MythTriage: Scalable Detection of Opioid Use Disorder Myths on a Video-Sharing Platform

Hayoung Jung, Shravika Mittal, Ananya Aatreya, Navreet Kaur, Munmun De Choudhury, Tanushree Mitra

Content Warning

Link to Paper

Background + Motivation

More than **108K** opioid drug-involved deaths in the United States [1]

Barriers to Harm Reduction

Myths around treatment, people with the opioid use disorder (OUD) [2]

Myths: Inaccurate and potentially harmful beliefs [3]

[1] NIDA. 2023. Drug Overdose Death Rates. https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates.
[2] Garett, R.; and Young, S. D. 2022. The Role of Misinformation and Stigma in Opioid Use Disorder Treatment Uptake. Substance Use & Misuse.
[3] 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.

Background + Motivation

Online platforms as alternatives for health information, especially in public health crises

Video-sharing Platforms
e.g., YouTube

Health Information

Recovery guidance

Peer Support

However, online myths towards OUD fuels hesitancy towards medication for addiction treatment (MAT), distrust in public health, and stigma

Background + Motivation

Understanding the prevalence of myths is crucial
for public health interventions

However, detecting misinformation on video platforms ***at scale*** is challenging.

- Requires domain expertise + intensive labeling. Very Expensive.

While LLMs show potential to address this scale challenge…

- High API cost on long-form video content limits their widespread use. Expensive.

MythTriage

Partnered with Clinical Experts

GOAL: Scalably measure the prevalence of Opioid Use Disorder Myths on a Video-Sharing Platform

YouTube Search + Recommendation

Methodology

1 Select Topics and Queries

Google Trends + YouTube Autocomplete

2 Data Collection

YouTube Search + Recommendations Dataset

3 Data Labeling for OUD-Related Myths

Collaborate with Clinical Experts for Large-Scale Labeling

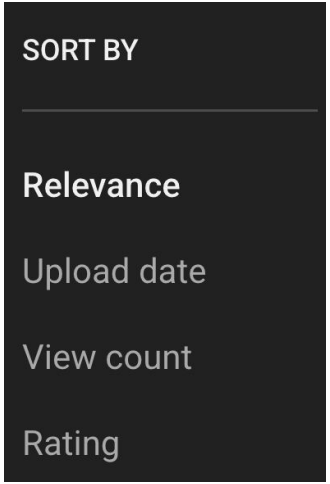
1 Select Topics and Queries

- We used Google Trends to identify 8 popular opioid and MAT topics.
- For each topic, we collected the top-10 trending search queries and top-10 YouTube Autocomplete suggestions, filtering out irrelevant and similar queries.

2 Data Collection: OUD Search Dataset

For each curated queries…

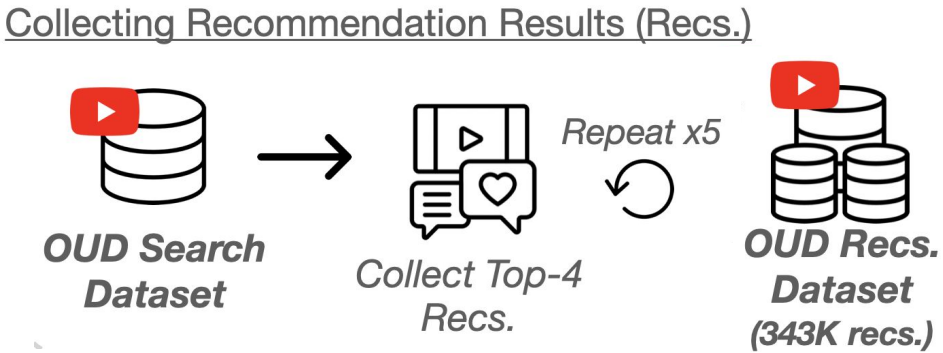
1. Search query
2. Collect top-10 results
3. Change search filters & Repeat #2
4. Repeat #1-3 for the next query



2 Data Collection: OUD Recommendation Dataset

Starting from each video in the OUD Search Dataset…

1. Recursively collect top-4 recommended videos per unique video
2. Repeat this process for 5 levels, collecting top-4 recommendations from the prior level.



