Identifying Self-Disclosures of Use, Misuse and Addiction in Community-based Social Media Posts

Chenghao Yang[1,*,†], **Tuhin Chakrabarty**[2,*], **Karli R Hochstatter**[3,†], **Melissa N Slavin**[4,†], **Nabila El-Bassel**[5], **Smaranda Muresan**[2,6]

[1]University of Chicago [2]Department of Computer Science, Columbia University
[3]Friends Research Institute [4]Fairleigh Dickinson University
[5]School of Social Work, Columbia University [6]Data Science Institute, Columbia University
chenghao@uchicago.edu, {tc2896, sm761}@columbia.edu

Abstract

In the last decade, the United States has lost more than 500,000 people from an overdose involving prescription and illicit opioids[1] making it a national public health emergency (USDHHS, 2017). Medical practitioners require robust and timely tools that can effectively identify at-risk patients. Community-based social media platforms such as Reddit allow self-disclosure for users to discuss otherwise sensitive drug-related behaviors. We present a moderate size corpus of 2500 opioid-related posts from various subreddits labeled with six different phases of opioid use: *Medical Use, Misuse, Addiction, Recovery, Relapse, Not Using*. For every post, we annotate span-level extractive explanations and crucially study their role both in annotation quality and model development.[2] We evaluate several state-of-the-art models in a supervised, few-shot, or zero-shot setting. Experimental results and error analysis show that identifying the phases of opioid use disorder is highly contextual and challenging. However, we find that using explanations during modeling leads to a significant boost in classification accuracy demonstrating their beneficial role in a high-stakes domain such as studying the opioid use disorder continuum.

1 Introduction

Extensive ongoing overuse of opioid medications, both from medical prescriptions and from illegal sources has led to a major public health crisis (Degenhardt et al., 2019; Krausz et al., 2021). There have been a total of 103,664 drug overdose deaths in the US in the 12-month period ending April 2022.[3] For individuals with opioid use disorder (OUD), targeted interventions need to be developed to better capture individuals' transitions at critical junctures (e.g., use to misuse; misuse to addiction; recovery to relapse) (Park et al., 2020).

I'm 18m and I've been taking norcos since I was 16 but just on and off. **Starting this year I've been taking it every day basically and now I'm tired of it. I still get high so ig my addiction isn't that bad** as others but I don't want to get to that point. I'm tired of chasing the high. I've spent at least 3k on norcos this year and **I can't control myself**. I try to go a day sober but my mind is telling me I need and then withdrawals starts [...]

Table 1: A self-disclosure from a user on Reddit going through the cycle of Opioid Addiction.

Due to their anonymous and real-time participation, community-based social media platforms such as Reddit, have been used by researchers to understand issues around mental health self-disclosure (Choudhury and De, 2014), suicide among youth (Sumner et al., 2019), marijuana regulations (Park and Conway, 2017), drug community analysis (Bouzoubaa et al., 2023) and Covid-19 impact on people who use opioids (El-Bassel et al., 2022). We choose Reddit for our research, specifically the popular opioid-related subreddits *r/Opiates, r/OpiatesRecovery* as well as *r/drugs* to collect our data (§ 2.1). Our research focuses on predicting the presence of self-disclosures related to OUD phases in users' Reddit posts (refer to Table 1 for an example). This task is critical in providing healthcare professionals and social workers with automated tools for detecting OUD indications in social media posts. Accurate identification of such self-disclosures can enable more effective, targeted interventions for individuals suffering from OUD, as supported by

*Equal Contribution.
†Work done when authors were at Columbia University.
[1]https://www.cdc.gov/drugoverdose/epidemic/index.html
[2]The codebase and dataset specification are available at https://github.com/yangalan123/OpioidID.
[3]https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2507–2521

prior research (Acion et al., 2017; Park et al., 2020; Hasan et al., 2021). Our goal is to establish an annotation framework based on addiction and substance use research, categorizing behaviors like Medical Use, Misuse, Addiction, Recovery, and Relapse. We also seek to demonstrate the effectiveness of recent NLP advancements, especially through the application of explanations and text-to-text models, in accurately identifying self-disclosures within the OUD continuum. We offer three primary contributions:

- *An annotation scheme amenable for both expert and novice annotations of self-disclosures.* The proposed scheme has three characteristics: 1) is grounded in research on addiction and substance use 2) aims to focus on self-disclosure of OUD phases by including a category Not Using that applies to posts that are not discussing the author's OUD experience; and 3) aims to provide reliable annotations by both experts and novices (§ 2).

- *High-quality dataset annotated with class labels and text explanations using expert and novice annotators.* Human annotations are essential, both to ensure that the NLP models can accurately learn to identify the various OUD phrases, and as an upper bound on the expected model performance. Towards this, we employ both substance use research experts and skilled crowd-workers to annotate our data based on our scheme (§ 2.1). To ground annotators' decisions towards a particular label, we also asked them to highlight the minimum span from the input that acts as an explanation for their chosen category/label.

- *Thorough experimental setup of zero-shot, few-shot, and supervised models with insights into the role of explanations for model performance, the impact of label uncertainty, and intriguing properties of users' self-disclosure.* Our experiments demonstrate that: 1) the model performance improves significantly when trained/prompted with explanations. A further ablation study on human-annotated explanations versus machine-generated explanations confirms that the quality of explanations is key to such improvement; 2) smaller models fine-tuned on our novice-annotated data with explanations works best, surpassing zero-shot and few-shot large models, including GPT-4, by a large margin (§ 4); 3) an ablation study taking into account label uncertainty sheds light on model errors for cases where humans agree or disagree on the label; 4) our error analysis shows preliminary insights in understanding users' self-disclosure (§ 6).

2 Data

2.1 Data Collection and Annotation

Data Source One of the greatest challenges in building models that are capable of identifying the appropriate category for opioid usage is the lack of publicly available large-scale datasets. Social media platforms such as Reddit often provide social support for people who use opioids, while allowing for anonymity when discussing stigmatized behaviors (Pandrekar et al., 2018; Bunting et al., 2021). We collect data from the popular opioid subreddits, *r/Opiates* and *r/OpiatesRecovery* as well the *r/drugs* subreddit. Since *r/drugs* can contain posts related to other drugs, we only select posts that are labeled with a flair(tag) "opioids" by the moderator.

