Reasoning, Contrastive, and In-Context Strategies for Opioid Use Stage Detection on Social Media

Vinu Ekanayake and Ramakanth Kavuluru
University of Kentucky, Lexington, KY USA
{vinu.ekanayake, ramakanth.kavuluru}@uky.edu

Abstract

The opioid epidemic has ravaged the US for the past two decades and is still a persistent threat. During the same time, the increasing use of social media has created a new avenue for people to share their journeys regarding opioid use. In this context, research in automatically determining opioid use stages (e.g., misuse, addiction, recovery) based on self disclosures in social media posts is gaining traction. In this paper, using a recent benchmark, we assess different supervised strategies for identifying self-disclosed opioid use stages from Reddit posts. We consider distilled reasoning traces from DeepSeek R1 (an open weights reasoning model), supervised contrastive learning (SCL), and few-shot in-context learning (ICL) with GPT-5 to conduct a variety of experiments with encoder and encoder-decoder models. We also conduct direct zero-shot (ZS) experiments with GPT 5 and GPT 5.2. Across different models and datasets, our strategies provide improvements in performance with some nuances that are too subtle to elaborate in the abstract. A surprising finding is that ZS results with GPT-5 are better than all supervised results, which ushers a new frontier for LLM-based classification of opioid use in social media posts. Our code is available for reuse and replication: https://github.com/bionlproc/Opioid-Stage.

Keywords: Opioid use stages, contrastive learning, test time reasoning, zero shot learning

1. Introduction

The opioid crisis remains a major public health concern in the United States. It has ravaged America for over two decades with drug overdose deaths reaching a peak of over 110K in 2022, 76% of which are attributed to opioids (Spencer et al., 2024). Opioid use disorder (OUD) is a related chronic condition characterized by repeated use in the past 12 months. Specifically, OUD is defined based on a variety of DSM criteria (National Center for Injury Prevention and Control, 2024) that are further converted to ICD-10 codes. OUD diagnosis facilitates identification, tracking, and intervention for individuals whose situation may exacerbate to overdoses and more serious long term disability. With regards to both overdoses and OUD, the situation appears to have stabilized since peak levels during the Covid-19 pandemic, as per the latest report from the US Substance Abuse and Mental Health Services Administration (2025).

Treatment for OUD often involves both behavioral and pharmaceutical strategies (Wakeman et al., 2020) and both aspects are discussed by users on social media at length, especially on Reddit, which has emerged as a key platform for peer support and recovery discussions (Laud et al., 2025; Carpenter et al., 2025). Reddit has several dedicated forums (e.g., r/OpiatesRecovery) where individuals openly share experiences and seek advice. Complementing the clinical definitions of OUD, researchers have identified discrete stages of opioid use to characterize the journey of a user, warranted by the need to design inter-stage transition-specific interventions (Park et al., 2020). Yang et al. (2024) used this need to come up with six stages: Not Using, Medical Use, Misuse, Addiction, Recovery, and Relapse, following the definitions by Smith et al. (2013) and National Institute on Drug Abuse (2007).

In their seminal effort, Yang et al. (2024) created a dataset and conducted baseline experiments to classify Reddit posts discussing opioids into one of these six stages. Our current paper deals with how we can improve over their baseline experiments. At this point, we note that their dataset included a so called "explanation" for each input Reddit post indicating the spans in the input that highlight the textual evidence for the annotator's decision to assign a stage. This explanation was used as part of the input to the model (along with the original post). Since these explanations are not available at test time, they designed a strategy to extract a "silver" explanation that is generated by a different supervised model. The silver explanation, in this sense, does not need human inputs at test time. However, our methods based on recent advances in reasoning LLMs obviate this need for these silver explanations. As such, for brevity and space limitations, we do not discuss or use the silver explanations in all reported results of this paper. Due to scarcity of expert time to create training labels, Yang et al. (2024) used Amazon Mechanical Turk to create significant portions of the dataset. The entire training dataset consists of these so called "novice" labeled examples while the test set has both expert and novice labeled instances.

Some of the main challenges on this task is the lack of expert labeled training examples, nontrivial

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 207–222
12 May 2026.

semantic overlap between the stages, and limited reasoning fidelity (and hence interpretability) of encoder and encoder-decoder models typically used in classification tasks. We next outline how we address these aspects in this paper.

Accurate stage detection often requires nuanced logical deduction beyond surface-level cues. Recent reasoning-capable LLMs like DeepSeek-R1 can produce detailed chain-of-thought traces or summarized reasoning at inference time, and we already have evidence that bigger more powerful models can improve downstream model performance when distilled into smaller models (Wei et al., 2022; Hsieh et al., 2023). However, it remains unclear whether such reasoning is necessary for opioid stage classification. We use DeepSeek R1 as a teacher model to generate two forms of reasoning: (a) *summarized reasoning* – concise 2–3 sentence justifications highlighting the most salient linguistic cues, and (b) *step-by-step reasoning* – detailed internal reasoning traces capturing the model's multi-step deductions. By distilling these into smaller student models (T5-3B, T5-11B, and DeBERTa v3), we test whether explicit reasoning can improve classification accuracy.

Furthermore, Yang et al. (2024) identified that opioid use stage classification suffers from significant semantic overlap, particularly between adjacent stages like Misuse and Addiction or Recovery and Relapse, as illustrated in Table 1. To address this, we integrate supervised contrastive learning (SCL), which, as shown by Gunel et al. (2020), enhances inter-class separability by explicitly pushing apart representations of different stages in the latent space. Finally, given the scarcity of expert labels in existing opioid use corpora and the persistent discrepancy between expert and novice annotations, we reduce our reliance on noisy novice data by employing in-context learning (ICL). We use GPT-5 as a synthetic annotator grounded in a small set of expert examples to realign novice-labeled training data toward expert standards. By systematically evaluating these three strategies, our work provides a comparison of how reasoning-centric, data-centric, and representation-centric techniques impact classification across different model scales.

