Inference Gap in Domain Expertise and Machine Intelligence in Named Entity Recognition: Creation of and Insights from a Substance Use-related Dataset

Sumon Kanti Dey,[1] Jeanne M. Powell,[1] Azra Ismail,[1] Jeanmarie Perrone,[2] Abeed Sarker[1,†]

[1]*Department of Biomedical Informatics, Emory University, Atlanta, GA 30322, USA*
[2]*Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[†]*E-mail: abeed.sarker@emory.edu*

Nonmedical opioid use is an urgent public health challenge, with far-reaching clinical and social consequences that are often underreported in traditional healthcare settings. Social media platforms, where individuals candidly share first-person experiences, offer a valuable yet underutilized source of insight into these impacts. In this study, we present a named entity recognition (NER) framework to extract two categories of self-reported consequences from social media narratives related to opioid use: *ClinicalImpacts* (e.g., withdrawal, depression) and *SocialImpacts* (e.g., job loss). To support this task, we introduce RedditImpacts 2.0, a high-quality dataset with refined annotation guidelines and a focus on first-person disclosures, addressing key limitations of prior work. We evaluate both fine-tuned encoder-based models and state-of-the-art large language models (LLMs) under zero- and few-shot in-context learning settings. Our fine-tuned DeBERTa-large model achieves a relaxed token-level F_1 of 0.61 [95% CI: 0.43–0.62], consistently outperforming LLMs in precision, span accuracy, and adherence to task-specific guidelines. Furthermore, we show that strong NER performance can be achieved with substantially less labeled data, emphasizing the feasibility of deploying robust models in resource-limited settings. Our findings underscore the value of domain-specific fine-tuning for clinical NLP tasks and contribute to the responsible development of AI tools that may enhance addiction surveillance, improve interpretability, and support real-world healthcare decision-making. The best performing model, however, still significantly underperforms compared to inter-expert agreement (Cohen's kappa: 0.81), demonstrating that a gap persists between expert intelligence and current state-of-the-art NER/AI capabilities for tasks requiring deep domain knowledge. The dataset, annotation guidelines, appendix, and training scripts are publicly available to support future research.*

Keywords: Named Entity Recognition; Substance Use; Clinical Impacts; Social Impacts; In-Context Learning; Large Language Models.

1. Introduction

Nonmedical use of opioids remains a pressing public health challenge in the United States (U.S.), with more than 8.6 million Americans affected [1]. Opioid-related overdoses have consistently remained a leading cause of accidental death in adults under 45 years of age, significantly reducing the average U.S. life expectancy [2]. In addition to fatal outcomes, nonmedical

*https://github.com/SumonKantiDey/Reddit_Impacts_NER

opioid use and addiction significantly disrupts social well-being and stability. People experiencing emotional distress or social instability, such as unemployment, housing insecurity, or family disruption, are more likely to initiate or escalate opioid use [3, 4]. Communities affected by the opioid epidemic often experience increased crime, educational challenges, and economic instability, further perpetuating cycles of disadvantage [5].

Nonmedical opioid use and subsequent impacts are underreported in clinical settings due to stigma, criminalization, and distrust of healthcare systems [6], leading to negative health outcomes [6, 7, 8]. Widespread underreporting creates significant blind spots in public health surveillance, impeding efforts to detect early warning signs and deliver timely interventions. In contrast, social media sites like Reddit provide a pseudonymous environment where people feel more comfortable disclosing sensitive information. Health-related disclosures can include patterns of substance use, clinical symptoms (e.g., withdrawal), overdose risk, and co-occurring mental health issues [9, 10]. Users also describe social consequences rarely captured in electronic health records [11], such as strained relationships, family disruption, financial hardship, unemployment, and social isolation. Analyzing these clinical and social impacts of opioid use reported on social media is crucial, as research has shown that the frequency and the content of opioid-related discussions online can mirror official epidemiological trends and provide timely insights for public health surveillance and intervention [12]. The enormity of the opioid crisis, which has ravaged the U.S. for almost three decades, requires innovative solutions [13], and the relative underutilization of social media data, which contains timely information directly posted by people with lived experiences, remains an untapped opportunity.

Extracting nuanced clinical and social impacts from informal, user-generated content poses significant challenges for natural language processing (NLP) models. Such content is highly unstructured, context-dependent, and contains abbreviations and ambiguities. Understanding the content, thus, requires deep contextual knowledge (e.g., through medical expertise or lived experience), which generic NLP systems, including large language models (LLMs) typically lack. Posts frequently express subjective, emotionally charged experiences, making it difficult for models to reliably map them to predefined categories [14]. There is a critical need to bridge the gap between expert-level domain-specific knowledge and NLP model capabilities in characterizing and extracting meaningful information from such user-generated content, so that surveillance systems can be deployed at scale to inform public health strategies, intervention planning, and, ultimately, reduce the burden of the substance-related overdose epidemic [15].

As an initial exploration of this challenge, our lab introduced a named entity recognition (NER) dataset—Reddit-Impacts [16]—which was the first to capture both clinical and social dimensions of substance use. Although promising, efforts employing transformer-based models and proprietary LLMs (GPT-3.5) revealed several critical limitations in the annotation and subsequent NLP system performances. A more detailed discussion about these limitations is provided in Appendix A. To address these shortcomings and enable the development/training of more effective NLP systems, we update the data set and develop an improved processing pipeline to automatically identify the clinical and social impacts of non-medical opioid use in Reddit narratives. Our contributions are summarized below:

(1) We release **RedditImpacts 2.0**, an improved, task-specific dataset featuring detailed an-

notation guidelines, consistent entity spans, and exclusive focus on first-person narratives.
(2) We propose an encoder-based framework for accurately extracting impact-related entities from unstructured social media narratives and systematically evaluate the effectiveness of various LLMs under zero- and few-shot in-context learning (ICL) settings.
(3) We introduce custom evaluation metrics designed to effectively measure the accuracy and reliability of models, ensuring they only identify self-reported social and clinical impacts.
(4) We conduct a targeted error analysis and data efficiency evaluation, demonstrating that strong, scalable performance can be achieved with substantially reduced labeled data, supporting responsible deployment in resource-limited and underserved settings.