Anonymization and Data Preprocessing To remove any personal identifying information (PII) that users might divulge in their posts (e.g., emails) and broken characters, we use *cleantext*[4] to preprocess raw social media posts. In addition, we manually investigated all samples prepared for annotation to make sure PII will not be exposed to annotators, and thus will not be released in the final dataset. After that, we check whether each post is of reasonable length (title + text), and filter the preprocessed posts having a length of less than 10 words (or more than 200 words for easier annotation). We sample 600 posts for expert annotation and $2,250$ posts for novice annotation.[5]

Annotation Guidelines To ensure the annotation quality, we worked closely with substance use research experts to develop comprehensive and precise annotation guidelines for different phases of opioid use. OUD has been recognized as a chronic, relapsing disorder in which individuals may begin at one stage, remain in that stage, grad-

[4] `www.github.com/prasanthg3/cleantext`

[5] Domain experts are postdoctoral and advanced doctoral students working in substance abuse research. We use MTurk for novice annotation.

Medical Use	**Oxycodone for wisdom teeth removal** I just got 4 wisdom teeth plus another tooth in my palette removed and got prescribed 1 or 2 5mg tablets of oxy (Endone) each time. He recommended to avoid it if I could since I'm 43kg and have no tolerance. [...]
Misuse	**Oxy nod but no euphoria?** Hi everyone, I tried oxy for the first time a few weeks back snorting a prolonged 20mg tablet and felt pretty good. Wednesday I dropped 9 of the 5mg capsules over a couple hours and was nodding strongly [...]
Addiction	**Well y'all were right.** The sickness came. And is the worst i've ever experienced. Took subs, went into pwd accidentally and jump started the methadone sickness. I am to the point that I just have to get off this godforsaken mountain and go back to my ex and get back in the clinic bc at this rate i'm afraid i'm gonna end up killing myself.[...].
Recovery	**It's my birthday!** One year off opiates It's been 365 days since I decided to take back control of my body. I was highly dependent and addicted to prescribed opiates [...]
Relapse	**So high. 18 hours later. Still so high.** So I'm pissed at myself. I was clean from heroin for 11 months and last night I did some. And for no reason too [...]
Not Using	**Partners of an Opiate addict in recovery** How do you guys do this? I feel like I am having an incredibly hard time "moving on". I have nightmares of my partner oding, dying, and pretty much anything else that involves drug use. I over analyze everything [...]

Table 2: Example for each Opioid Usage category. The underlined bold text represents the title of each post. Highlighted text represents salient spans annotated by humans as explanations for the label.

ually or rapidly advance to another stage, enter recovery, return to use, or even skip stages (Volkow, 2007). For this study, we adopted frequently used classifications to assign each post a stage in the continuum: Medical Use, Misuse, Addiction, Recovery, and Relapse (NIDA, 2007; Smith et al., 2013; Hanson et al., 2013a,b; Chan et al., 2015; Anderson et al., 2017; Phan et al., 2017; Hu et al., 2019). Our definitions for Medical Use, Misuse, and Addiction come from the systematic review (Smith et al., 2013), and our definitions for Recovery and Relapse come from National Institute on Drug Abuse guideline (NIDA, 2007). We also built a list of keywords, representative samples and FAQs to clarify the project background, ethical considerations, and how to handle uncertain cases. The guidelines aim to understand the opioid use experiences of *the author of the post* (self-disclosures). Thus, we introduced also a category of 'Not Using' that includes discussion about someone else who uses opioids or general questions about opioids, without evidence of use. Appendix A shows the definitions for each category and some examples of expert-authored FAQs for clarification. Table 3 shows the distribution of OUD categories in the annotation data.

Category	Novice	Expert
Misuse	22.10	20.0
Addiction	29.15	12.53
Recovery	18.89	25.49
Relapse	4.65	3.96
Medical Use	7.05	3.52
Not Using	18.17	34.51

Table 3: Distribution (%) of OUD categories in novice- and expert-annotated data.

Expert Annotation To build the expert evaluation dataset, we invited 4 substance use research experts to annotate 600 posts and paid them at a rate of $20/hour. To accommodate the experts' available timeslots, we split the posts into two equal batches and asked the experts to annotate the text and title of the post with both the label and the explanation. All four experts annotated the first batch. For label annotation, the inter-annotation agreement (IAA) was 0.46 Fleiss' kappa (Fleiss, 1971), indicating "moderate agreement". Only two experts were available to annotate the second batch, and the IAA was 0.62 Cohen's kappa (Cohen, 1960), indicating "substantial agreement". We filtered the posts that did not obtain majority agreement, obtaining an expert-annotated dataset of 455 posts. We further split the dataset into 13 samples as in-context samples of few-shot prompting (§ 3) and 442 samples for testing.

Novice Annotation Expert annotation, while being more accurate and trustworthy, is not feasible for scaling the process beyond a few hundred posts. Hence, we aimed to leverage novice annotators using Amazon MTurk. However, to obtain reliably annotated data, we need to ensure that novices are qualified and trained. Thus, we first

Is it safe to mix 20mg oxy with a few standard drinks? Seen online it's dangerous but I don't trust a lot of those harm reduction websites, most of them are whack. EDIT: **I didn't do it don't worry, thanks for the info**

Table 4: An annotation example demonstrating the role of explanation annotations for understanding annotator disagreement: the red is associated with "Misuse" and the blue with "Not Using".

Given the following title, text, and explanation from the text, please identify the appropriate opioid usage category among the following types: 'Medical Use', 'Misuse' ,'Recovery', 'Relapse', 'Addiction', 'Not Using'. Title: {{title}} Text: {{text}} Explanation: {{explanation}}

Table 5: Zero-shot instructional prompts for T0pp for *w/ Explanation* setting

conducted a qualification test where we recruited a total of 85 crowd-workers from the USA with a 98% success rate and asked them to annotate 250 randomly selected instances from the expert-labeled set. We qualified only 10 crowd-workers who obtained >60% accuracy in the qualification phase. In addition, for cases of disagreement with the experts, we further trained the novice annotators by providing them with follow-up explanations. We paid them $15/hour, which is in accordance with the minimum wage in the USA. Every post is labeled by three qualified novice annotators. We labeled 2,250 posts and kept 2,086 for which we could obtain a majority vote label. We split this set into 1,936 for training and 150 for testing. IAA was 0.47 based on Fleiss' kappa (Fleiss, 1971) ("moderate agreement").

Explanation Annotation Along with providing a label we also asked annotators to identify the minimum salient span from the text that justifies their decision towards labeling a post to a certain category. For cases where we have a majority vote and use the corresponding label as gold, we have to decide what explanation to include. We computed the max overlapping substring between the annotators' explanations. When the max overlapping substring is very short (typically $<$ 10 characters), we chose the longest explanation whose annotated label matches the majority vote label. For 63% of cases, there is significant overlap among annotators' selected explanation spans, while for 37% of cases the longest explanation is selected. Table 2 shows post examples in each category/label along with their annotated span-level explanations.