Our results show that the optimal reasoning format depends on the evaluation setting: summarized reasoning yields the strongest performance on the expert test set, while step-by-step reasoning achieves the best results on the novice test set. The integration of SCL improves performance by enhancing inter-class discrimination, while ICL-based realignment produces higher agreement with expert test sets. While none of our methods is a silver bullet, these techniques yield nontrivial gains in both accuracy and interpretability, demonstrating a promising pathway toward scalable, expert-aligned opioid use monitoring from social media. The dataset is only available from its creators (Yang et al., 2024) but all our code is available for reuse and replication: `https://github.com/bionlproc/Opioid-Stage`.

2. Related Works

As already discussed, Yang et al. (2024) demonstrated that explanation-augmented models improved detection of opioid use stages from Reddit posts. Earlier work examined substance-use detection more broadly from forums and Twitter, in identifying misuse, overdose risk, and relapse signals from user-generated content (Fodeh et al., 2021; Raza et al., 2023; Ahmad et al., 2026; Yang et al., 2022). Our lab has recently showed how Reddit posts can be used to glean evolving barriers to opioid recovery (Ekanayake et al., 2025).

Prior work has shown that models can benefit from human or model-generated rationales, which enhance both predictive accuracy and interpretability. Camburu et al. (2018) and Rajani et al. (2019) demonstrated that jointly training models to generate explanations alongside labels yields superior performance in reading comprehension and commonsense reasoning. Recent reasoning LLMs such as DeepSeek R1 (Guo et al., 2025) and chain-of-thought (COT) prompting have shown that decomposing complex problems into intermediate, step-by-step reasoning steps significantly improves robustness and logic in large language models (Wei et al., 2022). Subsequent work has further demonstrated that these reasoning rationales can be distilled to train smaller, more efficient, and effective task-specific models (Hsieh et al., 2023; Hancock et al., 2018). Our work applies these ideas to a high-stakes, fine-grained social health text task and compares the effectiveness of summarized vs step-by-step rationales.

Supervised contrastive learning has been shown to improve representation quality in classification problems by explicitly pulling together representations of same-label examples and pushing apart different labels (Khosla et al., 2020; Gunel et al., 2020; Kim et al., 2022). In health NLP, contrastive objectives have been used to better separate semantically adjacent classes, for example in automated diagnostic coding from clinical notes (Kailas et al., 2023). Separately, in-context learning and related prompting strategies with powerful LLMs have been used for expert-like few-shot annotation and label refinement in clinical and mental-health tasks, including clinical NER, suicidality phenotyping, and OUD identification from electronic health records (Baroian, 2025; Li et al., 2025; Huang et al., 2025; Molina et al., 2025). We build on this line of work by using GPT-5 to realign noisy novice labels

OUD Stage	Reddit Post Excerpt (Title + Text)
Medical Use	**[TITLE]** Codeine doses **[TEXT]** So I was recently prescribed Tylenol 3's with 30mg Codeine. Is it safe to take two, ingesting 60mg of codeine? Has anyone ever done it with this amount or more? Comment below!
Misuse	**[TITLE]** I'm on tramadol first time **[TEXT]** I took 350mg first time on this, it's kinda underwhelming but it's lght. I hit a juul and that took me way higher but idk I expected more from this
Addiction	**[TITLE]** Hydrocodone dosage **[TEXT]** Is 30mg too much if I just started taking it daily a week ago? I can tell my tolerance is going up so I'm just wondering if I can start to slowly increase my dosage.
Recovery	**[TITLE]** Venting **[TEXT]** I'm just ready to feel like myself again I won't give up though My life is not so bad but my mind is hopefully it a be over soon ...stay strong everybody it does get easier
Relapse	**[TITLE]** Withdrawal question please help? **[TEXT]** So I tried to get clean went 4 days and I was starting to get over the withdrawal then I used again will I feel the effects of the withdrawal all over again?

Table 1: Representative opioid use stage examples from the training dataset

toward a small expert subset and then measuring how this affects downstream models.

3. Methodology

Recall our goal is to improve the fine-grained classification of Reddit posts into six opioid use stages (*Not Using*, *Medical Use*, *Misuse*, *Addiction*, *Recovery*, and *Relapse*) using the dataset from Yang et al. (2024). The dataset is created using opioid related posts from three subreddits (r/OpiatesRecovery, r/Opiates, and r/Drugs) with sufficient accommodations to only retain opioid related posts. The inputs to the models consist of the Reddit post's title concatenated with its body text. To ensure comparisons with prior work, we followed the exact data partitioning from Yang et al. (2024):

- **Novice-annotated data**: 1936 training instances and 150 test instances,
- **Expert-annotated data**: 442 test instances (experts do not contribute to training).

All models are evaluated across the two test sets (novice and expert) to align with the evaluation framework introduced in Yang et al. (2024). The top class in novice annotated posts is *addiction* (29%) but for expert dataset it is *not using* (34%). To illustrate the nature of the self-disclosed Reddit posts and the fine-grained opioid use stages, we provide representative examples from the dataset introduced by Yang et al. (2024) in Table 1. These examples demonstrate how user narratives are categorized into distinct phases, highlighting the contextual nuances our models aim to capture.

3.1. Reasoning Distillation

Reasoning distillation aims to improve downstream model performance by providing student models with explicit intermediate rationales. These rationales reflect how a stronger teacher model justifies its predictions. To generate them, we use DeepSeek R1, a recent large language model optimized for structured reasoning (Guo et al., 2025). For each training instance, DeepSeek R1 is prompted in a zero-shot setting with the Reddit post and its ground-truth stage label, and asked to produce an explanation for why the post corresponds to that label. The full prompting template used to obtain reasoning is provided in Figure 1 of the Appendix.

3.1.1. Reasoning Formats

We generate two forms of teacher rationales:

- *Summarized reasoning*: A concise explanation (with 2-3 sentences) in which DeepSeek R1 provides a high-level justification for the label, highlighting the most salient cues or linguistic indicators present in the post. These summaries average approximately 62 words in length.
- *Step-by-step reasoning:* A detailed internal reasoning trace obtained from the model's `<think>` segment. These traces represent DeepSeek R1's hidden intermediate deliberations, including multi-step deductions, latent evidence extraction, and fine-grained decision heuristics. Compared to the summarized rationales, these internal reasoning sequences are significantly longer and more granular, av-

eraging approximately 322 words per post, reflecting the full chain-of-thought the model produces before emitting its final answer.