Collectively, these contributions advance the development of human-aligned and trustworthy NLP systems to accurately interpret first-person opioid use narratives from social media.

Findings: Our findings reveal that LLMs underperform in the token-level NER task to identify clinical and social impacts of nonmedical opioid use in social media posts, whereas encoder-based models can be fine-tuned to achieve substantially better performance. Via targeted error analysis, we highlight areas where fine-tuned encoders and LLMs differ, illustrating specific strengths and pitfalls of each. We also show that strong model performance can be achieved with significantly small training data, emphasizing the feasibility, scalability, and responsible development of fine-tuned NER models in low-resource and data settings.

2. Related Work

NER involves identifying specific terms or phrases within texts, referred to as "entities". Traditionally, NER methods have concentrated on identifying narrow sets of predefined entities. In the biomedical field, for instance, entities include genes, diseases, chemicals, and proteins [17]. Compared to general or open-domain NER tasks, biomedical NER studies have received comparatively less scholarly attention because of the paucity of labeled data, the need for specialized computational models, and gaps in evaluation and benchmarking standards [18, 19, 20]. Biomedical NER involving social media data adds an additional layer of difficulty. Texts from such sources are characteristically informal, noisy, and linguistically diverse, making automated NER particularly difficult [21, 22]. Factors such as non-standard terminology, abbreviations, lack of context, and misspellings further amplify NER challenges [22, 23].

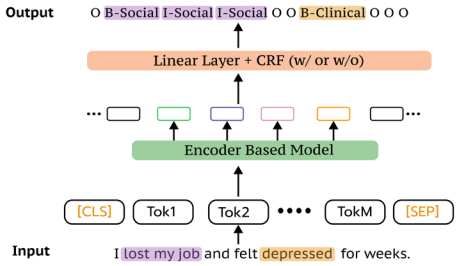
Recent advances have overcome many of the abovementioned challenges, demonstrating the successful use of advanced NER models, including neural network and transformer-based approaches, to extract entities such as drugs, diseases, and symptoms from posts on platforms like Reddit and online health forums [24, 16, 25]. Within the broader substance use sphere, transformer-based language models (BERT-base-NER, RoBERTa-base, BioBERT-base-cased, and Bio_ClinicalBERT) have been utilized to accurately identify opioid-related entities [25]. Scepanovic et al. [24] demonstrated that transformer-based models outperform traditional BiLSTM-CRF architectures in accurately extracting diverse medical entities such as symptoms, diseases, and drug names from social media posts, including Reddit.

The emergence of instruction-tuned LLMs, including Llama 3 [26], Gemma 3 [27], and GPT-4o [28], which have demonstrated exceptional performance in medical reasoning benchmarks, have also, in many cases, led to improvements in challenging NER tasks primarily

through innovative prompting strategies [29, 30, 31]. These prompting techniques enable models to generalize effectively to previously unseen scenarios with minimal contextual cues, reducing the need for extensive fine-tuning and thus facilitating efficient biomedical entity extraction. There is, however, still a research gap in NER for highly imbalanced datasets, descriptive entities, and in low-shot settings, as demonstrated by past work on the Reddit-Impacts dataset [16, 32].

3. Methodology

(a) Encoder-Based Pipeline for Impact Entity Recognition



(b) LLM-Based Pipeline for Impact Entity Recognition

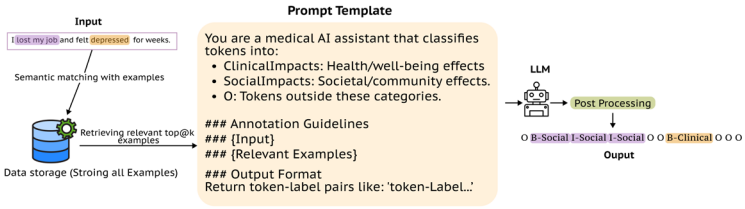


Fig. 1: NER framework for detecting *Social* and *Clinical Impacts* in user-generated text. In the output, Clinical corresponds to ClinicalImpacts and Social corresponds to SocialImpacts .

3.1. *Task Overview*

We formulate the NER task as a sequence-labeling problem using the BIO (Begin, Inside, Outside) tagging scheme. This scheme assigns one of three tags to each token: `B` to indicate the beginning of an entity, `I` for tokens that are inside an entity, and `O` for tokens that lie outside any entity span. Each tag is further associated with a specific entity type `X`, resulting in labels such as `B-X` (beginning of entity type `X`) and `O` (non-entity token).

Given a token sequence, $T = [t_1, t_2, \ldots, t_n]$ the model predicts a corresponding sequence of labels $L = [l_1, l_2, \ldots, l_n]$, where each $l_i \in \{\texttt{O}, \texttt{B-X}, \texttt{I-X}\}$. This formulation enables the model to learn both the boundary and the type of each entity span. In our study, we focus on two domain-specific entity types: **ClinicalImpacts**, representing physical or psychological consequences of substance use (e.g., *"hospitalized"*, *"depression"*), and **SocialImpacts**, which denote social, occupational, or relational consequences (e.g., *"lost job"*, *"arrested"*). An illustration of a BIO-tagged example with both entity types is shown in Table 1.