2.2 Disagreements in Annotation

Expert-Novice Disagreement During the qualification test, we observe a consistent labeling disagreement between our qualified novice annotators and the expert annotators (The confusion matrix is shown in Appendix B). The main disagreement between experts and novices are between **"Addiction" - "Recovery"** (22.35%), **"Not Using" - "Misuse"** (19.35%), **"Addiction" - "Misuse"** (12.90%), **"Medical Use" - "Misuse"** (10.75%) and **"Recovery" - "Relapse"** (8.60%).

Novice-Novice Disagreement. Even though we reach a majority vote for 2086 posts, an individual worker can still disagree on the collective label. Looking at these disagreements can help us better understand the difficulty of this task and the uncertainty in the annotated dataset. In total, 1165 out of 2086 (56%) posts in our final novice-annotated datasets fall in this category. The top-5 disagreements happen between **"Addiction"-"Misuse"** (34.84%), **"Recovery"-"Addiction"** (15.71%), **"Not Using"-"Addiction"** (12.27%), **"Not Using"-"Misuse"** (9.78%) and **"Not Using"-"Recovery"** (7.38%). This inherent uncertainty may inject wrong inductive bias into models, which we discuss in § 6.

A closer look at some examples of disagreement in annotations shows that selected explanations could shed some light. For example, Table 4, shows an example of disagreement between Misuse and Not Using, where the annotators selected two different explanations for the labels.

3 Modeling Strategies

As OUD status prediction is a high-stakes task with limited labeled data, we consider three different settings, gradually increasing the number of labeled data required to mimic real-world application scenarios: zero-shot, few-shot, and supervised learning. To understand the effectiveness of annotated span-level explanations, we conduct two experiments for each setting: i) *w/o Explanation*: where the explanation is not included in the input, and ii) *w/ Explanation*: otherwise.

Given the following title and text, please identify the appropriate opioid usage category among the following types: 'Medical Use', 'Misuse', 'Recovery', 'Relapse', 'Addiction', 'Not Using'. Please provide an explanation for your answer by extracting the relevant span from the text that justifies your choice. {13 in-context samples with the format below} Title: {{title}} Text: {{text}} Label: {{label}} Explanation: {{explanation}}

Table 6: Few-shot instructional prompt for GPT-3 for *w/ Explanation* setting.

Zero-Shot We first consider the extreme application scenario when zero training data is given. In order to measure zero-shot performance on our dataset, we prompt the widely-used instruction-tuned T0pp (Sanh et al., 2022) model for our task. The prompt with instructions are demonstrated in Table 5.[6] We use greedy search to generate the labels, then use exact match to compute accuracy after lowercasing both the output and label.

Few-Shot Now we relax the dataset size limitation to allow the few-shot setting. We use the GPT3-Davinci-002 model (Brown et al., 2020) and GPT-4 (OpenAI, 2023) for the few-shot learning method. Our prompts begin with the task instruction followed by 13 expert-annotated samples for in-context learning.[7] For in-context learning *w/ Explanation*, we place the explanation on a line after the answer, preceded by "Explanation:"(Lampinen et al., 2022). Table 6 shows an example prompt.[8] In this way, the evaluation can be performed regardless of whether explanations are provided in the prompt or not.[9]

Fully Supervised All of our training data comes from the novice-annotated set, while our test sets consist of expert or novice-annotated data. Our training data consists of 1936 examples, while our test sets consist of 442 expert-annotated examples and 150 novice-annotated examples. We consider two modeling variants: Masked Language Models (MLM) (as it is often used in traditional fine-tuning) and Generative Language Models (GLM) (as it is often used in instruction-tuning).

For MLM, we fine-tune DeBERTa-v3-large (He et al., 2021) on our training data.For input formatting, we use "[title] TITLE [text] TEXT [Rationale] RATIONALE" as the input for *w/ Explanation* settings and use "[title] TITLE [text] TEXT" to train models under *w/o Explanation* settings. The token in square brackets (e.g., "[title]") are special tokens and the tokens in all-caps (e.g., "TEXT") are actual text fields for each post. For GLM, we fine-tune T5-3B and T5-11B models (Raffel et al., 2020). We use the same instruction as input to the encoder for a given title, text and optionally explanation as the ones we used for zero-shot setting (see Table 5). The decoder generates the textual label autoregressively. More implementation details for fine-tuning can be found in Appendix E.

[6] We only presented *w/ Explanation* case to save space.

[7] See Appendix C for these in-context samples and the detailed explanations for selecting these examples.

[8] We only show *w/ Explanation* case to save space.

[9] Necessary post-processing during evaluation for GPT-3/4 output normalization is detailed in Appendix D.

4 Experiments

Table 7 summarizes our experimental results under three different learning settings (zero-shot, few-shot, and fully supervised) across two different modes i) *w/o Explanation* when only the Title and Text are a part of our input during training and testing, and ii) *w/ Explanation* when along with Title and Text, gold human annotated explanations are a part of our input during training and testing. We show results on both the expert-annotated test set and the novice-annotated test set. As the OUD category distribution in our dataset is unbalanced, we report both accuracy and macro F1 scores in Table 7. We find for the model-wise comparison, there is little difference in using accuracy or F1.

We highlight several takeaways. First, adding explanations helps the models both on expert and novice-annotated data (except for T0pp and GPT3 on Expert data), particularly in few-shot and fully supervised settings. In § 5, we will show additional experiments to study the role of explanations and their quality for model predictions. Second, supervised learning with small models outperform few-shot methods with larger models including GPT-4 by a large margin, on both expert and novice evaluation datasets, even if the training data is novice-annotated. T5-11B is the best overall model. While our training data is not annotated by experts, the quality of the data is still high. The accuracy on the expert evaluation set for a random baseline would be 17%, while a majority baseline would be 35%, which is significantly lower than 71.4% or 76.6% for the T5-11B model performance *w/o Explanation* and *w/ Explanation*,

Mode	Test Set	Zero-Shot	Few-Shot		Supervised		
		T0pp	GPT3	GPT4	DeBERTa	T5-3B	T5-11B
w/o Explanation	Expert	48.9 / 46.9	62.2 / 57.1	55.4 / 50.2	67.6 / 65.7	63.5 / 61.2	**71.4 / 70.4**
	Novice	62.0 / 60.8	66.4 / 65.4	63.3 / 60.0	74.0 / 74.4	72.7 / 71.4	**80.9 / 81.5**
w/ Explanation	Expert	48.9 / 47.4	61.1 / 54.5	63.2 / 59.1	73.8 / 72.6	64.4 / 65.5	**76.6 / 77.0**
	Novice	62.7 / 58.5	66.9 / 65.9	67.3 / 64.8	81.3 / 81.9	78.7 / 77.3	**84.0 / 84.0**

Table 7: Performance of different models on expert and novice-annotated test data in a zero-shot/few-shot/supervised setting. *w/o Explanation* and *w/ Explanation* models refers to the setting where *Explanations* are excluded or included as part of the input. Results are presented in "Accuracy/F1" format.