The two formats allow us to evaluate whether student models benefit more from concise rationales or from richer, model-internal reasoning traces.

3.1.2. Student Model Training

The reasoning outputs are incorporated directly into model supervision for T5-3B and T5-11B models (Raffel et al., 2020). During fine-tuning, the student models are trained to generate both the label y_i followed by its associated reasoning r_i (summarized or step-by-step) given input x_i, using a conditional language modeling objective:

$$\mathcal{L}_{\text{CE}} = -\sum_i \log P_\theta(y_i, r_i \mid x_i).$$

similar to other efforts (Wadhwa et al., 2023). This setup enables us to quantify the impact of distilled reasoning on classification accuracy and to compare the relative benefits of short versus detailed reasoning representations.

3.2. Supervised Contrastive Learning

To improve the separability of closely related opioid use stages, we incorporate supervised contrastive learning (SCL) as a two-stage training procedure applied uniformly across all model architectures evaluated in this work.

3.2.1. Stage 1: Contrastive Pretraining

In the first stage, the model encoder is optimized using a supervised contrastive objective. Each training batch is constructed using a balanced class sampling strategy, where a fixed number of examples are drawn from each of the 6 opioid use stages. Two examples per class are used in all models. This construction ensures that all batches contain well-defined positive and negative relationships.

View generation For each text instance x_i, we generate a single augmented variant $\tilde{x}_i$ using a random text-level operation (synonym substitution, token deletion, or token swapping). Both the original and augmented inputs are encoded independently:

$$v_i^{(1)} = f_\theta(x_i), \qquad v_i^{(2)} = f_\theta(\tilde{x}_i),$$

where $f_\theta(\cdot)$ denotes the encoder.

Projection and normalization The encoded representations are passed through a two-layer projection head (Linear $\rightarrow$ ReLU $\rightarrow$ Linear $\rightarrow$ LayerNorm). We apply L2 normalization to the resulting projection to produce unit vectors (z_i), suitable for contrastive learning.

Contrastive objective We adopt the supervised contrastive loss from Khosla et al. (2020). Let $I = \{1, \ldots, 2N\}$ be the set of indices for all augmented views in the batch. For an anchor embedding $i \in I$, let y_i be its corresponding opioid stage label. The set of positive indices is defined as

$$P(i) = \{p \in I | y_p = y_i \text{ and } p \neq i\},$$

representing all other views in the batch with the same label, while $A(i) = I \setminus \{i\}$ contains all remaining views. The loss is then defined as:

$$L = \sum_{i=1}^{2N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\text{sim}(z_i, z_p)/\tau\right)}{\sum_{a \in A(i)} \exp\left(\text{sim}(z_i, z_a)/\tau\right)},$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity computed using the dot product between L2-normalized embeddings, and $\tau = 0.07$ is the temperature parameter. This step yields an encoder whose latent space is structured to better reflect stage distinctions.

3.2.2. Stage 2: Task-specific Fine-tuning

Following contrastive pretraining, we load the resulting encoder weights into the downstream opioid stage classifier. To preserve the structure learned during SCL and mitigate catastrophic forgetting, we freeze the lower portion of the encoder and fine-tune only the upper k layers (determined based on validation experiments) alongside the prediction head (or decoder in seq2seq settings).

3.3. Zero-shot with GPT5

To establish a strong baseline we evaluated the most recent OpenAI large language models (GPT-5 and GPT-5.2) in a zero-shot setting, that is, without any task-specific fine-tuning or in-context examples. Two prompting strategies were explored: (1) Simple prompt (baseline), which directly mirrors the prompt used in the original study for zero-shot inference (Figure 3 in the Appendix). This minimal formulation tests the model's ability to map raw Reddit text to the six opioid use stages with only the label list as guidance. (2) Guideline-enhanced prompt, that incorporates the full class definitions that were provided to the human annotators in the original dataset (Figure 4 in the Appendix). Both prompts were submitted to GPT-5 and GPT-5.2 via the OpenAI API with default temperature (0) and greedy decoding to obtain a deterministic label.

3.4. Novice-to-Expert Realignment of Training Data

While reasoning distillation and supervised contrastive learning improve model learning dynamics and representation quality, their effectiveness remains ultimately bounded by the quality of the

underlying training labels. Note that the training split is exclusively novice-only; as per Yang et al. (2024) the novice turkers need only a 60% accuracy threshold, on a small sample of expert labeled instances, to be selected for this task. Thus testing on expert annotated test data is expected to give worse performance. To address this, we use few-shot in-context learning (ICL) with GPT-5 (OpenAI, 2025) to realign novice labels toward expert annotation standards.

We begin by selecting 40 expert-labeled instances, with 5-8 examples per class depending on availability. This selection counters the strong class imbalance in the expert subset, (e.g., Not Using: 34.51% vs. Medical Use: 3.52%), while ensuring that each stage is represented by multiple high-quality exemplars. From the pool of 40 expert examples, we construct 10 distinct ICL prompt sets, chosen to balance diversity of expert demonstrations with computational feasibility. Each set contains one demonstration per class (six total), providing GPT-5 with a balanced snapshot of expert decision patterns across all stages.

For each novice-labeled post, GPT-5 is prompted ten times, once with each of the ten ICL demonstration sets. Each prompt contains the full annotation gutidelines used by Yang et al. (2024), six expert few-shot exemplars, and the target Reddit post. GPT-5 is instructed to produce a JSON-formatted output containing a corrected label selected from the six opioid use stages and a rationale consisting of the minimal text span from the post that justifies the prediction. This process yields ten independent candidate labels for each instance. The final "aligned" label is determined through majority voting over these ten predictions. The complete prompt template is provided in Figure 2 of the Appendix.

