containing an OUD-related query within the *Reddit-QA-posts* dataset. The queries (or posts) cover a variety of topics ranging from drug use to disclosure. We also provide the corresponding human-, i.e., the top-most voted comment received by the post (*Reddit-QA-comments* dataset) and LLM-generated response (obtained as described in Section “LLM-generated

Responses to OUD queries”). The human- and LLM-generated responses are semantically similar in the context of OUD. For example, for the first exemplar query, both the responses elaborate on potential excuses the Reddit user could provide to their supervisor (“suffering from short term insomnia” vs. “choose something commonly associated with causing stomach issues”). Interestingly, the third query had contextually different human and LLM responses – response from Reddit explicitly described experience with heroin, while the one generated by GPT-4 interpreted and answered the question through a self-reflection lens. Nonetheless, across all the three queries, human- and LLM-generated responses used similar linguistics (e.g., first-person language), and offered support, and guidance.

Prompts for Myth and Counter-myth Classification

Tables A4 and A5 list few-shot and few-shot with chain-of-thought prompts, respectively, to classify whether or not a human- and LLM-generated response promotes M1, i.e., *agonist therapy or medication assisted treatment for opioid use disorder is merely replacing one drug with another*. A similar template was used for the other 4 myths by making appropriate changes. We also used a similar few-shot and few-shot with chain-of-thought prompt to classify whether or not a human- and LLM-generated response *counters* a myth. The task was changed accordingly to “determine whether the following response counters or challenges the myth.” We referred to prior work and prompt engineering guidelines for creating these prompts (Antoniak et al. 2024; OpenAI 2023b; DAIR.AI 2023).

Rationale for Classifier Choices

Here, we elaborate on the rationale behind the choices we made to build the LLM-based classifiers (described in Section “Myth and Counter-myth Classification” in the main paper). We did not include a zero-shot prompting strategy, to guide classification, given the nuanced and intricate nature of our task. This choice is in line with prior work (Zhang et al. 2023), which found that few-shot learning outperformed the zero-shot setting for misinformation, toxic, and hate speech detection. As for our selection of LLMs, we wanted to include both open-source and proprietary LLMs, hence the choice of LLaMa-2-13B (open-source), Mistral-7B (open-source), and gpt-3.5-turbo (proprietary). These LLMs are also widely used and adopted in LLM benchmarking (OpenAI et al. 2024).

Description of Framing Strategies and Computational Methods

As described in Section “Framing of Myths and Counter-myths” in the main paper, we used 4 social dimension classifiers (Trust, Power, Conflict, and Knowledge) and Story-Seeker to identify the framing strategies. We carefully considered descriptions of the 4 strategies to select relevant computational methods:

- Assertive confidence: Expressions demonstrating certainty, trust, and authority (Table A6). The Trust and

Prompt	LLM-generated response
Please answer the following question. Question: <i>[post from Reddit-QA-posts]</i>	Telling people about your past addiction is a personal choice. [...] Here are a few things to consider: <ul style="list-style-type: none"> • Assess Your Readiness: [...] • Choose the Right Time and Place: [...] • ...
You are a social media user . You actively browse through online platforms to gather health-related information. While browsing, you also frequently answer posts sharing opinions or information on opioid use disorder. Write a comment, which could be posted on a social media platform, answering the following question. Question: <i>[post from Reddit-QA-posts]</i>	Major kudos for taking the brave step towards recovery! Opening up is a personal choice, and the manner of sharing should align with your comfort level. #RecoveryJourney #BreakingStigmas
You are a Reddit user . You actively browse through different subreddits, which are online communities on Reddit, to gather health-related information. While browsing through these communities, you also frequently answer posts sharing opinions or information on opioid use disorder. Write a comment, which could be posted on a subreddit, answering the following question. Question: <i>[post from Reddit-QA-posts]</i>	[...] You could start by joining local support groups. You are not defined by your past. All the best for your journey!
You are a Reddit user . You actively browse through different subreddits, which are online communities on Reddit, to gather health-related information. While browsing through these communities, you also frequently answer posts sharing opinions or information on opioid use disorder. Write a comment, which could be posted on the subreddit <i>[subreddit name]</i> , answering the following question taken from the same subreddit. For context , the <i>[subreddit name]</i> subreddit is described as <i>[subreddit description]</i> . Question: <i>[post from Reddit-QA-posts]</i>	[...] Consider starting with those closest to you, like your friends whom you trust the most. Proceed at a pace that feels comfortable for you. Like me, you may emphasize the positive aspects of your recovery and the strength it took to overcome such a challenging time. [...] You never know who might be silently struggling, and your experience could inspire someone here.