Table 1: Example of BIO-tagged tokens. The phrase *"lost my job"* is annotated as a `SocialImpacts` entity and *"depressed"* as a `ClinicalImpacts` entity.

Tokens	I	lost	my	job	and	felt	depressed	for	weeks	.
BIO Tag	O	**B-SocialImpacts**	**I-SocialImpacts**	**I-SocialImpacts**	O	O	**B-ClinicalImpacts**	O	O	O

3.2. *Data Annotation*

Named entities for very specialized problems are expected to be context-dependent, making annotation of such entities inherently complex [33]. This challenge is exacerbated in the context of social media data, where language is informal, unstructured, and shaped by personal, lived experiences. Posts related to opioid misuse often contain emotionally expressive narratives, fragmented grammar, and colloquial terms, making it difficult to consistently identify clinical and social impacts. To meet these complexities, this study included two experienced annotators with formal linguistic training. Both were guided by a comprehensive and detailed annotation manual, refined by a subject matter expert, specially designed to address the specific complexities of clinical and social impact mentions in opioid-related narratives. The full annotation guideline is provided in Appendix C.

The annotation process followed an iterative design. Initially, a 10% subset of the dataset was independently annotated by both annotators. This step was critical for aligning their interpretation of the guidelines, especially in handling variations in language and subtle distinctions in meaning. Discrepancies were resolved through discussion, leading to refinement of the guideline. We conducted iterative co-annotation until an agreement accuracy exceeding 95% was achieved. Once this threshold was met, the remaining data were divided between the two annotators for independent annotation.

We used Cohen's Kappa [34] across overlapping annotated samples to compute inter-annotator agreement. The resulting score of 81% indicated substantial agreement, reflecting almost-perfect agreement [35]. Descriptive statistics of the annotated dataset, including the total number of posts, tokens, and labeled entities, are summarized in Table 2.

Table 2: Statistics of the annotated Reddit-Impacts NER dataset.

Split	**Total Posts**	**SocialImpacts Entities**	**ClinicalImpacts Entities**	**Total Entities**	**Total Tokens**
Train	842	408	616	1024	17.2K
Dev	258	167	223	390	5.2K
Test	278	256	108	364	6.2K
Total	1378	831	947	1778	28.6K

3.3. *Experimental Setup*

For training our models, we merged the original training and development datasets to ensure we utilized the maximum amount of available data. We then set aside approximately 10% of this combined dataset as a validation set, resulting in roughly a 90/10 split between training and validation. The test set was kept separate to provide a reliable assessment of how well our models perform on completely unseen data. To support fine-tuning these models, we leveraged

a GPU equipped with 48GB of memory (e.g., NVIDIA RTX A6000). Detailed fine-tuning procedures and settings are provided in Appendix D.1.

3.4. *Modeling Approaches*

We evaluate three modeling paradigms for our SocialImpacts and ClinicalImpacts detection NER task: fine-tuning pre-trained language models (PLMs), augmenting PLMs with conditional random fields (CRF), and applying few-shot prompting techniques with large language models (LLMs). Each of these strategies has been extensively validated in recent NER research, motivating their adoption in our study [36, 37, 38, 39]. Figure 1 provides a visual summary of our overall modeling pipeline.

3.4.1. *Pre-trained language models (PLMs)*

We fine-tuned several transformer-based encoder derivatives, namely BERT [40], RoBERTa [41], DeBERTa [42], RoBERTaNER [a], and BioBERT [43]. We replace each model's standard classification head with a linear token-level classification layer predicting our BIO-formatted labels. Models are optimized using cross-entropy loss, AdamW optimizer, learning rate scheduling with linear warm-up, and early stopping based on validation-set F_1 score.

3.4.2. *PLM with CRF (PLM + CRF)*

Motivated by prior research demonstrating that adding CRF (Appendix B) layers significantly improve NER performance by capturing inter-label dependencies and ensuring valid BIO-tag transitions [44, 45], we extend each PLM by adding a CRF layer. The CRF replaces the softmax outputs and decodes the most probable sequence of labels through Viterbi decoding. Training is performed by minimizing the negative log-likelihood of the true sequences, further improving tag consistency and producing more coherent spans.

3.4.3. *Few-Shot Prompting with LLMs*

We explore zero- and few-shot prompting strategies. For each input sequence requiring labeling, we perform three key steps. First, we compute the semantic embedding of all training samples using sentence-BERT [46]. Second, we identify and select the top@K most semantically similar training examples based on cosine similarity to the input sequence. We then construct a comprehensive prompt, integrating a concise task description, explicit annotation guidelines (BIO scheme and definitions for ClinicalImpacts and SocialImpacts), the top@K dynamically selected exemplar token-label sequences, and the target input tokens to be annotated. This in-context learning approach leverages the robust semantic understanding and few-shot inference capabilities of LLMs, eliminating the need for task-specific parameter updates and providing a flexible inference method suited for low-resource or rapid-deployment scenarios. The prompt is detailed in Appendix D.2.

3.5. *Evaluation Metrics: Relaxed F_1 Score*

We evaluate model performance using a relaxed, token-level F_1 scoring method designed for entity recognition tasks involving partial matches. Our approach is inspired by partial-matching

[a]https://huggingface.co/Jean-Baptiste/roberta-large-ner-english

techniques used in biomedical NLP tasks [47, 48]. Unlike strict span-level matching, which considers only exact boundary matches, relaxed evaluation accounts for partial overlap between predicted and ground truth entity spans of the same type, capturing cases where predictions are approximately correct. This evaluation is especially suitable for informal, user-generated text where exact span boundaries may be difficult to capture. The specific evaluation strategy formulation is outlined below.