The resulting relabeled dataset is used to train the same models evaluated in the main experiments, including both label-only and reasoning-augmented variants. We refer to this curated version as *ICL-aligned*. To avoid any form of data leakage, the 40 expert examples selected for the ICL demonstration sets were excluded entirely from both training and evaluation, and were removed from the original expert test split so that no item used in the ICL prompts appeared in the test set.

4. Results

We evaluate our framework across three dimensions: (1) reasoning distillation, comparing summarized vs step-by-step rationales; (2) supervised contrastive learning (SCL); and (3) expert realignment via ICL. All models were trained on novice-labeled dataset, and evaluated on both the novice and expert test sets using accuracy and macro F1 score. Each method is compared against baseline models that we reproduced (as described in Section 4.1) using the original input–output format (title + post $\rightarrow$ label), so that observed improvements can be attributed to methodological changes rather than optimization differences.

4.1. Reproduced Baselines

To enable fair comparison across all methods evaluated in this study, we first reproduced the baseline models from the original paper by Yang et al. (2024) using the same input-output with updated training configurations (see Table A in the Appendix for full hyperparameter specifications across all models). Table 4.2.1 reports the performance for all architectures, for both test sets. Most models showed modest improvements of 1-3 points over the originally reported numbers (Appendix's Table A), with T5-3B exhibiting the largest improvements on both expert and novice splits with improvements of 4-6 points. We interpret all reasoning distillation, SCL, and ICL results with respect to these reproduced baselines.

4.2. Reasoning Distillation Results

Our initial experiments evaluated the effectiveness of reasoning distillation using two distinct reasoning formats generated by DeepSeek R1: `summarized` (Table 4.2.1) and `step-by-step traces` (Table 4.2.1), evaluated for T5-3B and T5-11B.

4.2.1. Comparison with Reproduced Baselines

Relative to the reproduced baselines, reasoning distillation has a clear positive effect on the smaller T5-3B model and a consistently negative effect on T5-11B. For T5-3B, adding reasoning supervision improves performance on novice test set: step-by-step rationales increase novice F1 from 0.755 in the baseline to 0.791, an improvement of 3.6 F1 points, while summarized reasoning also improves novice F1 to 0.784. On the expert test set, both reasoning variants slightly reduce F1 relative to the baseline (0.703), producing 0.698 (step-by-step) and 0.683 (summarized), although accuracy increases marginally in both cases (1-2 percentage points). For the T5-11B model, reasoning distillation consistently degrades performance over standard fine-tuning. On the expert test set, F1 drops from 0.739 to 0.723 (summarized) and 0.674 (step-by-step), and on the novice test set from 0.814 to 0.763 and 0.809, respectively. These results indicate that, at this scale, additional reasoning supervision does not provide clear benefits and can interfere with the model's pre-trained decision boundaries.

4.2.2. Impact of Reasoning Format

The comparison between summarized and step-by-step formats reveals that the optimal reasoning

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.681	0.675
	Novice	0.707	0.712
DeBERTa-v3-Large	Expert	0.708	0.707
	Novice	0.727	0.737
T5-3B	Expert	0.690	0.703
	Novice	0.747	0.755
T5-11B	Expert	0.732	0.739
	Novice	0.813	0.814

Table 2: Reproduced baseline results for DeBERTa and T5 Models

Model	Dataset	Accuracy	F1
T5-3B	Expert	0.713	0.683
	Novice	0.787	0.784
T5-11B	Expert	0.733	0.723
	Novice	0.760	0.763

Table 3: Models trained with summarized reasoning

Model	Dataset	Accuracy	F1
T5-3B	Expert	0.708	0.698
	Novice	0.787	0.791
T5-11B	Expert	0.683	0.674
	Novice	0.807	0.809

Table 4: Models trained with step-by-step reasoning

format depends on student model size and dataset. For T5-3B, step-by-step reasoning is consistently more effective than summarized reasoning on both expert and novice test sets. On expert annotations, step-by-step achieves higher F1 than summarized (0.698 vs 0.683, +1.5 points), and on novice annotations it achieves the best overall performance with F1 0.791 compared to 0.784 for summarized reasoning. For T5-11B, step-by-step reasoning substantially outperforms summarized reasoning on the novice test set (F1 0.809 vs. 0.763), whereas summarized reasoning is less harmful on the expert set (F1 0.723 vs. 0.674), though both formats remain below the baseline. Taken together, these results show that explicit reasoning supervision is most helpful for the smaller T5-3B model, especially on the noisier novice data where step-by-step traces produce the largest gains over the baseline. For the larger T5-11B model, which already performs strongly under standard fine-tuning, reasoning distillation reduces performance on both expert and novice test sets, suggesting that reasoning supervision may be ill-suited for larger models.

4.3. Supervised Contrastive Learning

We next evaluate the impact of integrating an SCL objective on all our model architectures (Table 4.3). Each SCL-enhanced model is directly comparable to its baseline counterpart in Table 4.2.1.

SCL generally improves performance on novice-labeled test sets for smaller and mid-sized models, while its effect on very large models is mixed. For DeBERTa-v3-Base, DeBERTa-v3-Large, and T5-3B, SCL yields substantial F1 gains on novice data, whereas T5-11B experiences slight degradation in this regime. On expert test set, the impact of SCL is more modest and model-dependent.

On the expert test set, SCL produces small but noticeable changes. For T5-3B, SCL improves expert F1 from 0.703 to 0.724 (+2.1 points), alongside a corresponding 2 point increase in accuracy. DeBERTa-v3-Base and DeBERTa-v3-Large show only modest shifts in expert F1 under SCL, remaining close to their respective baselines. In contrast, T5-11B experiences a clear degradation: expert F1 decreases from 0.739 to 0.690 (-5 points), and accuracy falls from 0.732 to 0.686.

On the novice test set, the impact of SCL is considerably larger particularly for smaller and mid-sized models. For T5-3B, SCL increases novice F1 from 0.755 to 0.820 (+6.5 points), with accuracy rising from 0.747 to 0.800. DeBERTa-v3-Base and DeBERTa-v3-Large also achieve strong novice performance under SCL, with F1 scores of 0.726 and 0.756, respectively indicating robustness to noisy labels compared to their non-SCL baselines. By contrast, T5-11B exhibits a slight reduction in novice F1, from 0.814 to 0.803, suggesting that SCL is less beneficial at this scale in the realistic inference setting.