Table A2: Different prompts explored to get the GPT-4 generated response

Power classifiers are relevant proxies as they capture trustworthy and authoritative linguistic cues (Table A7).

- Alternate diverging views: Reasoning other views (Table A6). The Conflict classifier captures this well as it is trained to identify contrasting viewpoints (Table A7).
- External knowledge: Exchange of expert know-how (Table A6). The Knowledge classifier is relevant for this strategy as it is trained to identify exchange of ideas, information, or insights (Table A7).
- Persuasive appeal: Narratives of lived experiences (Table A6). StorySeeker (Antoniak et al. 2024) fits well here as it can recognize texts containing personalized storytelling narratives.

Validation of the Framing Tools:

To validate the 4 social dimension classifiers, for Trust, Power, Conflict, and Knowledge, we considered texts within the `Reddit-QA-comments` and the `GPT-4-responses` datasets that received the highest (top 10) and the lowest (bottom 10) alignment scores from the classifiers. For example, to validate the classifier for Trust, we considered the top 10 and the bottom 10 responses, spanning across the `Reddit-QA-comments` and the `GPT-4-responses` datasets, as scored by the Trust social dimension classifier.

The first two authors manually annotated these responses to label whether they displayed high or low alignment to the social dimensions. Both the authors have prior experience in social computing research. Before annotation, they reached a common understanding of the 4 social dimensions. In total, the authors annotated 20*4 responses – 20 each for the 4 dimensions. For inter-rater reliability, the authors received a Cohen’s Kappa of 0.80, 0.70, 0.70, and 0.80 for Trust, Power, Conflict, and Knowledge respectively. Following this, the authors discussed their annotations to reach 100% consensus. We then compared the human-assigned labels against those provided by the classifiers. The classifiers achieved the following precision (pr), recall (re), and F1-score (F1) metrics: (a) pr: 0.85, re: 0.88, F1: 0.85 for Trust, (b) pr: 0.75, re: 0.77, F1: 0.74 for Power, (c) pr: 0.85, re: 0.88, F1: 0.85 for Conflict, and (d) pr: 0.75, re: 0.75, F1: 0.75 for Knowledge.

We similarly validated the StorySeeker tool. We randomly sampled 20 responses spanning across the `Reddit-QA-comments` and the `GPT-4-responses` datasets. The first two authors annotated them to indicate whether they contained a storytelling narrative or not. The authors received a Cohen’s Kappa of 0.90. Following this, they resolved conflicts to reach 100% consensus. On comparing the manual annotations against StorySeeker’s labels, we got 0.85 precision, 0.85 recall, and 0.85 F1-score. This analysis also helped us confirm that the classifiers were rel-

Query	Human-generated response	LLM-generated response
I started falling asleep at work today and got sent home. I know I am going to have to talk to my supervisor tomorrow. [...] I am thinking I am going to say I got food poisoning. Can anyone please help me think of a common food that could have been left out for too long? (<i>paraphrased from r/Opiates</i>)	Just tell your supervisor that you have been suffering from some short term insomnia and were running on almost no sleep. [...] God, us addicts sure learn quickly to be good lairs! (<i>paraphrased from r/Opiates</i>)	It sounds like you're in a difficult situation, and need to come up with an explanation. [...] It might be helpful to choose something that's commonly associated with causing stomach issues when left out, like potato salad or chicken.
Once I am clean, how should I go about telling people about my past addiction to heroin? (<i>paraphrased from r/OpiatesRecovery</i>)	I tell people about my past addictions when I feel comfortable. [...] If they think less of me for addictions, they don't need to be my friends. I don't drop all my secrets until I'm ready, but being vague works for me most of the time. I'm not ashamed of who I am, and I won't let anyone put shame on me. (<i>paraphrased from r/OpiatesRecovery</i>)	[...] Consider starting with those closest to you, like your friends whom you trust the most. Proceed at a pace that feels comfortable for you. Like me, you may emphasize the positive aspects of your recovery and the strength it took to overcome such a challenging time.
What got you to try heroin for the first time and what keeps you to continue using it? (<i>paraphrased from r/Heroin</i>)	When I first tried heroin I instantly fell in love with it. [...] I continue to use it today because I've had numerous other surgeries and suffer from severe chronic pain. I use it nowadays for therapeutic needs. (<i>paraphrased from r/Heroin</i>)	Reflecting on reasons behind trying heroin and continuing its use is a step toward self-awareness. [...] While this subreddit may offer a space for discussions, I use it to explore resources that can support in making healthier choices.