Explanation	Test Set	T5-11B	DeBERTa
Gold	Expert	**76.6**	73.8
	Novice	**84.0**	81.3
Silver	Expert	70.3	70.3
	Novice	78.0	78.0
Random	Expert	69.6 ($\pm$ 1.3)	65.8 ($\pm$ 1.1)
	Novice	68.3 ($\pm$ 1.7)	67.9 ($\pm$ 1.4)

Table 8: Accuracy of T5-11B and DeBERTa *w/ Explanation* model on expert and novice annotated test sets by varying the quality of explanations. We can observe the importance of including gold explanations.

respectively. Moreover, we notice that the performance gap between expert-annotated test data and novice-annotated test data is reduced using supervised models. A closer look at GPT-4 errors shows that GPT-4 is particularly struggling with the “Not Using” category, which covers a diverse range of topics that can look very different from posts in other categories, and more analysis on this category will be further studied in § 6. Third, model capabilities improve with scale under the same family in a supervised setting. The T5-11B model, on average, is about 8.4 points better than the T5-3B model in accuracy and 9.3 points better in F1. However, when models belong to different families (i.e., Generative vs. MLM), the scaling law might not hold as the DeBERTa-v3-large model (1.5B) outperforms T5-3B across both settings (*w/o and w/ Explanation*).

5 The Role of Explanations

To test the quality and helpfulness of the annotated explanations on model prediction, we conduct three different experiments using our two best-performing models trained *w/ Explanation* (T5-11B and DeBERTa). All these experiments are conducted at inference time on top of a model fine-tuned on <title, text, E^{gold}>. For convenience, from here on, we will refer to this model as M1.

Gold Explanations at Inference. In the first experiment, we use the gold explanations from our test sets (expert and novice). In particular, during inference, we prompt the two best-performing models (T5-11B and DeBERTa) with an input that consists of <title, text, E^{gold}>. Table 8 shows that models that use gold explanations at inference time are the best. We analyze whether the explanation contains words that refer to the label (e.g., addiction or addicted), a problem referred to as *leakage* (Sun et al., 2022). We notice that there is 5.6% leakage on expert-annotated test data and 8% leakage on novice-annotated test data, which means that most of our annotated explanations do not give away the label easily.

Silver Explanations at Inference. In a real-world setting, it is not possible to expect gold explanations at inference time. Thus, in this setting, we investigate whether model-generated explanations can still be helpful for final label prediction. Prior works in explainability have trained two types of models: 1) *Pipeline* model, which maps an input to an explanation (I $\rightarrow$ E), and then an explanation to an output (E $\rightarrow$ O); and 2) *Joint Self Explaining* models that map an input to an output and explanation (I $\rightarrow$ OE). The latter has been shown to be more reliable (Wiegreffe et al., 2021). Thus, we first train a T5-11B model (M2) that can jointly generate <label, explanation> given any <title, text>. At inference time, we first generate a silver explanation E_i^{silver} by prompting M2 with a given <title$_i$, text$_i$> from the test set. We then prompt M1 with <title$_i$, text$_i$, E_i^{silver}> to generate label$_i$. While these explanations are not as high quality as gold explanations, they still outperform random explanations. It should be noted that the goal of this paper is not to build models that facilitate extracting accurate explanations. However, such models might improve the silver quality explanations and thereby improve overall classification results. We leave this for future work.

Random Explanations at Inference. As a baseline, we use a randomly selected sentence from the post as the explanation. We repeat the random selection for five random seeds and report the mean and standard deviation of these five runs (Table 8). Both silver and gold explanations outperform the random explanation baseline, indicating the need for informative, high-quality explanations.

6 Error Analysis

To understand the challenges and limitations of our best models, we perform an error analysis.

Model Errors We compute the confusion matrices for DeBERTa and T5-11B on the expert evaluation dataset, as it is arguably more reliable and contains more examples. We generally find that both *w/ Explanation* and *w/o Explanation* models struggle with confusion between **1) Not Using - Misuse; 2) Recovery - Addiction; and 3) Not Using - Addiction** (Table 9, some of which we also noticed in the disagreement among annotators. We notice that these problems surface in a very asymmetrical pattern – one direction (e.g., "Not Using" $\rightarrow$ "Misuse") matters more in this confusion. Recall that our focus is on self-disclosures, so if a post discusses misuse (either a question or someone else misusing behavior), the expert label is Not Using, which might be difficult for models to capture in some cases. Adding explanations mostly helps the model by reducing the confusion on the 'Not Using' and 'Recovery' labels, the two dominant as well as the top-2 most difficult categories in expert-annotated data.

Error Annotations To better understand why the model makes mistakes we did a thorough fine-grained error case annotation for the T5-11B *w/ Explanation* model. When analyzing why our models misclassified the **Recovery** class, we notice that "Recovery" can be a long process, and it is very common for users to express their eagerness to get opiates (47.4%), and/or to talk about their history of addiction (21.1%). There are also some hard cases, such as a post showing the patient undergoing repeated recovery-relapse-recovery cycles (the model predicts "relapse" in this case). When analyzing the cases where our model misclassified the **Not Using** label, several cases emerge: 1) *asking a question about use/misuse/addiction (57.69%)* (e.g., "how much [Drug 1] should I take to get high (safely)? Can I use it with [Drug 2]?", or asking questions about whether using drugs for certain syndromes is legal in some states and how much they should use); 2) *irrelevant topics (23.08%)* ("Merry Christmas!"); 3) *Others' overdose (7.69%)* (discussing the addiction of their friends or family members; since we are interested in self-disclosures, this is labeled as Not Using, but models fail to recognize such subtle differences); and 4) *other drugs/substances, not opiates (3.85%)* (as we focus on OUD, these posts are labeled "Not Using").

Influence of Dataset Annotation Uncertainty As we have already seen in the previous sections, annotators found it difficult to annotate several edge cases, which in turn brings uncertainty in the final annotation. To investigate how such uncertainty influences model performance, we do a further ablation study to test the model performance on data with unanimous agreement (all annotators give the same label) (47% on the novice test set, 44% on the expert test set)[10] and data where some disagreement exists, although majority voting can be reached (we call it arguable data). For the latter, we consider as gold label either the majority vote or any label chosen by at least one annotator. The results are shown in Table 10.

We notice that: 1) models perform better on data with unanimous agreement than on arguable data (15%-32%); 2) given the difficulty of the annotation task, if we consider all annotators' labels as gold (Arguable, all annotations), we can see the model can improve (14%-25%); 3) by comparing the performance on the first batch of expert-annotated data and the novice-annotated data, our models achieve very similar performance on instances with unanimous agreement and also when considering all annotators' labels as gold. In arguable cases with majority voting, however, models trained on novice-annotated data cannot perform as well on expert test sets where experts cannot reach a unanimous agreement or where we do not consider all labels. This confirms the fact that the disagreement among annotators will influence the model performance and roughly quantify the performance bottleneck resulted from using majority voting as the gold label.