These patterns suggest that SCL is especially beneficial for DeBERTa-v3-Base and T5-3B under noisy novice supervision, where it yields multi-point F1 gains and more stable performance. Larger models such as DeBERTa-v3-Large and T5-11B exhibit more limited benefits.

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.677	0.660
	Novice	0.720	0.726
DeBERTa-v3-Large	Expert	0.710	0.703
	Novice	0.753	0.756
T5-3B	Expert	0.710	0.724
	Novice	0.800	0.820
T5-11B	Expert	0.686	0.690
	Novice	0.800	0.803

Table 5: Effect of Supervised Contrastive Learning (SCL) on model performance.

Model	Dataset	Prompt w/o guidelines		Prompt w/ guidelines	
		Accuracy	F1	Accuracy	F1
GPT-5	Expert	0.649	0.647	0.79	0.796
	Novice	0.625	0.593	0.749	0.739
GPT-5.2	Expert	0.652	0.644	0.739	0.722
	Novice	0.622	0.588	0.749	0.739

Table 6: Zero-shot performance of GPT-5 models with/without guidelines, averaged across 3 runs.

4.4. Zero-shot Performance with GPT-5

Table 4.3 lists zero-shot performances of GPT-5 and GPT-5.2 on both expert and novice test sets. With the concise prompt from the original paper, GPT-5 attained an accuracy of 0.649 and an F1 score of 0.647 on the expert set, and 0.625 accuracy (F1 0.593) on the novice set. GPT-5.2 performed similarly on both sets, achieving 0.652 accuracy (F1 0.644) for expert data and 0.622 accuracy (F1 0.588) for novice data. But introducing the elaborate guideline-enhanced prompt consistently improved performance for both GPT-5 and GPT-5.2 across both expert and novice datasets. GPT-5 achieved an F1 score of 0.796 on the expert dataset, a 15 point bump compared to the simple prompt. Notably, for GPT-5, the performance on the expert-labeled data was higher than on the novice-labeled data, even surpassing the results of the newer GPT-5.2 by 7.4 points (F1 0.722 vs 0.796). This comparison suggests that for zero-shot stage classification, GPT-5 demonstrated stronger performance on expert data with the enhanced prompting strategy compared to its successor.

4.5. Realignment via ICL

Our final set of experiments assessed whether few-shot ICL with GPT-5 could realign noisy novice-labeled data toward expert annotation standards and hence produce better performance on the expert test set.

On the expert test set, ICL-based realignment gave mixed but often positive effects across all architectures, particularly in accuracy (Table 4.5). For DeBERTa-v3-Base, expert F1 increases from 0.675 in the baseline to 0.694, a gain of around 2 points. T5-3B also shows a modest improvement in expert F1, rising from 0.703 to 0.718 (+1.5 points), with accuracy increasing from 0.690 to 0.744. Conversely, DeBERTa-v3-Large experiences a slight decline in expert F1, from 0.707 to 0.682 (-2.5 points). However, the accuracy rises by 2 points. T5-11B also sees a nearly 2 point decrease in F1 score, with the accuracy slightly increasing from 0.732 to 0.736.

However, the improvements in the expert-set performance were accompanied by substantial declines in novice test set performance. Most notably, T5-3B's accuracy dropped from 0.840 to 0.727 after ICL realignment (-11.3 points). So, ICL-based label refinement successfully achieves its stated objective, improved alignment with expert evaluation standards using only 40 expertly annotated exemplars. This efficiency is operationally valuable for scenarios where expert evaluation is the primary validation metric. The reduction in performance on the novice evaluation set is expected, as the realigned labels impose stricter expert-style decision boundaries that differ from the noisier patterns present in novice annotations.

5. Discussion

Our work addresses three persistent challenges in fine-grained opioid use stage classification from social media data: (1) limited high-quality expert annotations, (2) semantic overlap across stages, and (3) the lack of interpretable, reasoning-driven models suitable for public health applications. Our find-

Model	Dataset	Accuracy	F1
DeBERTa-v3-Base	Expert	0.734	0.694
	Novice	0.647	0.620
DeBERTa-v3-Large	Expert	0.729	0.682
	Novice	0.693	0.663
T5-3B	Expert	0.744	0.718
	Novice	0.747	0.739
T5-11B	Expert	0.736	0.720
	Novice	0.713	0.697

Table 7: Performance of models trained on ICL-realigned data.

ings showed that reasoning distillation, supervised contrastive learning, and ICL-based label realignment, each address different challenges in stage classification, with distinct benefits depending on model size and label quality.

5.1. Reasoning Distillation

The reasoning distillation experiments highlight the importance of explanation format in transferring knowledge from a larger teacher model into smaller student architectures. We find that for smaller models such as T5-3B, both summarized and step-by-step explanations provide performance gains, but their effectiveness varies by explanation quality. Summarized reasoning proves more helpful on the expert test set, while step-by-step reasoning provides the largest improvements on the noisier novice test set. We hypothesize that explicit reasoning steps help smaller models disentangle ambiguous decision boundaries by exposing the underlying decision-making process, which is valuable when training labels are noisy. These patterns suggest that detailed reasoning helps a smaller model handle noisy novice labels by exposing underlying decision patterns, whereas concise rationales better align with expert decision boundaries.

However, reasoning distillation does not uniformly benefit larger models. For T5-11B, both summarized and step-by-step explanations tend to degrade performance relative to standard fine-tuning. Summarized reasoning produces only marginal changes on the expert set and step-by-step reasoning can substantially reduce expert accuracy. Since larger models such as T5-11B already learn to infer implicit reasoning patterns during standard fine-tuning, adding explicit teacher rationales may over-constrain their learned representations or introduce conflicting supervision signals, leading to performance degradation. This trade-off suggests that reasoning augmentation is most effective for smaller models on noisy data, where the explicit signal provides genuine benefit.