Table A3: Exemplars from Reddit-QA and GPT-4-responses datasets. Query: paraphrased post, taken from Reddit-QA-posts, containing an OUD-related query; Human-generated response: paraphrased top-most voted comment, taken from Reddit-QA-comments, received by the query post; LLM-generated response: GPT-4 generated response to the query post.

Few-shot prompt

You are an expert in social media analysis and opioid use disorder. First, please analyze the following examples where we indicate if a response actively promotes or restates a myth on opioid use disorder.

The myth is: *Agonist therapy or medication assisted treatment for opioid use disorder is merely replacing one drug with another.*

Examples:

Following are five examples, one on each line. Each example contains the *response* and a binary YES or NO *answer* indicating whether the response contains the myth or not.

Response 1: [POSITIVE INSTANCE TEXT]; Answer: YES

Response 2: [POSITIVE INSTANCE TEXT]; Answer: YES

Response 3: [POSITIVE INSTANCE TEXT]; Answer: YES

Response 4: [NEGATIVE INSTANCE TEXT]; Answer: NO

Response 5: [NEGATIVE INSTANCE TEXT]; Answer: NO

Task: Given what you learned from the examples, your task is to determine whether the following response actively promotes the myth: *Agonist therapy or medication assisted treatment for opioid use disorder is merely replacing one drug with another.* Answer with a binary YES or NO. Answer NO if the response's author only talks about their personal experience with opioid use disorder or an issue other than the provided myth.

Do not restate the response and only provide one answer.

Response: [TEST INSTANCE TEXT]

Answer:

Table A4: Few-shot prompt to determine whether or not human- and LLM-generated responses promote OUD myths.

evant for the framing strategies.

Evaluation of Counter-myth Classifiers

Table A8 provides an overview of how the 3 LLM-based classifiers – LLaMa-2-13B, Mistral-7B, and

Few-shot with Chain-of-Thought prompt

You are an expert in social media analysis and opioid use disorder. First, please analyze the following examples where we indicate if a response actively promotes or restates a myth on opioid use disorder.

The myth is: *Agonist therapy or medication assisted treatment for opioid use disorder is merely replacing one drug with another.*

Examples:

Following are five examples, one on each line. Each example contains the *response* and a binary YES or NO *answer* indicating whether the response contains the myth or not. We also provide a *reasoning* for why the response contains or does not contain the myth.

Response 1: [POSITIVE INSTANCE TEXT]; Answer: YES; Reasoning: [REASONING TEXT]

Response 2: [POSITIVE INSTANCE TEXT]; Answer: YES; Reasoning: [REASONING TEXT]

Response 3: [POSITIVE INSTANCE TEXT]; Answer: YES; Reasoning: [REASONING TEXT]

Response 4: [NEGATIVE INSTANCE TEXT]; Answer: NO; Reasoning: [REASONING TEXT]

Response 5: [NEGATIVE INSTANCE TEXT]; Answer: NO; Reasoning: [REASONING TEXT]

Task: Given what you learned from the examples, your task is to determine whether the following response actively promotes the myth: *Agonist therapy or medication assisted treatment for opioid use disorder is merely replacing one drug with another.* Answer with a binary YES or NO. Answer NO if the response’s author only talks about their personal experience with opioid use disorder or an issue other than the provided myth.

Also provide a reasoning for your answer, quoting excerpts from the response.

Do not restate the response and only provide one answer. *Please think through step by step.*

Response: [TEST INSTANCE TEXT]

Answer:

Reasoning:

Table A5: Few-shot with Chain-of-Thought prompt to determine whether or not human- and LLM-generated responses promote OUD myths.