Let each labeled sequence consist of predicted and gold spans represented as:

$$G = \{g_1, g_2, \ldots, g_m\}, \quad P = \{p_1, p_2, \ldots, p_n\},$$

where each span g_i or p_j is defined as a triple (t, s, e), denoting the entity type t, start token index s, and end token index e.

We define token-level overlap between two spans g_i and p_j of the same entity type as:

$$\text{Overlap}(g_i, p_j) = \max(0, \min(e_g, e_p) - \max(s_g, s_p) + 1),$$

where (s_g, e_g) and (s_p, e_p) are the token span boundaries of the gold and predicted entities, respectively.

For each entity type T, we compute:

$$\text{TP}_T = \sum_{(g,p)\in\mathcal{M}_T} \text{Overlap}(g, p); \text{P}_T = \sum_{p_j\in P_T} (e_p - s_p + 1); \text{R}_T = \sum_{g_i\in G_T} (e_g - s_g + 1)$$

where $\mathcal{M}_T$ denotes the set of span pairs (g, p) of type T with non-zero token-level overlap.

Precision, recall, and F_1 score for entity type T are then defined as:

$$\text{Precision}_T = \frac{\text{TP}_T}{\text{P}_T}, \quad \text{Recall}_T = \frac{\text{TP}_T}{\text{R}_T}, \quad \text{F1}_T = \frac{2 \cdot \text{Precision}_T \cdot \text{Recall}_T}{\text{Precision}_T + \text{Recall}_T}.$$

We additionally compute a micro-averaged overall F_1 score by aggregating overlapping token counts across all entity types:

$$\text{Overall Precision} = \frac{\sum_T \text{TP}_T}{\sum_T \text{P}_T}, \quad \text{Overall Recall} = \frac{\sum_T \text{TP}_T}{\sum_T \text{R}_T},$$

$$\text{Overall F}_1 = \frac{2 \cdot \text{Overall Precision} \cdot \text{Overall Recall}}{\text{Overall Precision} + \text{Overall Recall}}$$

This relaxed overlap-based metric for NER is tailored to informal user-generated content, where exact boundary prediction is often difficult.

3.6. *Error Analysis*

To better understand the strengths and limitations of our models, we conducted a qualitative error analysis comparing the fine-tuned DeBERTa-large model (the best-performing model among fine-tuned PLMs) and GPT-4o with 3-shot prompting (the top-performing model among LLMs) on the task of extracting social and clinical impacts.

4. Results

Table 3 summarizes the overall token-level relaxed precision, recall, and F_1 scores with 95% confidence intervals (CI) for all models described in Section 3.4.

Among fine-tuned PLMs (Table 3(a)), **DeBERTa-large** achieved the highest overall performance, with an F_1 score of **0.61** (95% CI: [0.43, 0.62]), precision **0.75**, and recall 0.52, demonstrating significant improvement over prior approaches reported on this difficult NER task. Incorporating a CRF layer yielded mixed results across models. For DeBERTa-large, the F_1 score was lower with CRF at 0.57 [0.43-0.59], with overlapping confidence intervals suggesting statistically insignificant differences. For BERT-large, adding CRF similarly reduced F_1 from 0.56 [0.38–0.56] to 0.52 [0.36–0.52]. In contrast, RoBERTaNER-large with CRF achieved an F_1 of 0.56 [0.45–0.60], slightly higher than its no-CRF counterpart at 0.51 [0.39–0.53], reflecting modest improvements in recall. Overall, the effect of CRF was minimal and appeared to depend on the underlying architecture. BioBERT models generally achieved lower F_1 scores compared to other PLMs, reflecting possible limitations in transferring their specialized clinical language understanding to the informal language context of social media.

Table 3(b) presents the in-context learning performance of LLMs under zero-, 3-, and 5-shot settings. Among these LLMs, **GPT-4o** achieved the best overall few-shot prompting performance, with an F_1 score of **0.44** [95% CI: 0.37, 0.51] in the 3-shot setting, demonstrating balanced precision (0.42) and recall (0.46). Llama 3-70b also performed competitively, reaching an F_1 score of 0.41 in the 3- and 5-shot settings. Gemma 3-27b, while lower in absolute performance, showed a clear benefit from in-context examples, improving from an F_1 score of 0.32 [0.27-0.40] in a zero-shot setting to 0.34 [0.33-0.45] in the 3-shot setting.

These results collectively demonstrate that few-shot prompting using semantically similar examples enhances the performance of LLMs on token-level prediction tasks compared to zero-shot settings. Nevertheless, even the best-performing LLM (GPT4o, $F_1 = 0.44$) remained below the best fine-tuned encoder (DeBERTa-large, F_1=0.61), emphasizing the advantage of domain-specific fine-tuning for token-level NER tasks in this domain.

We also report entity-specific token label performance for the best-performing fine-tuned PLM (DeBERTa-large) and the top-performing few-shot LLM (GPT-4o with 3-shot prompting) in Table 4. In these two combinations, DeBERTa-large yielded the strongest performance on both ClinicalImpacts ($F_1 = 0.60$) and SocialImpacts ($F_1 = 0.50$). GPT-4o (3-shot) performed comparably on ClinicalImpacts ($F_1 = 0.51$), but underperformed on SocialImpacts ($F_1 = 0.26$), indicating varying levels of difficulty across entity types.

4.1. *Qualitative Error Analysis Results*

We identified four major categories of errors: label confusion, missed implicit entities, false positives due to negation/context errors, and violations of annotation guidelines.