3 Data Labeling for OUD-Related Myths

Collaborated with clinical experts to

- Identify 8 OUD-related myths
- Develop the Data Annotation Guidelines. Three classes:
  - Opposing OUD Myth(-1)
  - Neither (0)
  - Supporting OUD Myth (1)
- Create the Expert-Labeled Gold Standard Dataset

3 Data Labeling for OUD-Related Myths

8 Myths surrounding patient characteristics and treatment models. Recognized by major health organizations and validated by clinical experts.

Myth 1 (M1): Medication-assisted treatments (MAT) are merely replacing one drug with another

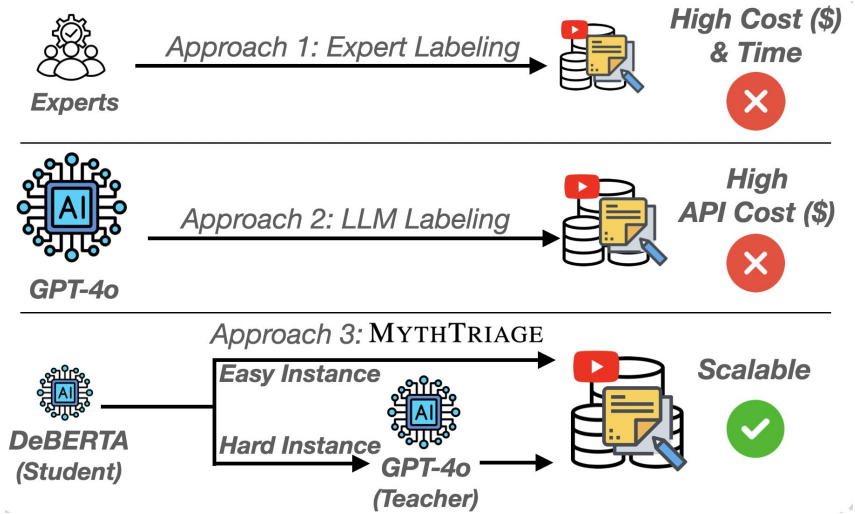
"Being on [suboxone] and you know… **it is an opioid so I don't count that as clean time**."
— **(supports M1)**

"Buprenorphine… was later adopted for treatment of opioid use disorder because… **it was so helpful in treating addiction.**" — **(opposes M1)**

3 Data Labeling for OUD-Related Myths

Introducing **MythTriage**, a cascaded triage classification pipeline**.**

- Using lightweight model distilled on GPT-4o-generated labels for routine cases
- Uses confidence-based deferral to route harder cases to a more accurate, but costlier LLM



3 Data Labeling for OUD-Related Myths

Results: Evaluated on the ~2.5K Expert Labels.

Findings: MythTriage provides a competitive performance while estimated to reduce annotation time and financial cost by over 76% compared to expert and full LLM labeling. Scalable solution for annotating large datasets.

Myth	GPT-4o	DeBERTA	MythTriage
M1	**0.87** (1)	0.77 (0)	0.86 (0.60)
M2	**0.85** (1)	0.70 (0)	0.80 (0.57)
M3	**0.86** (1)	0.76 (0)	**0.86** (0.67)
M4	**0.82** (1)	0.62 (0)	0.76 (0.31)
M5	**0.82** (1)	0.60 (0)	0.68 (0.28)
M6	**0.86** (1)	0.76 (0)	0.83 (0.52)
M7	**0.85** (1)	0.74 (0)	0.81 (0.44)
M8	**0.87** (1)	0.78 (0)	0.81 (0.05)

- Left number indicates Macro F1-score
- Parentheses indicate the proportion of examples handled by GPT-4o (lower is better)

Prevalence of Myths in YouTube Search Results

Findings

Distribution of labels across myths + overall

Label	M1	M2	M3	M4	M5	M6	M7	M8	Over.
Oppose	0.15	0.23	0.14	0.16	0.11	0.16	0.11	0.04	0.30
Neither	0.77	0.69	0.78	0.81	0.85	0.76	0.82	0.91	0.51
Support	0.08	0.09	0.09	0.03	0.05	0.08	0.07	0.05	0.20

Overall, **~20%** of the videos support OUD myths.