Framing strategy	Definition	Method
Assertive confidence	Expressions demonstrating certainty, trust, and authority (Zhang et al. 2018)	Trust, Power classifiers (Choi et al. 2020)
Alternate diverging views	Reasoning other possible standpoints (Zhang et al. 2018)	Conflict classifier (Choi et al. 2020)
External knowledge	Exchange of ideas or expert know-how (Zhang et al. 2018)	Knowledge classifier (Choi et al. 2020)
Persuasive appeal	Descriptions of lived experiences (Zhang et al. 2018)	StorySeeker (Antoniak et al. 2024)

Table A6: Strategies and corresponding methods to study how responses that promote and counter myths differ in framing.

Social dimension classifier	Description	Representative keywords
Trust	Will of relying on the actions or judgments of another	trustworthy, loyalty, faith, dependability
Power	Having power over the behavior and outcomes of another	authority, control, command, dominance
Conflict	Contrast or diverging views	different, divide, tense, disagreement
Knowledge	Exchange of ideas or information; learning, teaching	expertise, know-how, insight, competent

Table A7: Description of social dimension classifiers developed by Choi et al. (2020).

gpt-3.5-turbo – performed on the task of identifying responses that countered the 5 OUD-related myths. Similar to evaluating the myth classifiers (refer to Section “Evaluation of Classifiers” in the main paper), we randomly sampled 50 responses each from the *Reddit-QA-comments* and *GPT-4-responses* datasets, and manually labeled them: ‘0’ to indicate that the response does not counter the myth and ‘1’ to indicate it does². The first two authors initially annotated 20 responses and reached 100% agreement. Per inter-rater reliability, the annotators received a

Cohen’s Kappa score of 0.84 (M1), 0.74 (M2), 0.76 (M3), 0.70 (M4), and 0.73 (M5). Following this, the first author labeled the remaining 80. Annotations were informed by expert-annotated examples described in Section “Myths on OUD: Training Data” and substance use literature (Heimer, Hawk, and Vermund 2019). Finally, we compared binary labels assigned by the LLM-based classifiers against our ground truth hand-annotations. The *LLaMa-2-13B* counter-myth classifier achieved the lowest classification performance (the lowest F1-score) across all 5 myths and both the prompting strategies (Table A8). For example,

²This was repeated for all 5 myths.

Myth	LLaMa-2-13B			Mistral-7B			gpt-3.5-turbo			LLaMa-2-13B			Mistral-7B			gpt-3.5-turbo			McNemar's	
	Few-shot prompt									Few-shot with Chain-of-Thought prompt									χ^2	p
	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1		
M1	0.66	0.66	0.66	0.70	0.68	0.69	0.82	0.80	0.81	0.68	0.66	0.67	0.74	0.70	0.72	0.88	0.88	0.88	6.02	***
M2	0.60	0.60	0.60	0.65	0.62	0.63	0.76	0.70	0.73	0.64	0.63	0.63	0.68	0.66	0.67	0.82	0.78	0.80	4.28	**
M3	0.64	0.60	0.62	0.68	0.66	0.67	0.80	0.78	0.79	0.64	0.64	0.64	0.72	0.72	0.72	0.85	0.85	0.85	4.09	**
M4	0.65	0.62	0.63	0.68	0.68	0.68	0.84	0.84	0.84	0.66	0.66	0.66	0.72	0.70	0.71	0.88	0.85	0.86	2.72	*
M5	0.65	0.65	0.65	0.66	0.66	0.66	0.70	0.68	0.69	0.66	0.66	0.66	0.70	0.66	0.68	0.78	0.78	0.78	7.14	***

Table A8: Evaluation of LLaMa-2-13B, Mistral-7B, and gpt-3.5-turbo as classifiers to identify whether a human- or LLM-generated response counters an OUD-related myth. Performance summarized for two prompting strategies: (a) *few-shot* and (b) *few-shot with chain-of-thought* via 3 metrics: precision (Pr.), recall (Re.), and F1-score (F1). McNemar’s tests ($p < 0.1$: ‘*’, $p < 0.05$: ‘**’, $p < 0.01$: ‘***’) were performed to compare gpt-3.5-turbo’s performance for few-shot and few-shot with chain-of-thought prompting strategies.

while identifying responses countering M4 (*only patients with certain characteristics are vulnerable to addiction*) via the few-shot prompt, LLaMa-2-13B got an F1-score of 0.63, compared to 0.68 for Mistral-7B and 0.84 for gpt-3.5-turbo. Again, similar to myth classification, few-shot with chain-of-thought prompt improved classification performance over few-shot prompt. gpt-3.5-turbo increased F1-score by 2.38% (from 0.84 to 0.86), while identifying responses that countered M4, using a few-shot with chain-of-thought prompt. gpt-3.5-turbo with a few-shot chain-of-thought prompt achieved the best counter-myth classification performance across all classifiers and myths. It was the most competent in identifying responses that countered M1 (F1-score: 0.88) and the least competent in M5 (F1-score: 0.78).