Label confusion (Social vs. Clinical): GPT-4o frequently struggled to differentiate between social and clinical contexts. For instance, in the sentence *"When I was 21... on a therapeutic community"*, the model misclassified the phrase *"therapeutic community"* as a SocialImpact, whereas it was correctly annotated as a ClinicalImpact.

Missed implicit entities: Both models exhibited limitations in capturing impacts that

Table 3: Token-level relaxed precision, recall, and F_1 scores with 95% confidence intervals (CI) across two evaluation settings: **(a)** *Fine-tuned PLMs*–evaluated with and without Conditional Random Fields (CRF). **(b)** *In-Context Learning Performance of LLMs*–evaluated under zero-shot, 3-shot, and 5-shot settings using the most similar examples for prompting. The best-performing model is highlighted in **bold**, and the second-best is underlined.

(a) Fine-tuned Pretrained Language Models (PLMs).

Model	Precision	Recall	F_1	95% CI
BERT(large-uncased)	0.65	0.49	0.56	[0.38, 0.56]
BERT(large-uncased) + CRF	0.67	0.42	0.52	[0.36, 0.52]
RoBERTaNER(large)	0.67	0.42	0.51	[0.39, 0.53]
RoBERTaNER(large) + CRF	0.59	**0.54**	0.56	[0.45, 0.60]
DeBERTa(large)	**0.75**	0.52	**0.61**	[0.43, 0.62]
DeBERTa(large) + CRF	0.68	0.50	0.57	[0.43, 0.59]
BioBERT(large-cased)	0.60	0.39	0.47	[0.34, 0.52]
BioBERT(large-cased) + CRF	0.63	0.42	0.51	[0.35, 0.53]

(b) In-Context Learning Performance of LLMs under Zero-, 3-, and 5-shot Settings.

Model	Prompting	Precision	Recall	F_1	95% CI
Gemma 3-27b-it	0-shot	0.29	0.35	0.32	[0.27, 0.40]
	3-shot	0.26	**0.48**	0.34	[0.33, 0.45]
	5-shot	0.25	**0.48**	0.33	[0.32, 0.46]
Llama 3-70b instruct	0-shot	0.46	0.27	0.34	[0.28, 0.37]
	3-shot	**0.47**	0.37	0.41	[0.35, 0.45]
	5-shot	0.46	0.37	0.41	[0.33, 0.48]
GPT-4o	0-shot	0.43	0.37	0.40	[0.31, 0.43]
	3-shot	0.42	0.46	**0.44**	[0.39, 0.51]
	5-shot	0.39	**0.48**	0.43	[0.37, 0.51]

Table 4: Entity-specific token-level relaxed precision, recall, F_1-score with 95% CI for the top-performing fine-tuned PLM (DeBERTa-large) and few-shot LLM (GPT-4o, 3-shot).

Model	**Entity**	**Precision**	**Recall**	**F_1-score**	**95% CI**
DeBERTa(large)	SocialImpacts	0.63	0.41	0.50	[0.25, 0.62]
	ClinicalImpacts	0.80	0.56	0.66	[0.45, 0.67]
GPT-4o(3-shot)	SocialImpacts	0.28	0.25	0.26	[0.18, 0.34]
	ClinicalImpacts	0.46	0.56	0.51	[0.44, 0.60]

were implied rather than explicitly stated. In the sentence *"I was shocked... when I told them about my addiction and that I was seeking/needed help"*, DeBERTa successfully identified the ClinicalImpact *"addiction"*, but failed to detect the SocialImpact embedded in the phrase

"seeking/needed help". GPT-4o similarly failed, incorrectly labeling isolated words and missing the intended SocialImpacts entirely.

False positives due to context or negation errors: Both models occasionally produced false positives when failing to interpret negation or surrounding context accurately. In the sentence *"I am a recovering heroin addict with no criminal record but..."*, both DeBERTa and GPT-4o incorrectly labeled the phrase *"criminal record"* as a SocialImpact, despite the explicit negation ("no criminal record").

Guideline violations and overgeneralization errors: GPT-4o sometimes failed to adhere to annotation guidelines, including the instruction to annotate only first-person experiences. For example, in the sentence *"Helps with the restlessness and anxiety"*, the model labeled *"restlessness"* and *"anxiety"* as ClinicalImpacts, despite the absence of first-person framing. Moreover, GPT-4o exhibited overgeneralization, labeling ambiguous or emotionally charged terms, such as *"blindsided"*, *"pressuring"*, and *"treatment"*, as ClinicalImpacts, even when these were contextually neutral or not linked to direct impacts.

5. Discussion

5.1. *Bridging Expert Knowledge and Model Intelligence Remains a Challenge for Complex NER*

While our NER models demonstrated improvement over past efforts on the same, challenging dataset, there remains a significant gap between human-level agreement and model performance. Systematically improving the annotation guideline improved IAA, but models failed to fully capture the nuances of first-person reporting of clinical and social impacts of substance use. While our study explored strategies to bridge this gap using the Reddit-Impacts dataset, it generalizes to other NER tasks where deep domain expertise is necessary and in problems where annotated examples are lexically diverse and low due to dataset characteristics. In the context of substance use surveillance from social media data, improving NER performance may lead to the timely detection of emerging impacts of the continuously evolving overdose epidemic in the U.S. Our work may also shed light on how social media content reflects and shapes public perceptions of substance use over time, helping identify high-risk groups and the types of content they engage with, thereby informing development of more targeted and effective prevention strategies. Automated impact detection can further support the development of real-time support systems that connect users with peer or professional support during moments of crisis, thus strengthening clinical decision-making and public health intervention.