M2 (e.g., *OUD is a personal failure rather than a treatable disease*) had the highest levels of support + opposition among myths, indicating its contentiousness. **Need Targeted Intervention.**

Prevalence of Myths in YouTube Search Results

Findings

Distribution of labels

Bias Score	Topic	Support	Neither	Oppose
0.15	Kratom	0.36	0.42	0.22
-0.02	Heroin	0.22	0.53	0.25
-0.03	Codeine	0.04	0.90	0.07
-0.04	Methadone	0.36	0.24	0.40
-0.14	Percocet	0.03	0.79	0.18
-0.20	Fentanyl	0.13	0.54	0.33
-0.21	Narcan	0.02	0.74	0.24
-0.31	Suboxone	0.25	0.19	0.56

Bias Score: *Positive indicates videos lean towards supporting misinfo, negative indicates video opposes misinfo*

Kratom has the highest prevalence of myths (36% of search results) across topics.

Kratom is widely—but falsely—touted as an OUD treatment, potentially misleading users toward unsafe alternatives compared to evidence-based treatments like MAT.

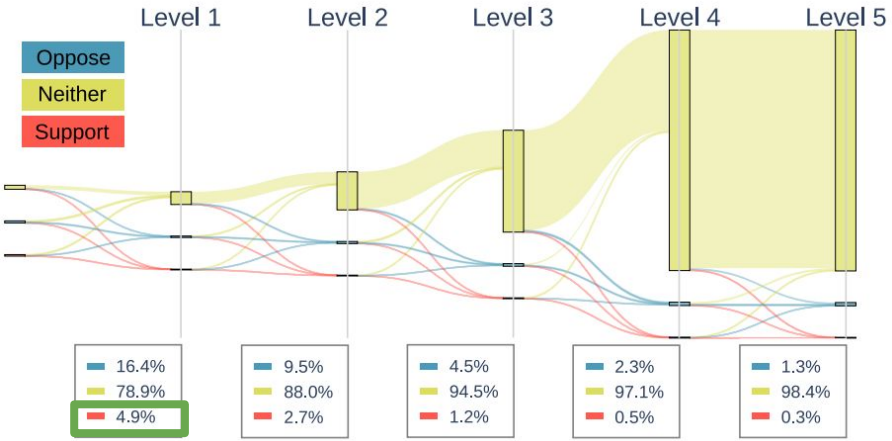
Heroin and Methadone show high levels of myth supporting content (22% and 36%).

These findings can help platforms prioritize moderation on high-risk topics and inform health officials on where myths are most prevalent.

Prevalence of Myths in YouTube Recommendations

Findings

Recommendation transition across levels.



Level 1 recommendations had the highest proportion of myth-supporting videos (4.9%).

This is concerning as early recommendations can shape user engagement and viewing trajectories.

At level 1, 12.7% recommendations to myth-supporting videos lead to other supporting videos, rising to 22% by level 5.

Continued engagement with such videos may increase exposure to more supporting videos.

These findings can inform YouTube's moderation efforts in recommendation pathways that may reinforce users to myths.

Takeaways

MythTriage achieves strong performance on detecting OUD myths against expert labels, while greatly reducing annotation time and cost.

1 **MythTriage** can be integrated into platform moderation workflows and empower public health practitioners monitor misinformation trends.

Takeaways

Using **MythTriage**, we reveal concerning levels of myth-supporting content across YouTube and offer actionable insights.

2 **Actionable insights can empower platform moderators and public health officials improve content moderation and launch targeted health interventions.**

Thank you!

MythTriage: Scalable Detection of Opioid Use Disorder Myths on a Video-Sharing Platform

Hayoung Jung, Shravika Mittal, Ananya Aatreya, Navreet Kaur, Munmun De Choudhury, Tanushree Mitra

PRINCETON UNIVERSITY

UNIVERSITY of WASHINGTON

Georgia Institute of Technology

EMNLP 2025
Suzhou, China | 中国苏州