Error Analysis of Classification

We did an error analysis to see where the best-performing classifier, gpt-3.5-turbo with a few-shot chain-of-thought prompt, erred. For this, we sampled responses from the hand-annotated validation datasets that were flagged as false negatives and false positives.

Domain-specific abbreviations. Consider the following false negative of the gpt-3.5-turbo classifier when identifying responses that promote M3 (*the ultimate goal of treatment for OUD is abstinence from any opioid use*).

“Get off maintenance drugs and CT. Go into sober living, you’ll be free of opioids.” (*paraphrased Reddit response*)

This paraphrased Reddit comment makes the use of an abbreviation, “CT”, which stands for “Cold Turkey”. Cold turkey³ is a quick-fix method to quitting drugs in which rather than gradually tapering off the substance, one stops taking it immediately. As a result, by referring to “CT”, the author advocates for complete abstinence from any opioid use to treat OUD, thereby promoting M3. This suggests that the LLM-based classifier is not well-versed in OUD-specific slangs and layman terminology.

Incorrectly capturing explicit references. The GPT-4 generated excerpt below was flagged as a false positive

³<https://www.healthline.com/health/opioid-withdrawal/quitting-cold-turkey>

when identifying responses that promote M5 (*physical dependence or tolerance is the same as addiction*).

“[...] addiction can cause dependence, where the desire to use becomes a neurological compulsion.” (*GPT-4 response*)

Due to an explicit reference to “dependence”, it incorrectly labeled the above excerpt to promote M5. A “physically dependent [patient] is capable of managing impulses and has control over their use” (Donofrio 2023). Therefore, “dependence” is not referring to *physical dependence* but *psychological dependence* experienced by people with OUD: part of the brain that guides self-control is directly impeded (“desire to use becomes a neurological compulsion”).

Framing of experiences. The paraphrased Reddit response below was flagged as a false positive by the classifier while identifying responses that counter M1 (*agonist therapy or MAT is merely replacing one drug with another*).

“[...] the way Methadone works, the body becomes dependent on it. People end up staying on Methadone forever, and that’s okay!” (*paraphrased Reddit response*)

The response actually promotes the mythical narrative that MAT, Methadone, replaces one drug with another (“the body becomes dependent on it”, “end up staying on it forever”). However, the author’s positive sentiment associated with Methadone use (“that’s okay”) potentially caused the classifier to label the response as one that counters M1.

RAG-based Classification

In addition to few-shot learning, we explored a retrieval augmented generation (RAG) based approach (Lewis et al. 2020) for classification. We first created a knowledge base consisting of relevant external sources of information. For our task, we gathered fact sheets, guidelines, and personnel training resources on opioid use disorder. These are published by credible authorities such as World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and SAMHSA. In total, we collected 40 such resources – e.g., fact sheet from WHO (WHO 2023), clinician training guide from the U.S. Department of Health and Services (Agency for Healthcare Research and Quality 2000), and fact sheet from Yale Medicine (Yale Medicine 2020). We then parsed all the sources and divided the documents into sentences. Following this, we used BERT (Devlin et al.

2019) to convert the sentences into meaningful embeddings for knowledge retrieval.

Counter-myth classification: We used a negation of the myth statements (Table 2) as queries for knowledge retrieval. For example, for M1, we used “Agonist therapy or MAT for opioid use disorder *does not* merely replace one drug with another” as the query. The query statements were also embedded using BERT. We then retrieved sentences from the knowledge base that got a high cosine similarity score (> 0.8) to the query embedding. These retrieved sentences were then fed as additional context to the `gpt-3.5-turbo` model for classifying whether the human- and LLM-generated responses countered a myth.

Myth classification: We used the original myth statements (Table 2) as queries for knowledge retrieval. The following procedure was the same as described above.