5.2. *Error Analysis*

Label confusion, prominent in GPT-4o, highlights its difficulty in distinguishing between structured clinical environments and broader social circumstances, particularly when terminologies overlap, although these subtle differences are evident to human experts. GPT-4o's overgeneralization errors also revealed its difficulty in handling subtle context-dependent biomedical terms. Model errors in detecting implicit entities revealed challenges in grasping context-dependent, nuanced language. Labeling errors in both models' ability to process negated constructs and assess contextual validity prior to labeling. Overall, while GPT-4o offers impressive generalization through prompting, it tends to misinterpret nuanced or implied information,

and is more prone to guideline violations and contextual errors. In contrast, the fine-tuned DeBERTa-large model demonstrated stronger alignment with domain-specific guidelines and span accuracy. It may be possible to employ ensembling approaches, such as adding a first-person report classifier prior to processing by an LLM, rather than a single-module pipeline to improve inference performance.

5.3. *Impact of In-Context Learning in LLMs*

Our findings indicate that ICL offers marginal performance gains in certain cases. For example, using 3-shot prompting resulted in a modest improvement of less than 4% in F_1 score across all LLMs compared to the zero-shot setting. However, when we increased the number of examples to 5, the overall performance declined by ≈1–2%, suggesting that additional examples may introduce noise or confuse the model's decision-making process in the NER detection task. The differences were not statistically significant, as indicated by overlapping 95%-CIs. Overall, we observe that while ICL may enhance performance to a limited extent, particularly when using well-selected and minimal examples, increasing the amount of text in the context window may overwhelm the model, leading to reduced inference performance.

5.4. *Sample Size Consideration for Fine-Tuning a Language Model*

Training a model for specialized tasks can be expensive due to the need for domain expert annotation—leading to substantial time and financial costs [49, 50]; expert annotators take 10-30 seconds per sentence to label named entities [51]. These challenges are further amplified for of user-generated biomedical data, which can be context-dependent, and emotionally nuanced, and where the demand for annotation quality is especially high. Consequently, developing models that can perform well even with limited annotated data is paramount.

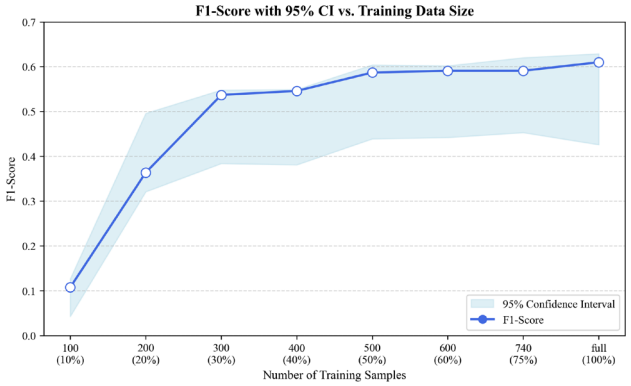


Fig. 2: F_1 score with 95% CIs across training data sizes. x-axis shows the number of training samples with the corresponding percentage of the full dataset; shaded area represents the 95% CI for each point.

We investigated the amount of annotated data required to achieve expert-level performance using the DeBERTa-large model, which attained the highest F_1 score. Our findings (Figure 2) demonstrate that F_1 score plateaus at approximately 50% of training data, with additional data yielding only minor improvements. This suggests that better strategies for incorporating domain knowledge may be more effective in improving performance than annotating more data.

5.5. *Domain-Specific Fine-Tuning Still Matters for Biomedical NER*

Our findings reinforce the growing evidence that prompting alone is insufficient for achieving state-of-the-art performance in biomedical NER. Despite the flexibility and generalization capabilities of LLMs, models like GPT-4o and Llama 3-70B underperformed compared to

domain-specific encoders fine-tuned on task-relevant labeled data. In our experiments, even the best-performing LLM using in-context learning (few-shot prompting) failed to match the F_1 score of the top fine-tuned encoder model, DeBERTa-large. These results align with some recent studies showing that LLMs, when used without task-specific fine-tuning, struggle to surpass traditional encoder architectures in specialized biomedical settings [39, 36]. One underlying limitation is that token-level NER tasks—particularly in biomedical contexts—require models that can interpret the structured label schemes such as BIO tags and manage domain-specific terminology, which may include rare and unseen tokens. Prompting strategies, which do not modify the model's internal weights, are often insufficient to capture these nuances. In contrast, fine-tuning involves updating the model's parameters through supervised training, which allows the model to learn domain-specific representations and labeling conventions more effectively. The relatively stable and accurate performance of fine-tuned PLMs in structured biomedical NER tasks highlights the continued importance of fine-tuning, particularly in domains where general-purpose LLMs still fall short.

5.6. *Limitations*

We acknowledge some limitations in our study. First, the dataset used for training is relatively small, resulting in suboptimal performance. Second, while our fine-tuned encoder-based models achieved strong performance, they are still struggle with capturing nuanced, context-dependent expression–especially when entities are implied rather than explicitly stated. Incorporating additional contextual signals or auxiliary tasks could further enhance the performance of the model. We also observed that social impacts are harder to detect than clinical impacts. This might be due to the fact that social impact expressions have higher lexical variability and non-clinical concepts are not as well defined as clinical ones. Lastly, though we evaluated LLMs under zero- and few-shot in-context learning settings, we did not explore instruction-tuned version of these models or feedback from error analysis to refine prompts. In future work, we aim to explore these approaches, along with multi-agent architectures, to better adapt LLMs for domain-specific entity recognition applicable to other types of texts (e.g., clinical notes).