To illustrate the impact of reasoning-aware supervision, we present a representative example in Figure 5 in the Appendix, where the baseline fine-tuned model fails to predict the correct stage, while the reasoning-distilled T5-3B model succeeds. In this instance, the baseline model misclassifies the post as Addiction, likely due to strong lexical associations with terms such as “morphine”, “high dose”, and “high”, which are frequently correlated with the Addiction class. This behavior indicates a reliance on surface-level cues rather than a deep understanding of the user’s behavioral context. In contrast, the model trained with summarized reasoning correctly predicts Misuse. Using rationales distilled from the DeepSeek-R1 teacher model, the student model may have internalized the logical distinctions required for this task. The rationale generated at inference time demonstrates that the model prioritizes key factors, such as non-therapeutic intent (seeking a “good buzz”) and limited usage frequency, rather than over-weighting the presence of high-risk keywords. This example highlights how reasoning distillation enables the model to capture subtle distinctions between semantically overlapping classes. In particular, it shows that the approach aligns the model’s decision boundaries with annotation guidelines by distinguishing intent-driven, non-therapeutic use (`Misuse`) from compulsive or dependence-driven behavior (`Addiction`).

5.2. Supervised Contrastive Learning

SCL consistently strengthened the discriminative capabilities of many models, with the largest benefits observed for smaller to mid-sized architectures particularly on the noisy novice test set. Both DeBERTa-v3-Base and DeBERTa-v3-Large exhibit clear novice F1 gains in this setting, indicating that SCL effectively helps untangle closely related opioid use stages such as Misuse versus Addiction or Recovery versus Relapse which often exhibit overlapping linguistic cues. T5-3B sees the largest relative improvement on novice data, substantially closing the performance gap to the T5-11B baseline. In fact, with SCL, T5-3B slightly surpasses the performance of the fine-tuned T5-11B baseline. On the expert set, SCL produced more modest and model-

dependent effects; T5-3B and DeBERTa-v3-Base showed small F1 gains, whereas DeBERTa-v3-Large was largely unchanged and very large models (T5-11B) showed a clear performance degradation. Overall, these results indicate that SCL acts as a noise-robust regularizer that is particularly helpful for moderate-capacity models trained on noisy novice annotations. The benefits tend to diminish or even reverse for very large models whose baselines are already strong, especially on expert-labeled data.

5.3. Zero-shot Performance with GPT-5

GPT-5 had substantial gains (+15 points in F1 on expert test set) with the detailed prompt containing full guidelines provided to human annotators compared to a brief prompt just listing the class names. It appears that current frontier LLMs equipped with reasoning capabilities are capable of digesting longer instructions and leverage them in excelling at off-the-shelf zero-shot tasks. More importantly, the expert set zero-shot F1 of 0.796 by GPT-5 is almost six points above the best supervised F1 of 0.739 by T5-11B. The GPT-5 score was based on average from three different runs and hence it is unlikely it's a fluke. Also, we do not suspect any label leakage from the dataset into GPT-5's training data considering the dataset creators did not publicly release the dataset and we could not find any other online availability based on Web search. Thus, we believe this could be a new frontier in the abilities of frontier LLMs for social media based analytics.

On the flip side, GPT-5 obtained considerable gains compared against the newer GPT-5.2 model. This outcome suggests that newer LLM versions do not always guarantee improved reliability or accuracy across all specialized applications. Without insight into proprietary training data or architectural changes, explaining these performance differences remains challenging. This highlights that newer LLM iterations may not consistently enhance or replicate the performance of prior versions on specific tasks.

5.4. Label Realignment via ICL

The ICL realignment experiments used GPT-5 as an offline oracle to shift novice-labeled training data towards expert annotation standards. This is operationally valuable because it avoids the need for extensive expert annotation to improve label quality. The relabeled training data improved expert-set performance for smaller models, specifically DeBERTa-v3-Base and T5-3B. However, it produced less consistent improvements for larger models. T5-11B, for instance, consistently showed lower F1 compared to its already strong baseline. As expected, models trained on the realigned labels showed reduced performance on the novice evaluation set; F1 scores dropped across architectures with significant drops in the T5 models. This reflects the substantive differences between novice and expert annotation practices, despite being provided the same guidelines.

6. Conclusion

This paper provides a comprehensive evaluation of reasoning distillation, supervised contrastive learning, and expert-guided ICL realignment for fine-grained opioid use stage classification using a new benchmark dataset introduced in 2024. This is the first effort to attempt and improve upon the results from the original paper by Yang et al. (2024). Our results show that each method offers distinct benefits, improving interpretability, latent separability, or expert alignment depending on model size and the quality of supervision. In the end, none of our methods is a silver bullet offering gains across the board for all model sizes. Surprisingly, zero shot classification with the GPT-5 model (which already has inbuilt reasoning) with the full annotation guidelines as part of the prompt improves over best supervised results. Ultimately, method selection should be driven by the target evaluation, infrastructure, and operational constraints (e.g., privacy leakage concerns might rule out API calls to frontier LLMs in the healthcare domain). Together, our findings offer practical ideas for building more reliable, interpretable, and expert-aligned public-health NLP systems under realistic constraints.

7. Limitations

Several limitations warrant consideration. First, reasoning distillation depends on the quality of teacher explanations; alternative teacher models or more structured prompting strategies may yield further improvements. Second, while SCL improves class separability for many models, its interaction with larger models (e.g., T5-11B) remains less clear, with diminishing returns on expert data. Third, the ICL realignment process depends entirely on the quality of the teacher model used. Moreover, one methodological consequence of our expert realignment procedure is that the 40 expert-annotated examples used to construct the ICL demonstration sets were removed from the original expert test split to prevent data leakage. As a result, our expert evaluation set is not identical to that used in Yang et al. (2024), and direct numerical comparison to their expert-test results should be interpreted with caution (just for the ICL results in Table 4.5). With all our strategies, while improvements on smaller models are nice, the gains are simply not there or minimal with larger models such as T5-11B. What would it take to improve larger models would be part of our future work.