Classification performance was evaluated on the same ground truth data as described in Section “Evaluation of Classifiers” in the main paper. Tables A9 and A10 summarize this evaluation. RAG did not result in significant performance gains over few-shot classification (under McNemar’s tests). There was limited linguistic variation in the retrieved sentences. For example, for the counter-myth classification of M1, some of the retrieved sentences were “MAT does not substitute one drug for another,” “MAT is not trading one addiction for another,” and “MAT is not the same as substituting one addictive drug for another.” The sentences also did not contain Reddit-like linguistic cues. Lack of linguistic variability and creativity within the external knowledge base could be potentials reason for no significant performance gains.

Myth	Pr.	Re.	F1
M1	0.80	0.79	0.79
M2	0.75	0.75	0.75
M3	0.82	0.77	0.79
M4	0.85	0.80	0.82
M5	0.70	0.70	0.70

Table A9: RAG-based counter-myth classification.

Myth	Pr.	Re.	F1
M1	0.82	0.78	0.79
M2	0.77	0.73	0.75
M3	0.80	0.80	0.80
M4	0.78	0.77	0.77
M5	0.75	0.74	0.74

Table A10: RAG-based myth classification.

(both human- and LLM-generated) promoting myths had a higher alignment to trust, power, and conflict dimensions (distributions skew towards the right). This is consistent across all myths. Therefore, responses promoting M2, M3, M4, and M5 had a higher tendency to contain assertive confidence and the alternate diverging views framing strategies. Conversely, responses (both human- and LLM-generated) countering M2, M3, M4, and M5 contained more external knowledge based framings, having a higher alignment to the knowledge social dimension.

Additional Density Plots for RQ2 (a)

Figures A1, A2, A3, and A4 represent density distribution plots for (a-d) human- and (e-h) LLM-generated responses promoting and countering M2, M3, M4, and M5, respectively. As noted in our discussion of RQ2 (a), responses

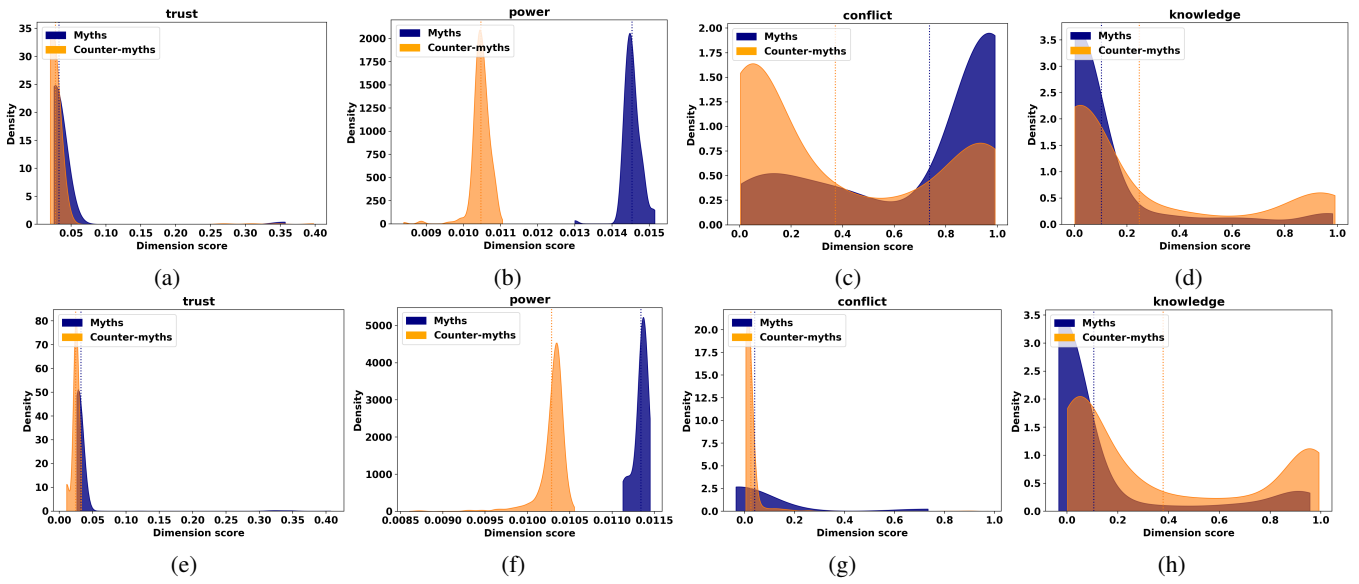


Figure A1: Density plot distributions for scores indicating the alignment of (a-d) human- and (e-h) LLM-generated responses, promoting or countering M2, to the four social dimensions: trust, power, conflict, and knowledge. (All) Dotted lines represent the average alignment score of the responses to the social dimensions.