6. Conclusion

We explored strategies for extracting social and clinical impacts from first-person narratives related to substance use, using the refined and high-quality RedditImpacts 2.0 dataset. Our best-performing fine-tuned encoder model, DeBERTa-large, achieved a relaxed token-level F_1 score of 61%, significantly outperforming the best-performing LLM (GPT-4o), which achieved an F_1 score of 44%—a 17% gap. This performance gap highlights the current limitations of prompting-based LLMs in specialized NER tasks without fine-tuning. Furthermore, our data efficiency analysis revealed that using only 50% of the balanced dataset was sufficient to achieve performance comparable to training on the full dataset, suggesting that high-quality, domain-specific annotations, even in moderate quantities, can yield strong results.

Acknowledgments

This publication was supported by the National Institute on Drug Abuse (NIDA) of the National Institutes of Health (NIH); award number R01DA057599. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] American Psychiatric Association, Opioid use disorder (2023), Accessed: 2025-06-06.

[2] Mayo Clinic, The role of healthcare professionals in opioid addiction prevention (n.d.), Accessed: 2025-06-06.

[3] M. Cerdá, N. Krawczyk, L. Hamilton, K. E. Rudolph, S. R. Friedman and K. M. Keyes, A critical review of the social and behavioral contributions to the overdose epidemic, *Annual review of public health* **42**, 95 (2021).

[4] C. Lin, S. J. Cousins, Y. Zhu, S. E. Clingan, L. J. Mooney, E. Kan, F. Wu and Y.-I. Hser, A scoping review of social determinants of health's impact on substance use disorders over the life course, *Journal of substance use and addiction treatment* , p. 209484 (2024).

[5] R. Darolia and C. Heflin, The social and community consequences of the opioid epidemic, *The ANNALS of the American Academy of Political and Social Science* **703**, 7 (2022).

[6] C. Strike, S. Robinson, A. Guta, D. H. Tan, B. O'Leary, C. Cooper, R. Upshur and S. Chan Carusone, Illicit drug use while admitted to hospital: Patient and health care provider perspectives, *Plos one* **15**, p. e0229713 (2020).

[7] S. Cooper and S. Nielsen, Stigma and social support in pharmaceutical opioid treatment populations: A scoping review, *International Journal of Mental Health and Addiction* **15**, 452 (2017).

[8] A. Cheetham, L. Picco, A. Barnett, D. I. Lubman and S. Nielsen, The impact of stigma on people with opioid use disorder, opioid treatment, and policy, *Substance abuse and rehabilitation* , 1 (2022).

[9] U. Lokala, O. C. Phukan, T. G. Dastidar, F. Lamy, R. Daniulaityte and A. Sheth, Detecting substance use disorder using social media data and the dark web: time-and knowledge-aware study, *JMIRx Med* **5**, p. e48519 (2024).

[10] A. Sarker, A. DeRoos and J. Perrone, Mining social media for prescription medication abuse monitoring: a review and proposal for a data-centric framework, *Journal of the American Medical Informatics Association* **27**, 315 (2020).

[11] S. Giorgi, D. B. Yaden, J. C. Eichstaedt, L. H. Ungar, H. A. Schwartz, A. Kwarteng and B. Curtis, Predicting us county opioid poisoning mortality from multi-modal social media and psychological self-report data, *Scientific reports* **13**, p. 9027 (2023).

[12] K. A. Carpenter, A. T. Nguyen, D. A. Smith, I. A. Samori, K. Humphreys, A. Lembke, M. V. Kiang, J. C. Eichstaedt and R. B. Altman, Which social media platforms facilitate monitoring the opioid crisis?, *PLOS Digital Health* **4**, p. e0000842 (2025).

[13] K. Humphreys, R. Saitz, N. D. Volkow, C. L. Shover, T. F. Babor, B. G. Carr, W. N. Evans, D. A. Fiellin, S. A. Glantz, W. Hall, D. Heller, D. H. Jernigan, R. J. MacCoun, B. K. Madras, S. Satel, B. Vicknasingam, S. E. Wakeman, R. West, D. P. Wilson, D. Ziedonis, L. R. Zindel, P. Das, O. A. Olukoya, S. L. Proctor, S. J. Tye, E. Wakeman and C. H. Wilkins, Responding to the opioid crisis in north america and beyond: recommendations of the stanford–lancet commission, *The Lancet* **399**, 555 (Feb 2022).

[14] G. Coppersmith, R. Leary, P. Crutchley and A. Fine, Natural language processing of social media as screening for suicide risk, *Biomedical informatics insights* **10**, p. 1178222618792860 (2018).

[15] D. Beyer, Joint economic committee releases new report on the economic toll of the opioid crisis (2023), Accessed: 2025-06-15.

[16] Y. Ge, S. Das, K. O'Connor, M. A. Al-Garadi, G. Gonzalez-Hernandez and A. Sarker, Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media, *arXiv preprint arXiv:2405.06145* (2024).

[17] H. Cho and H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, *BMC bioinformatics* **20**, 1 (2019).