8. Acknowledgments

This work is supported by the U.S. National Institute on Drug Abuse through grant R01DA057686. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

9. Bibliographical References

Muhammad Ahmad, Rita Orji, Maaz Amjad, et al. 2026. Automated risk assessment of opioid use: Analysis using pre-trained transformers on social media data. *JMIR infodemiology*, 6:e77783.

Andrei Baroian. 2025. Supervised fine-tuning or in-context learning? evaluating llms for clinical ner. *arXiv preprint arXiv:2510.22285*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, et al. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Kristy A Carpenter, Anna T Nguyen, Delaney A Smith, et al. 2025. Which social media platforms facilitate monitoring the opioid crisis? *PLOS Digital Health*, 4(4):e0000842.

Vinu Ekanayake, Md Sultan Al Nahian, and Ramakanth Kavuluru. 2025. Mining social media for barriers to opioid recovery with llms. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 83–99.

Samah Jamal Fodeh, Mohammed Al-Garadi, Osama Elsankary, et al. 2021. Utilizing a multi-class classification approach to detect therapeutic and recreational misuse of opioids on twitter. *Computers in biology and medicine*, 129:104132.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Daya Guo, Dejian Yang, Haowei Zhang, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Braden Hancock, Paroma Varma, Stephanie Wang, et al. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, et al. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Ming Huang, Zehan Li, Yan Hu, et al. 2025. Multi-label classification with generative ai models in healthcare: A case study of suicidality and risk factors. *ArXiv*, pages arXiv–2507.

Prajwal Kailas, Max Homilius, Rahul C Deo, et al. 2023. Notecontrast: Contrastive language-diagnostic pretraining for medical text. In *Machine Learning for Health (ML4H)*, pages 201–216. PMLR.

Prannay Khosla, Piotr Teterwak, Chen Wang, et al. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Jaewon Kim, Hyukjong Lee, Jooyoung Chang, et al. 2022. Generalized supervised contrastive learning. *arXiv preprint arXiv:2206.00384*.

Tanmay Laud, Akadia Kacha-Ochana, Steven A Sumner, et al. 2025. Large-scale analysis of online questions related to opioid use disorder on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1068–1084.

Zehan Li, Wanjing Wang, Lokesh Shahani, et al. 2025. Explainable suicide phenotyping from initial psychiatric evaluation notes using reasoning large language models. *medRxiv*, page 03.

Melanie Molina, Cynthia Fenton, Kathy T LeSaint, et al. 2025. Comparing computable structured phenotype-versus large language model-identification of opioid use disorder using electronic health record data. *medRxiv*, page 12.

National Center for Injury Prevention and Control. 2024. Opioid Use Disorder: Diagnosis. https://www.cdc.gov/overdose-prevention/hcp/clinical-care/opioid-use-disorder-diagnosis.html. Centers for Disease Control and Prevention. Accessed: 2026-02-25.

National Institute on Drug Abuse. 2007. *Drugs, Brains, and Behavior: The Science of Addiction*. National Institute on Drug Abuse, National Institutes of Health, U.S. Department of Health and Human Services. NIH Publication.

OpenAI. 2025. Introducing gpt-5.

Ju Nyeong Park, Saba Rouhani, LEO Beletsky, et al. 2020. Situating the continuum of overdose

risk in the social determinants of health: a new conceptual framework. *The Milbank Quarterly*, 98(3):700–746.

Colin Raffel, Noam Shazeer, Adam Roberts, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, et al. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4932–4942.

Shaina Raza, Brian Schwartz, Sahithi Lakamana, et al. 2023. A framework for multi-faceted content analysis of social media chatter regarding non-medical use of prescription medications. *BMC digital health*, 1(1):29.

Shannon M Smith, Richard C Dart, Nathaniel P Katz, 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.

Merianne R. Spencer, Matthew F. Garnett, and Arialdi M. Miniño. 2024. Drug Overdose Deaths in the United States, 2002–2022. https://www.cdc.gov/nchs/products/databriefs/db491.htm. NCHS Data Brief, no. 491. National Center for Health Statistics. Accessed: 2026-02-25.

Substance Abuse and Mental Health Services Administration. 2025. 2024 Companion Infographic Report: Results from the 2021 to 2024 National Surveys on Drug Use and Health. https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/national-releases. SAMHSA Publication No. PEP25-07-006. Center for Behavioral Health Statistics and Quality. Accessed: 2026-02-25.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Sarah E Wakeman, Marc R Larochelle, Omid Ameli, et al. 2020. Comparative effectiveness of different treatment pathways for opioid use disorder. *JAMA network open*, 3(2):e1920622.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenghao Yang, Tuhin Chakrabarty, Karli Hochstatter, et al. 2024. Identifying self-disclosures of use, misuse and addiction in community-based social media posts. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2507–2521.

Zhou Yang, Spencer Bradshaw, Rattikorn Hewett, et al. 2022. Discovering opioid use patterns from social media for relapse prevention. *Computer*, 55(2):23–33.

A. Appendix

Model	Explanation	Dataset	Accuracy	F1
Original Results (Yang et al. 2024)				
DeBERTa-v3-Large	wo_explanation	Expert	0.676	0.657
		Novice	0.740	0.744
	w_explanation	Expert	0.738	0.726
		Novice	0.813	0.819
T5-3B	wo_explanation	Expert	0.635	0.612
		Novice	0.727	0.714
	w_explanation	Expert	0.644	0.655
		Novice	0.787	0.773
T5-11B	wo_explanation	Expert	0.714	0.704
		Novice	0.809	0.815
	w_explanation	Expert	0.766	0.770
		Novice	0.840	0.840
Reproduced Results (This Work)				
DeBERTa-v3-Base	wo_explanation	Expert	0.681	0.675
		Novice	0.707	0.712
	w_explanation	Expert	0.674	0.661
		Novice	0.733	0.739
DeBERTa-v3-Large	wo_explanation	Expert	0.708	0.707
		Novice	0.727	0.737
	w_explanation	Expert	0.735	0.739
		Novice	0.793	0.800
T5-3B	wo_explanation	Expert	0.690	0.703
		Novice	0.747	0.755
	w_explanation	Expert	0.738	0.750
		Novice	0.840	0.843
T5-11B	wo_explanation	Expert	0.732	0.739
		Novice	0.813	0.814
	w_explanation	Expert	0.772	0.788
		Novice	0.853	0.854