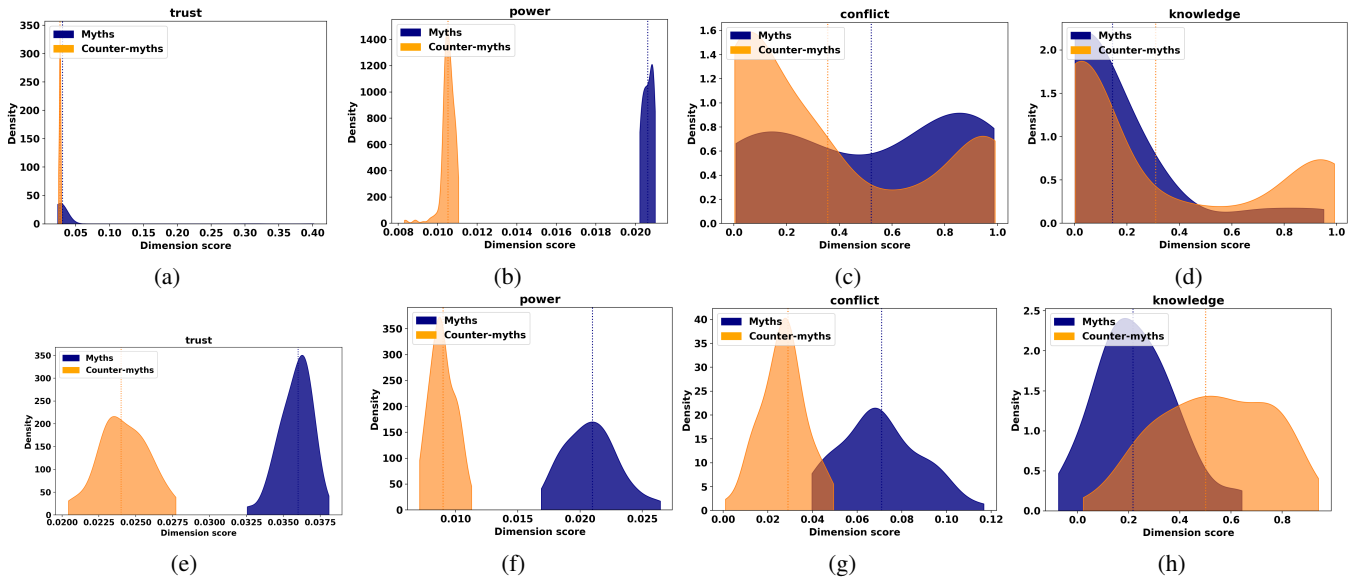


Figure A2: Density plot distributions for scores indicating the alignment of (a-d) human- and (e-h) LLM-generated responses, promoting or countering M3, to the four social dimensions: trust, power, conflict, and knowledge. (All) Dotted lines represent the average alignment score of the responses to the social dimensions.