[18] X. Dai, Recognising biomedical names: Challenges and solutions, *arXiv preprint arXiv:2106.12230* (2021).
[19] L. Luo, C.-H. Wei, P.-T. Lai, R. Leaman, Q. Chen and Z. Lu, Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning, *Bioinformatics* **39**, p. btad310 (2023).
[20] S. Liu, A. Wang, X. Xiu, M. Zhong, S. Wu *et al.*, Evaluating medical entity recognition in health care: Entity model quantitative study, *JMIR Medical Informatics* **12**, p. e59782 (2024).
[21] P. Karisani and E. Agichtein, Did you really just have a heart attack? towards robust detection of personal health mentions in social media, in *Proceedings of the 2018 World Wide Web Conference*, 2018.
[22] A. Magge, A. Klein, A. Miranda-Escalada, M. A. Al-Garadi, I. Alimova, Z. Miftahutdinov, E. Farre, S. Lima-López, I. Flores, K. O'Connor *et al.*, Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021, in *Proceedings of the sixth social media mining for health (# SMM4H) workshop and shared task*, 2021.
[23] D. Weissenbacher, A. Sarker, A. Magge, A. Daughton, K. O'Connor, M. Paul and G. Gonzalez, Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019, in *Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task*, 2019.
[24] S. Scepanovic, E. Martin-Lopez, D. Quercia and K. Baykaner, Extracting medical entities from social media, in *Proceedings of the ACM conference on health, inference, and learning*, 2020.
[25] G. Sidorov, M. Ahmad, I. Ameer, M. Usman and I. Batyrshin, Opioid named entity recognition (oner-2025) from reddit, *arXiv preprint arXiv:2504.00027* (2025).
[26] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, The llama 3 herd of models, *arXiv preprint arXiv:2407.21783* (2024).
[27] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, Gemma 3 technical report, *arXiv preprint arXiv:2503.19786* (2025).
[28] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, Gpt-4o system card, *arXiv preprint arXiv:2410.21276* (2024).
[29] Z. Zhan, S. Zhou, M. Li and R. Zhang, Ramie: Retrieval-augmented multi-task information extraction with large language models on dietary supplements, *Journal of the American Medical Informatics Association* **32**, 545 (March 2025).
[30] W. Zhou, S. Zhang, Y. Gu, M. Chen and H. Poon, Universalner: Targeted distillation from large language models for open named entity recognition, *arXiv preprint arXiv:2308.03279* (2023).
[31] C. Shyr, Y. Hu, L. Bastarache, A. Cheng, R. Hamid, P. Harris and H. Xu, Identifying and extracting rare diseases and their phenotypes with large language models, *Journal of Healthcare Informatics Research* **8**, 438 (2024).
[32] Y. Ge, Y. Guo, S. Das, M. A. Al-Garadi and A. Sarker, Few-shot learning for medical text: A review of advances, trends, and opportunities, *Journal of Biomedical Informatics* **144**, p. 104458 (August 2023), Epub 2023 Jul 23.
[33] N. Ding, G. Xu, Y. Chen, X. Wang, X. Han, P. Xie, H.-T. Zheng and Z. Liu, Few-nerd: A few-shot named entity recognition dataset, *arXiv preprint arXiv:2105.07464* (2021).
[34] J. Cohen, A coefficient of agreement for nominal scales, *Educational and psychological measurement* **20**, 37 (1960).
[35] A. J. Viera and J. M. Garrett, Understanding interobserver agreement: the kappa statistic, *Family Medicine* **37**, 360 (May 2005).
[36] Q. Lu, R. Li, A. Wen, J. Wang, L. Wang and H. Liu, Large language models struggle in token-level clinical named entity recognition, in *AMIA Annual Symposium Proceedings*, 2025.
[37] I. Keraghel, S. Morbieu and M. Nadif, A survey on recent advances in named entity recognition, *arXiv preprint arXiv:2401.10825* (2024).

[38] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li and G. Wang, Gpt-ner: Named entity recognition via large language models, *arXiv preprint arXiv:2304.10428* (2023).
[39] M. S. Obeidat, M. S. A. Nahian and R. Kavuluru, Do llms surpass encoders for biomedical ner?, *arXiv preprint arXiv:2504.00664* (2025).
[40] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019.
[41] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
[42] P. He, X. Liu, J. Gao and W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, *arXiv preprint arXiv:2006.03654* (2020).
[43] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* **36**, 1234 (2020).
[44] T. H. Dang, H.-Q. Le, T. M. Nguyen and S. T. Vu, D3ner: biomedical named entity recognition using crf-bilstm improved with fine-tuned embeddings of various linguistic information, *Bioinformatics* **34**, 3539 (2018).
[45] Y. Qiu, L. Dong, W. Zhang, H. Xing and J. Huang, A diffusion enhanced crf and bilstm framework for accurate entity recognition, *Scientific Reports* **15**, p. 19670 (2025).
[46] N. Reimers and I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, *arXiv preprint arXiv:1908.10084* (2019).
[47] G. H. B. Andrade, S. Yada and E. Aramaki, Comparative evaluation of boundary-relaxed annotation for entity linking performance, in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.
[48] I. Segura-Bedmar, P. Martínez and M. Herrero-Zazo, Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013), in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 2013.
[49] Z. P. Majdik, S. S. Graham, J. C. Shiva Edward, S. N. Rodriguez, M. S. Karnes, J. T. Jensen, J. B. Barbour and J. F. Rousseau, Sample size considerations for fine-tuning large language models for named entity recognition tasks: methodological study, *Jmir ai* **3**, p. e52095 (2024).
[50] I. Lopez, A. Swaminathan, K. Vedula, S. Narayanan, F. Nateghi Haredasht, S. P. Ma, A. S. Liang, S. Tate, M. Maddali, R. J. Gallo *et al.*, Clinical entity augmented retrieval for clinical information extraction, *npj Digital Medicine* **8**, p. 45 (2025).
[51] Y. Chen, T. A. Lask, Q. Mei, Q. Chen, S. Moon, J. Wang, K. Nguyen, T. Dawodu, T. Cohen, J. C. Denny *et al.*, An active learning-enabled annotation system for clinical named entity recognition, *BMC medical informatics and decision making* **17**, 35 (2017).