[10] Since only the first batch of expert data contains more than two annotators, for this study we only report ablation for this expert-annotated dataset.

Setting	Not Using - Misuse		Recovery - Addiction		Not Using - Addiction	
	→	←	→	←	→	←
T5-11B-*w/o Explanation*	20%	4.5%	20%	0%	14%	0%
DeBERTa-*w/o Explanation*	18%	0%	18%	3.6%	16%	5.5%
T5-11B-*w/ Explanation*	16%	8%	14%	0%	13%	0%
DeBERTa-*w/ Explanation*	15%	2.3%	15%	3.6%	15%	0%

Table 9: Model error analysis over expert annotation data. → means the expert-annotated label is on the left side and the predicted label is on the right side, and ← vice versa. Percentages in the table represent the error rate in each expert labeled category. The results demonstrate that the main confusion for the models exists in "Not Using - Misuse", "Recovery - Addiction", and "Not Using - Addiction". These problems surface in an asymmetrical pattern – one mis-classification direction matters more in the confusion.

Dataset	Agreement	T5-11B (w/o expl)	T5-11B (w/ expl)	DeBERTa (w/o expl)	DeBERTa (w/ expl)
Novice	Unanimous Consent	95.5 / 94.0	97.1 / 95.7	87.1 / 85.9	90.0 / 89.6
	Arguable (Majority Vote)	68.0 / 66.7	72.5 / 71.3	62.5 / 60.8	73.8 / 75.2
	Arguable (All Annotations)	84.0 / 81.4	92.5 / 91.8	86.3 / 84.3	91.3 / 92.1
Expert (FirstBatch)	Unanimous Consent	92.7 / 85.0	96.4 / 97.1	83.6 / 72.7	89.1 / 83.2
	Arguable (Majority Vote)	60.3 / 59.9	64.0 / 66.0	56.9 / 58.2	65.0 / 66.2
	Arguable (All Annotations)	84.6 / 83.3	86.8 / 85.8	82.5 / 82.7	86.9 / 87.3

Table 10: Ablation study on dataset annotation disagreement. Results are presented in "Accuracy/F1" format. "Unanimous Consent": all annotators agree on the same label; "Arguable (Majority Vote)": annotators have some disagreements, and majority voting is used as the correct label; "Arguable (All annotations)": disagreements exist, and any annotator label is considered correct. We observe that models perform better on data with unanimous agreement than on arguable data.

7 Related Research

Machine Learning for Substance Use Machine learning methods' application to substance use research is growing (Bharat et al., 2021). Several studies have attempted to predict substance use treatment completion among individuals with substance use disorders (Gottlieb et al., 2022; Acion et al., 2017; Hasan et al., 2021). This study takes advantage of anonymous data to identify treatment needs among individuals who may not currently be in formal substance use treatment. Researchers have also used natural language processing to identify substance misuse in electronic health records (Afshar et al., 2019; Riddick and Choo, 2022) and to classify substances involved in drug overdose deaths (Goodman-Meza et al., 2022). MacLean et al. (2015) collect user-level data on a social platform, Forum 77, to build a CRF model predicting three phases of drug use: using, withdrawing, and recovering. Our work is different in several aspects: 1) we propose an annotation scheme grounded in research on addiction and substance use that defines behaviors such as Medical Use, Misuse, Addiction, Recovery, Relapse (and Not Using), that enable us to code self-disclosures of such behaviors using both expert and novice annotators; 2) we develop explanation-infused accurate models to identify self-disclosure at the post level. These two innovations will enable future research on using these models for a reliable global, user-level analysis across time.

Learning from Explanations There have been works focusing on learning from human-annotated explanations. Wiegreffe et al. (2021) investigates how free-form explanations and predicted labels are associated and use it as a property to evaluate the faithfulness of explanations. Different from that, our work focuses more on the utility of extractive span-level explanations as an additional source of supervision in a high-stakes domain and further shows how the quality of explanations impacts inference time results (Sun et al., 2022). Similar to our work, Carton et al. (2022) leverages extractive explanations and shows a consistent trend that using explanations can improve model performance in reading comprehension. Our work is most similar to Huang et al. (2021), who noticed that the quality of explanations could have a huge impact on model performance and explore the utility of extractive explanations, and to Sun et al. (2022), who perform similar studies using free-form explanations.

Understanding the OUD continuum Scientists have explained how opioids produce changes in brain structure and function that promote and sustain addiction and contribute to relapse (Koob and

Volkow, 2010; Abuse et al., 2016). Now recognized as a chronic but treatable disease of the brain, OUD is characterized by clinically significant impairments in health and social function and influenced by genetic, developmental, behavioral, social, and environmental factors (Volkow et al., 2016). The HEALing Communities Study implemented the Opioid-overdose Reduction Continuum of Care Approach (ORCCA) to reduce opioid-overdose deaths across the OUD continuum (Winhusen et al., 2020). Taking advantage of self-disclosures on community-based social media, as this study aims to do, could lead to the development of interventions that better address risks associated with OUD.

8 Conclusions

We presented a novel task aimed to deepen our understanding of how people move across the OUD continuum: given a user's post in an opioid-related Reddit, predict whether it contains a self-disclosure of various phases of OUD. We provided an annotation scheme grounded in research on addiction and substance use, which enables us to code self-disclosures of such behaviors using both expert and novice annotators. Following the annotation scheme, we created a high-quality dataset annotated with class labels and text explanations. We presented several state-of-the-art explanation-infused models, showing they can achieve accurate results in identifying self-disclosures of use, misuse, addiction, recovery, and relapse. Accurate models will enable further research in this space by considering a global user-level analysis across time. Our error analysis showed that explanations could provide insights both into annotator disagreement and errors in model predictions. In addition, our findings shed light on how annotation uncertainty impacts model performance.

Acknowledgements

This research is supported by the National Institutes of Health (NIH), National Institute on Drug Abuse, grant #UM1DA049412. We want to thank our expert and novice annotators. We also thank the anonymous reviewers and chairs for providing insightful and constructive feedback to make this work more solid.

Limitations

This study's results are not without limitations. The anonymity of Reddit users does not allow us to characterize the demographics or geographic extent of the study population. Moreover, the current study looks at identifying self-disclosures at the message level without taking a global (user-level) and temporal view. In our future work, we plan to apply our models to study users' posts in opioid-related Reddits and observe their behavior over time. In addition, we will work on improving our models to both predict a label and provide a textual explanation for the prediction.