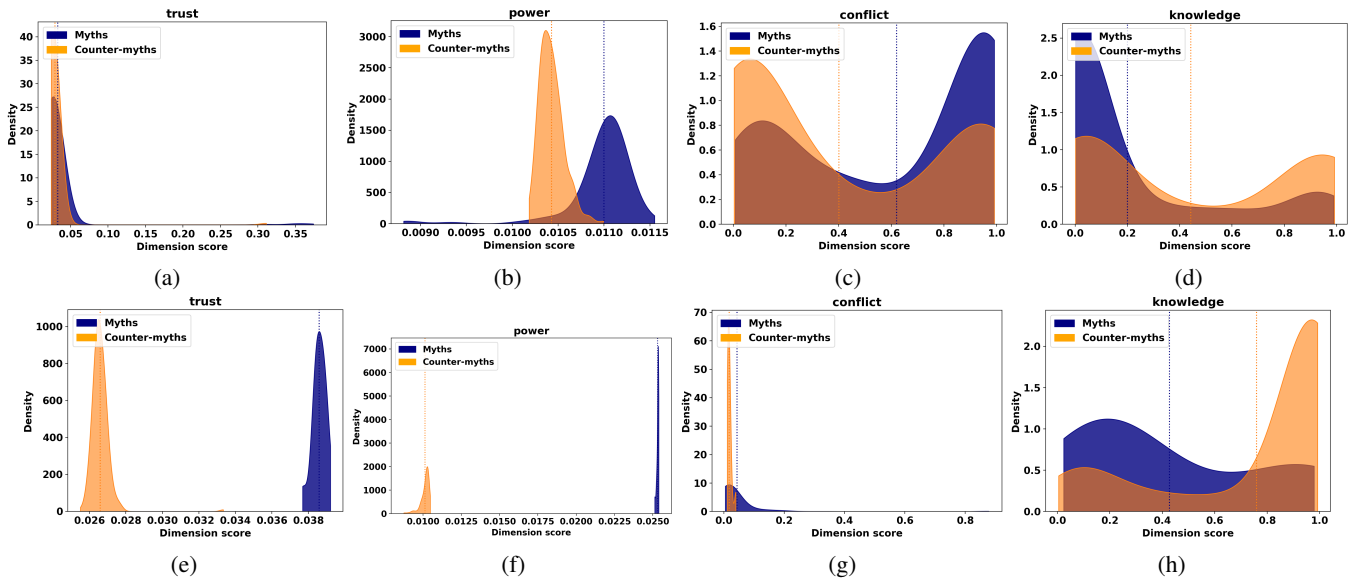


Figure A3: Density plot distributions for scores indicating the alignment of (a-d) human- and (e-h) LLM-generated responses, promoting or countering M4, to the four social dimensions: trust, power, conflict, and knowledge. (All) Dotted lines represent the average alignment score of the responses to the social dimensions.

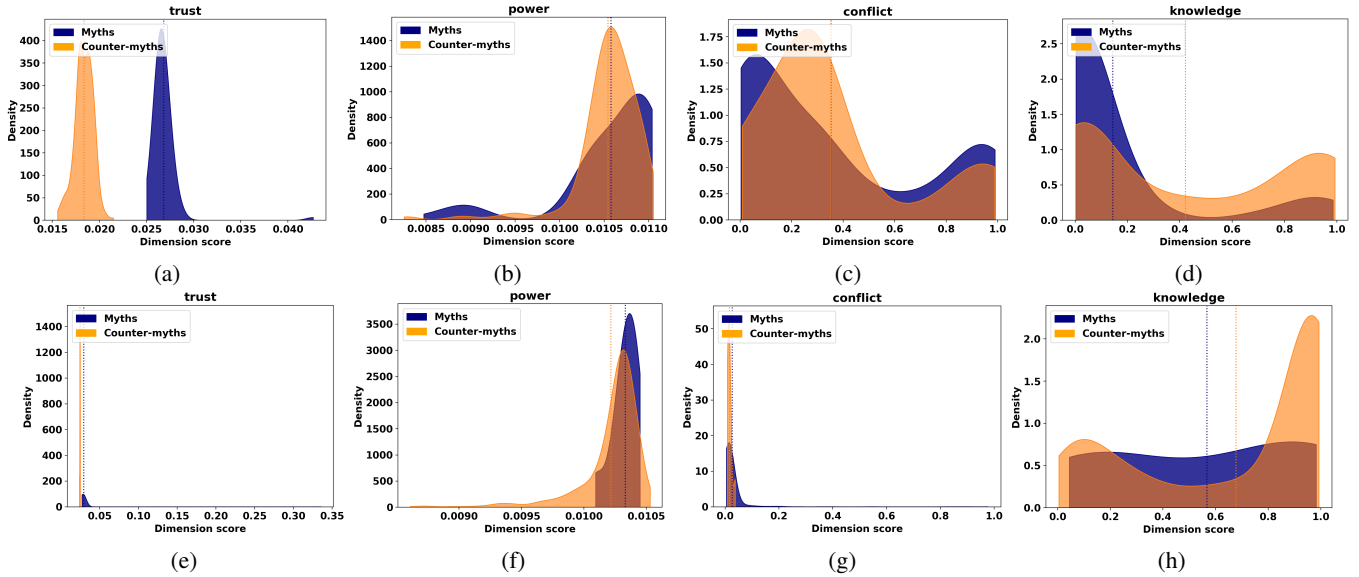


Figure A4: Density plot distributions for scores indicating the alignment of (a-d) human- and (e-h) LLM-generated responses, promoting or countering M5, to the four social dimensions: trust, power, conflict, and knowledge. (All) Dotted lines represent the average alignment score of the responses to the social dimensions.