Table 8: Comparison of Original and Reproduced Baseline Results for DeBERTa and T5 Models

DeepSeek R1 Prompt

Classify the following post into one of these categories: **"Medical Use", "Misuse", "Addiction", "Recovery", "Relapse", "Not Using"**

- **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 effects (euphoric, sedative, or anxiolytic). Misuse is not restricted to prescription opioids.
- **Addiction:** Addiction is defined as compulsive opioid use that occurs despite personal harm or negative consequences. It 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' when they discuss substances other than opioids, another person's opioid use, general opioid questions without evidence of personal use, or irrelevant information.

A reference answer is provided solely to help guide accurate reasoning toward the correct conclusion. However, the reasoning should not reference or rely on the presence of a reference answer in any explicit way.
Respond ONLY with a JSON object:

```
{
  "reasoning": "Because ...",
  "label": "Recovery"
}

Post:
{text}

Reference Answer:
{label}
```

Figure 1: The prompt used with DeepSeek R1 to obtain reasoning traces along with the label

ICL-based relabeling prompt with GPT-5

Annotation Guidelines:

- **Medical Use:** Use of prescription opioids prescribed by a medical professional for a medical condition.
- **Misuse:** Use of a substance not following medical indications or prescribed dosing. Includes non-therapeutic use (e.g., euphoria, sedation). Misuse not restricted to prescription opioids.
- **Addiction:** Compulsive opioid use despite harm or negative consequences. Includes impaired control, craving, dependence, withdrawal.
- **Recovery:** Process of change where individuals stop using opioids, improve health and wellness, and live without opioids.
- **Relapse:** Return to opioid use after an attempt to quit.
- **Not Using:** Posts about substances other than opioids (e.g., marijuana), another person's opioid use, general questions without evidence of personal use, or irrelevant info.

Rules:

1. Output must be in JSON format: label: <one of [Medical Use, Misuse, Addiction, Recovery, Relapse, Not Using]>, rationale: <exact span(s) of text from the Reddit post that justify the label>
2. Rationale must be verbatim from the post (no paraphrasing).
3. Use the minimum sufficient span that clearly justifies the label.

Prompt Template:
You are a substance use research expert tasked with labeling Reddit posts about opioid use disorder (OUD).

```
{guidelines}

### Few-shot Examples
{examples_str}

### Task
Label the following post using the specified JSON output format and
rules:

{post_text}
```

Figure 2: The prompt used to relabel the posts using GPT-5 with few-shot ICL examples.

Simple zero-shot Prompt

Given the following title and text, please identify the appropriate opioid usage category among the following types: **"Medical Use", "Misuse", "Addiction", "Recovery", "Relapse", "Not Using"**

```
{post_text}
```

Figure 3: The short prompt used with GPT-5 models in the zero-shot setting.

Guideline-enhanced zero-shot Prompt

You are a substance-use research expert tasked with labeling Reddit posts about opioid use disorder (OUD) into one of the following OUD stages.

- **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 effects (euphoric, sedative, or anxiolytic). Misuse is not restricted to prescription opioids.
- **Addiction:** Addiction is defined as compulsive opioid use that occurs despite personal harm or negative consequences. It 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' when they discuss substances other than opioids, another person's opioid use, general opioid questions without evidence of personal use, or irrelevant information.

#Task

Determine the single OUD stage that describes the author's current situation at the time of writing. Choose exactly one: **"Medical Use", "Misuse", "Addiction", "Recovery", "Relapse", "Not Using"**

#Answer

Return only the label name on one line.

{post_text}

Figure 4: The guideline-enhanced prompt used with GPT-5 models in zero-shot setting.

Representative opioid stage classification example where reasoning helps

Title: Opiates on amitriptyline

Post Text: Is there anybody who uses opiates on amitripty line? I used morphine like three times when on this drug (I take it for 2.5 months now, I take 60mg every evening) but even though I took high dose of morphine (on second try I had problems with breathing!) I don't feel good buzz, no satisfaction at all. Anyone knows how long do I have to give up the amitriptyline for the opiates to work? I have problems with sleeping and depression and amitriptyline helps me with it, but I miss my morphine high and at the moment it feels like waste of money.

Baseline prediction: Addiction

Reasoning-distilled prediction: Misuse. **Reasoning:** *The post discusses using morphine three times while on amitriptyline. The user is taking high doses of morphine (60mg) to achieve a 'good buzz', indicating non-therapeutic use for psychotropic effects. This aligns with the definition of Misuse, which involves using substances outside medical indications for purposes like achieving euphoria.*

Gold label: Misuse

Figure 5: Example illustrating reasoning leading to correct prediction compared with the baseline T5-3B.

Model	Stage	LR	Epochs	Batch	Grad Acc	Weight Decay
DeBERTa Base	Baseline	1e-5	15	8	1	–
	SCL Pretraining	3e-5	50	12	8	–
	SCL Finetuning	2e-5	10	8	1	–
DeBERTa Large	Baseline	1e-5	15	8	1	–
	SCL Pretraining	3e-5	50	12	8	–
	SCL Finetuning	2e-5	10	8	1	–
T5-3B	Baseline	1e-4	5	1	8	–
	SCL Pretraining	2e-5	50	12	8	–
	SCL Finetuning	1e-4	5	1	8	0.01
	Summarized Reasoning	1e-4	8	1	8	0.001
	Step-by-Step	1e-4	15	1	8	0.01
T5-11B	Baseline	5e-5	5	1	8	–
	SCL Pretraining	2e-5	50	12	8	–
	SCL Finetuning	1e-4	10	1	8	0.01
	Summarized Reasoning	8e-5	20	1	8	–
	Step-by-Step	5e-4	15	1	8	–

Table 9: Hyperparameter configurations for all models and training setups