Ethical Considerations

For our data collection and annotation, we have obtained IRB approval. The source data comes from Reddit (r/opiates, r/OpiatesRecovery and r/-Drugs), and is thus publicly available and anonymous. In addition, we preprocess the data to additionally remove any potentially identifiable information (see Section 2.1). All data is kept secure and online userIDs are not associated to the posts. For the expert annotation we compensated the experts with $20 per hour, and the novice annotators with $15 per hour.

Our intention of developing datasets and models for predicting the stages of opioid use disorder is to help health professionals and/or social workers to both understand personal experiences of people across the opioid used disorder continuum and potentially to identify people that might be at risk of overdose. The inclusion of explanations both in the annotation and in the prediction of our models could help the health professional better assess the models predictions. We emphasize that our models should be used with a human in the loop — for example a medical professional, or a social worker, who can look at the predicted labels and the explanations to decide whether or not they seem sensible. We note that because most of our data were collected from Reddit, a website with a known overall demographic skew (towards young, white, American men), our conclusions about what explanations are associated with various OUD stages cannot necessarily be applied to broader groups of people. This might be particularly acute for vulnerable populations such as people with opioid use disorder (OUD). We hope that this research stimulates more work by the research community to consider and model ways in which different groups self-disclose their experiences with OUD.

References

Substance Abuse, Mental Health Services Administration US, Office of the Surgeon General (US, et al. 2016. The neurobiology of substance use, misuse, and addiction. In *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health [Internet]*. US Department of Health and Human Services.

Laura Acion, Diana Kelmansky, Mark van der Laan, Ethan Sahker, DeShauna Jones, and Stephan Arndt. 2017. Use of a machine learning framework to predict substance use disorder treatment success. *PloS one*, 12(4):e0175383.

Majid Afshar, Andrew Phillips, Niranjan Karnik, Jeanne Mueller, Daniel To, Richard Gonzalez, Ron Price, Richard Cooper, Cara Joyce, and Dmitriy Dligach. 2019. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *Journal of the American Medical Informatics Association*, 26(3):254–261.

Laurie S Anderson, Heidi G Bell, Michael Gilbert, Julie E Davidson, Christina Winter, Monica J Barratt, Beta Win, Jeffery L Painter, Christopher Menone, Jonathan Sayegh, et al. 2017. Using social listening data to monitor misuse and nonmedical use of bupropion: a content analysis. *JMIR public health and surveillance*, 3(1):e6174.

Chrianna Bharat, Matthew Hickman, Sebastiano Barbieri, and Louisa Degenhardt. 2021. Big data and predictive modelling for the opioid crisis: existing research and future potential. *The Lancet Digital Health*, 3(6):e397–e407.

Layla Bouzoubaa, Jordyn Young, and Rezvaneh Rezapour. 2023. Exploring the landscape of drug communities on reddit: A network study. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 558–565.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Amanda M. Bunting, David Frank, Joshua Arshonsky, Marie A. Bragg, Samuel R. Friedman, and Noa Krawczyk. 2021. Socially-supportive norms and mutual aid of people who use opioids: An analysis of reddit during the initial covid-19 pandemic. *Drug and alcohol dependence*, page 108672.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to learn, and how: Toward effective learning from rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088, Dublin, Ireland. Association for Computational Linguistics.

Brian Chan, Andrea Lopez, and Urmimala Sarkar. 2015. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. *PloS one*, 10(8):e0135072.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *ICWSM*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Louisa Degenhardt, Jason Grebely, Jack Stone, Matthew Hickman, Peter Vickerman, Brandon DL Marshall, Julie Bruneau, Frederick L Altice, Graeme Henderson, Afarin Rahimi-Movaghar, et al. 2019. Global patterns of opioid use and dependence: harms to populations, interventions, and future action. *The Lancet*, 394(10208):1560–1579.

Nabila El-Bassel, Karli R Hochstatter, Melissa N Slavin, Chenghao Yang, Yudong Zhang, and Smaranda Muresan. 2022. Harnessing the power of social media to understand the impact of covid-19 on people who use drugs during lockdown and social distancing. *Journal of addiction medicine*, 16(2):e123.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

David Goodman-Meza, Chelsea L Shover, Jesus A Medina, Amber B Tang, Steven Shoptaw, and Alex AT Bui. 2022. Development and validation of machine models using natural language processing to classify substances involved in overdose deaths. *JAMA network open*, 5(8):e2225593–e2225593.

Assaf Gottlieb, Andrea Yatsco, Christine Bakos-Block, James R Langabeer, and Tiffany Champagne-Langabeer. 2022. Machine learning for predicting risk of early dropout in a recovery program for opioid use disorder. In *Healthcare*, volume 10, page 223. MDPI.

Carl L Hanson, Scott H Burton, Christophe Giraud-Carrier, Josh H West, Michael D Barnes, and Bret Hansen. 2013a. Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students. *Journal of medical Internet research*, 15(4):e2503.

Carl Lee Hanson, Ben Cannon, Scott Burton, and Christophe Giraud-Carrier. 2013b. An exploration of social circles and prescription drug abuse through twitter. *Journal of medical Internet research*, 15(9):e2741.

Md Mahmudul Hasan, Gary J Young, Jiesheng Shi, Prathamesh Mohite, Leonard D Young, Scott G Weiner, et al. 2021. A machine learning based two-stage clinical decision support system for predicting patients' discontinuation from opioid use disorder treatment: Retrospective observational study. *BMC Medical Informatics and Decision Making*, 21(1):1–21.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Han Hu, NhatHai Phan, James Geller, Stephen Iezzi, Huy T Vo, Dejing Dou, and Soon Ae Chun. 2019. An ensemble deep learning model for drug abuse detection in sparse twitter-sphere. In *MedInfo*, pages 163–167.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring distantly-labeled rationales in neural network models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5571–5582, Online. Association for Computational Linguistics.

George F Koob and Nora D Volkow. 2010. Neurocircuitry of addiction. *Neuropsychopharmacology*, 35(1):217–238.

R Michael Krausz, Jean Nicolas Westenberg, and Kimia Ziafat. 2021. The opioid overdose crisis as a global health challenge. *Current Opinion in Psychiatry*, 34(4):405–412.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1511–1526.

NIDA. 2007. *Drugs, brains, and behavior: The science of addiction*. National Institute on Drug Abuse, National Institutes of Health, US.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Sheetal Pandrekar, Xin Chen, Gaurav Gopalkrishna, Avi Srivastava, Mary Saltz, Joel Saltz, and Fusheng Wang. 2018. Social media based analysis of opioid epidemic using reddit. In *AMIA Annual Symposium Proceedings*, volume 2018, page 867. American Medical Informatics Association.

Albert Park and Mike Conway. 2017. Tracking health related discussions on reddit for public health applications. In *AMIA annual symposium proceedings*, volume 2017, page 1362. American Medical Informatics Association.

Ju Nyeong Park, Saba Rouhani, Leo Beletsky, Louise Vincent, Brendan Saloner, and Susan G Sherman. 2020. Situating the continuum of overdose risk in the social determinants of health: a new conceptual framework. *The Milbank Quarterly*, 98(3):700–746.

Nhathai Phan, Soon Ae Chun, Manasi Bhole, and James Geller. 2017. Enabling real-time drug abuse detection in tweets. In *2017 IEEE 33rd international conference on data engineering (ICDE)*, pages 1510–1514. IEEE.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Tyne A Riddick and Esther K Choo. 2022. Natural language processing to identify substance misuse in the electronic health record. *The Lancet Digital Health*, 4(6):e401–e402.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Shannon M Smith, Richard C Dart, Nathaniel P Katz, Florence Paillard, Edgar H Adams, Sandra D Comer, Aldemar Degroot, Robert R Edwards, J David Haddox, Jerome H Jaffe, et al. 2013. Classification and definition of misuse, abuse, and related events in clinical trials: Acttion systematic review and recommendations. *Pain®*, 154(11):2287–2296.

Steven A. Sumner, Stacey Galik, Jennifer Mathieu, Megan Ward, Thomas R. Kiley, Bradford N. Bartholow, Alison Dingwall, and Peter Mork. 2019. Temporal and geographic patterns of social media posts about an emerging suicide game. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 65 1:94–100.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. *arXiv preprint arXiv:2206.11083*.

USDHHS. 2017. Hhs acting secretary declares public health emergency to address national opioid crisis. *Washington, DC: USDHHS*.

ND Volkow. 2007. How science has revolutionized the understanding of drug addiction. *Drugs, Brains and Behavior: The Science of Addiction*.

Nora D Volkow, George F Koob, and A Thomas McLellan. 2016. Neurobiologic advances from the brain disease model of addiction. *New England Journal of Medicine*, 374(4):363–371.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284.

Theresa Winhusen, Alexander Walley, Laura C Fanucchi, Tim Hunt, Mike Lyons, Michelle Lofwall, Jennifer L Brown, Patricia R Freeman, Edward Nunes, Donna Beers, et al. 2020. The opioid-overdose reduction continuum of care approach (orcca): evidence-based practices in the healing communities study. *Drug and alcohol dependence*, 217:108325.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Annotation Guideline

A brief annotation guideline created by experts is shown in Table 11, which explains the definition for each OUD category. This guideline also comes with example posts picked by experts that help annotator under the definitions and we show them in Table 12. The full guideline is too large to put in this paper so we will release it in our GitHub project.

Experts also help draft FAQs for clarification in the initial trial of annotations. Examples of FAQs are shown below:

QUESTION: What if the post described family, friend, or peer opioid use and there is no evidence that the person posting used opioids?

ANSWER: This post should be labeled 'not using' because there is no evidence that the individual posting the comment used opioids.

QUESTION: What if the post discusses using stimulants, marijuana, or other drugs that are not opioids?

ANSWER: This post should be labeled 'not using' because this study is specifically focused on understanding the development and advancement of opioid use disorder.

QUESTION: Is 'misuse' restricted to prescription opioids?

ANSWER: We have decided for the purpose of this study that misuse will NOT be restricted to prescription opioids. Therefore, if someone describes trying a synthetic or semi-synthetic opioid (e.g., heroin) or using it infrequently, but does not display signs of being addicted, this post should be labeled 'misuse.'

QUESTION: What if the post asks a question about opioid use, but does not provide evidence that the individual posting the comment used opioids?

ANSWER: This post should be labelled 'not using' because there is no evidence that the individual posting the comment used opioids. They may just be curious.

QUESTION: If someone reports using drugs that are NOT opioids during a period of time when they are attempting to quit (i.e., when they are in recovery), should this be considered 'relapse?'

ANSWER: Because this study is focused on opioid use disorder, we have defined relapse as use of opioids after an attempt to quit. Thus, if the individual used other drugs that are not opioids during recovery, we will not consider this relapse.

B Heatmap for Worker-Expert labels over the Qualification Test

The heatmap summarizes the difference in annotations between workers and experts over the qualification test is shown in Fig. 1.

Medical Use	Medical use is defined as the use of prescription opioids that were prescribed by a medical professional for the purpose of treating a medical condition
Misuse	Misuse is defined as the use of a substance that does not follow medical indications or prescribed dosing. Substances are commonly used for nontherapeutic purposes to obtain psychotropic (eg, euphoric, seditative, or anxiolytic) effects. Misuse is not restricted to prescription opioids.
Addiction	Addiction is defined as compulsive opioid use that occurs despite personal harm or negative consequences. Addiction may involve impaired control and craving, neurobiologic dysfunction, physical and psychological dependence, and withdrawal.
Recovery	Recovery is a process of change through which individuals improve their health and wellness, live a self-directed life, and strive to reach their full potential without using opioids.
Relapse	Relapse is defined as the return to opioid use after an attempt to quit.
Not Using	Posts should be labeled 'Not Using' which are about substances other than opioids (e.g., marijuana), another person who uses opioids (e.g., family or friend), general questions about opioids without evidence that the persons use opioids, and irrelevant information.

Table 11: Expert guidelines on how to assign each post one of the six stages of the OUD continuum

Medical Use	I got pretty decent surgery on my feet and was prescribed 400 mg of oxy after takeing that In about 10 days as needed due to pain (never takeing more then prescribed) but I have had minor withdrawal symptoms I took a 3 day break when do you think i can start taking it agian when my foot hurts and not withdrawal
Misuse	So I was given vicoprofen (7.5 hydrocodone to 200tylenol) for a severe toothache. I have been using it as prescribed but dumb ass me decided to take quite a large dose last night after missing a few normal doses. If I go back to using the normal doses now, after one large one, is it still going to be effective? Or should I wait and if so how long."
Addiction	I have been on opiates (oxycodone/contin) for like 5-6 years. Started off really small, got really big, now at like medium use- compared to before. I spent the last year or so very slowly tapering from my high of 330mg/day to now about 80mg/day. At this point is just maintenance to be able to function properly in my everyday life w out being sick or too tired.
Recovery	"7 days clean from heroin today after having been IV'ing it on my daily basis since August, 2020"
Relapse	"i made it 70 days clean. now i'm back to square one. i wish i could stop but i can't. now i'm shooting 2 grams a day, plus 2-4 grams of coke a day. everytime i relapse i get more and more addicted. anyone else experience this ? that when you relapse it gets more out of control. but godam i love it, i love the feeling, the lifestyle. "
Not Using	"How do you feel about Oxford houses/halfway houses/sober houses?" "Dreary , rainy day here , thought about using , now binge watching Reno 911 instead . It's so funny lol

Table 12: Expert guidelines on example posts for each category

Category	T5-11B w/ Explanation	T5-11B w/o Explanation	GPT-4 w/ Explanation	GPT-4 w/o Explanation
Addiction	88%/98%	84%/94%	28%/62%	20%/49%
Medical Use	84%/87%	76%/87%	88%/87%	84%/87%
Misuse	88%/70%	88%/75%	76%/89%	72%/82%
Not Using	68%/66%	68%/63%	36%/30%	32%/27%
Recovery	92%/83%	84%/67%	88%/81%	84%/70%
Relapse	84%/81%	88%/69%	88%/81%	88%/69%

Table 13: Class-wise performance decomposition for different models. Results are presented in a format of "Accuracy on Novice Test Set/Accuracy on Expert Test Set".

C In-Context Samples in Few-Shot Learning Settings

The 13 in-context samples we used for prompting in the few-shot learning setting are shown in Table 14. These in-context samples are selected as the representative samples under each category after discussions with experts providing the annotations. The distribution of classes is decided based on preliminary experiments on held-out data.

D Post-processing needed for processing GPT-3 outputs

In our experiments, we generally find that GPT-3 outputs cannot be taken as exact match as outputs and can contain some typos, we provide the following post-processing for it:

1. We ignore any content after a newline symbol (i.e., "$\backslash n$").
2. If GPT3 responses are like "1) ... 2) ...", we

Title	Text	Opioid Usage Label	Explanation
Advice welcome	I am a 23yr old female, been addicted to H for 3 y...	relapse	I am a 23yr old female, been addicted to H for 3 y...
Nearly threw two months down the drain today.	Well everyone I've been clean from Heroin for the ...	recovery	Well everyone I've been clean from Heroin for the ...
2weeks clean from all opiates.i just want to vent a bit.	So two weeks ago I quit my job, opiate use, and go...	recovery	2weeks clean from all opiates
Listening to Christmas music....	I could have been someone "Well so could anyone.....	not using	I could have been someone "Well so could anyone.....
Heroin use	Hi, non user here, just curious as to what heroin ...	not using	Hi, non user here, just curious as to what heroin ...
Anyone here either a lawyer or have solid solid knowledge or experience on drug	Not really comfortable discussing this publicly, b...	not using	Anyone here either a lawyer or have solid solid kn...
Supeudol oxycodone, sniffable?	My doctor changed my oxy prescription to supeudol ...	medical use	My doctor changed my oxy prescription to supeudol ...
Back in the cycle..	I started using more again (daily when I can) afte...	addiction	I started using more again (daily when I can) afte...
ROA 30mg roxis (blues)	I am currently on 7 blues. That i have done over t...	misuse	I am currently on 7 blues. That i have done over t...
Hey guys! Kinda worried!	Hey, I took abour 4 lines of heroin at 6pmAnother ...	misuse	Hey, I took abour 4 lines of heroin at 6pmAnother ...
Really wanting to try heroin :/	I just wanna say before I start, Ik how bad it is ...	misuse	I've used weed, Xanax, coke, I'm off of 2 Kpins ri...

Table 14: Thirteen in-context examples for each Opioid Usage category.

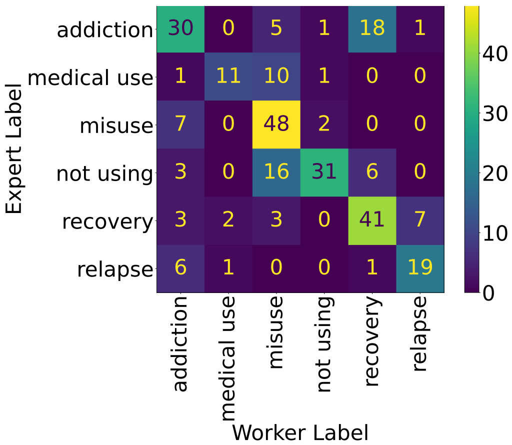

Figure 1: Heatmap for all worker labels and expert labels over the qualification test.

only take the term between "1)" and "2)".

3. For morphological changes like predicting "misuse" as "misusing", we manually recover these changes.

4. For typos like "misue", we would manually correct it to be "misuse".

We tried to apply the same processing for GPT-4 as well, but we did not find significant changes. This may indicate GPT-4 has better instruction-following capability while GPT-3 does not.

E Fully Supervised Fine-Tuning Details

In this section, we give details for fine-tuning language models under fully supervised setting.

For fine-tuning DeBERTa-v3-large (He et al., 2021), we adopt the widely-used huggingface transformers fine-tuning implementation (Wolf et al., 2020) with the learning rate of $2e-5$ and fine-tune the model for 10 epochs. For optimizer, we use AdamW (Loshchilov and Hutter, 2018).

For fine-tuning T5 (Raffel et al., 2020), we adopt the huggingface transformers implementation (Wolf et al., 2020) to fine-tune two versions of T5, the 3B model and the 11B model, respectively. We hold out 100 examples for validation from our training set to tune our models and find the best checkpoint. We use a batch size of 1024 for the 3B model and 512 for the 11B model. Further, we maintain a learning rate of 1e-4 and AdamW optimizer (Loshchilov and Hutter, 2018) for both 3B and 11B models. We fine-tune all models on 4 A100 GPUs and use Deepspeed (Rasley et al., 2020) integration for the 11B model. We fine-tune the 3B model for 20 epochs and the 11B model for eight epochs. During fine-tuning, we restrict the maximum sequence length of the source to 1024 (via truncation), while our target length is less than the default 128 tokens.

F Class-Wise Performance Decomposition

In § 4, we show the model average performance w/ and w/o explanations over all categories in Table 7. As there exist significant differences between OUD categories and their individual importance can vary depending on application purposes, we further show the class-wise performance decomposition in Table 13 for both expert and novice annotated test sets.

G Scientific Artifacts

In this paper, we use the following artifacts:

cleantext[11] (v1.1.4): is an open-source python package to clean raw text data. We use it to preprocess raw social media posts. This toolkit is released under an MIT license.

Transformers (Wolf et al., 2020)[12] (v4.35.0): provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio. We use it for model training and inference. This toolkit is released under an Apache-2.0 license.

OpenAI-python[13] (v1.0.0): provides convenient access to the OpenAI REST API from any Python

[11] www.github.com/prasanthg3/cleantext
[12] https://github.com/huggingface/transformers
[13] https://github.com/openai/openai-python

3.7+ application. The library includes type definitions for all request params and response fields, and offers both synchronous and asynchronous clients powered by httpx. We use it for prompting the GPT-series models. This toolkit is released under an Apache-2.0 license.

In addition, we plan to release our codebase and dataset under an MIT license in the formal version.