The Point of View of a Sentiment: Towards Clinician Bias Detection in Psychiatric Notes

Alissa Valentine[*†], **Lauren Lepow**[†], **Lili Chan**[†], **Alex Charney**[†], **Isotta Landi**[†]
[*]Copenhagen University [†]Mount Sinai School of Medicine
alissa.valentine@di.ku.dk

Abstract

Negative patient descriptions and stigmatizing language can contribute to generating healthcare disparities in two ways: (1) read by patients, they can harm their trust and engagement with the medical center; (2) read by physicians, they may negatively influence their perspective of a future patient. In psychiatry, the patient-clinician therapeutic alliance is a major determinant of clinical outcomes. Therefore, language usage in psychiatric clinical notes may not only create healthcare disparities, but also perpetuate them. Recent advances in natural language processing systems have facilitated the efforts to detect discriminatory language in healthcare. However, such attempts have only focused on the perspectives of the medical center and its physicians. Considering both physicians' and non-physicians' subjective points of view is a more equitable approach to identifying harmful language in clinical notes. By leveraging large language models (LLMs), this work aims to characterize potentially harmful language usage in psychiatric notes by identifying the sentiment expressed in sentences describing patients based on the reader's point of view. First, we curated a psychiatric lexicon containing words commonly used to describe patients in psychiatry. Sentences (N=39) were extracted from clinical text containing psychiatric lexicon at a medical center, with which a set of physicians (N=10) and non-physicians (N=10) annotated them as negative, neutral, or positive. Three LLMs (GPT-3.5, Llama-3.1, and Mistral) used zero-shot/few-shot in-context learning (ICL) approaches to classify the sentiment of the sentences according to the physician or non-physician point of view. Results showed that GPT-3.5 aligned best to physician point of view and Mistral aligned best to non-physician point of view, both with an ICL approach. These results underline the importance of recognizing subjectivity in clinical annotation tasks, not only for improving the note writing process, but also for the quantification, identification, and reduction of bias in computational systems for downstream analyses.

Keywords: sentiment analysis, psychiatry, subjectivity, bias detection, LLMs, clinical annotation

1. Introduction

Psychiatric notes document the clinical signs, symptoms, and behaviors of patients from the perspective of the physician. When documenting the clinical encounter, the language used by clinicians can be classified as neutral, negative, or positive (Park et al., 2021). Negative patient descriptors include those that question patient credibility, reasoning, insight, or judgment; portray the patient as noncompliant or as a threat; remark on the patient's poor self-care; or generally conveys disapproving feelings towards the patient and their presentation. In contrast, positive patient descriptors include patient strengths, minimization of blame, and language that conveys of approval and positive feelings towards the patient and their presentation.

The use of negative language in clinical notes carries two distinct downstream harms which may sway patient outcomes. When read by future providers, harmful language use can impact their perspective of a patient and decrease their quality of care (P Goddu et al., 2018). When read by patients, inaccurate or negative patient descriptions foster mistrust and harms the therapeutic alliance (Fernández et al., 2021; Flickinger et al., 2013; Silva et al., 2023), a major determination of positive outcomes in psychiatry (Flückiger et al., 2018; Horvath and Symonds, 1991; Karver et al., 2018). Therefore, in both scenarios, while not necessarily intended by the writer, there is potential for harm by perpetuating biases from the medical system, in a way that hinders mutual engagement and connection between patients and their physicians.

With the ubiquitous use of natural language processing (NLP) systems in healthcare research (He et al., 2023), the use of harmful language in clinical notes also threatens equitable deployment of artificial intelligence (AI) in medical contexts. Recent work has shown that societal biases are often reflected in AI. Namely, if biased language is embedded in the corpora used to train large language models (LLMs), models learn to perpetuate societal biases across gender, language, race, ethnicity, and insurance status (Navigli et al., 2023; Zhang et al., 2020; Omar et al., 2025). Without taking appropriate action, these models risk contributing to perpetuating health disparities, as seen in the statistically different performance rates on clinical prediction tasks between sociodemographic groups (Omar et al., 2025; Zack et al., 2024; Pal et al., 2023).

To this end, we investigated how LLMs capture the subjective point of view of physicians and non-physicians towards sentences containing patient descriptions from psychiatric clinical notes. Using

Proceedings of Identity-Aware AI (IAAI) @ LREC 2026, pages 1–11
16 May 2026.

a sentiment analysis task, we explored which language models performed best at classifying sentences as negative, neutral, or positive from the physician or non-physician point of view. We built upon a real-world lexicon of patient descriptions used in naturalistic settings to aid our work (Sun et al., 2022), tailoring it for the psychiatric setting. We evaluated the performance of three LLMs (GPT-3.5, Llama-3.1, and Mistral) on the same sentiment analysis task via prompt-based approaches and in-context learning (ICL). In doing so, we seek to address the following questions:

1. Do physicians and non-physicians perceive the sentiment of clinical text describing psychiatric patients differently?
2. Which perspectives do out-of-the-box LLMs align better with in the psychiatric domain: physicians or non-physicians?
3. Can ICL approaches inject subjectivity into LLM behavior to mimic the point of view of physicians versus non-physicians?

To summarize, this work demonstrates that the successful deployment of LLM methods in healthcare depends on addressing the differing points of view between physicians and non-physicians and the subjectivity of sentiment labels that reflect real-word scenarios in psychiatry. In doing so, we learn how to optimize LLMs as-is for tasks using psychiatric text and move towards understanding how bias is represented in the clinical text input into LLMs. This work not only aims at informing the clinician note writing process, but also at providing insight into bias quantification, identification, and removal. There are considerable ethical dilemmas in this project, many of which stem from concerns about using LLMs in a high risk domain like psychiatry, which we discuss in section 7.

2. Related Work

Current definitions of harmful language include words from discriminatory and stigmatizing lexicons (Himmelstein et al., 2022; P Goddu et al., 2018; Park et al., 2021; Sun et al., 2022). Attempts to assess harmful language in health care settings have recently increased, leveraging such lexicons and benefiting from the advances in NLP systems (Bilotta et al., 2024; Boley et al., 2024; Kelly et al., 2023). These approaches and lexicons rely on a consensus perspective, which has yet to be dissected into how physicians and non-physicians perceive the same words and their usage. One explanation could be that taking the patient's perspective into account is a newer issue. In 2020, the 21st Century Cures Act went into effect in the United States, granting patients the right to immediately access their electronic health record data, including clinicians' notes, during a clinical encounter (Rodriguez et al., 2020). Patient-facing interfaces, such as MyChart, furthermore make clinical notes easily accessible for reading—some even alert the user that a new note is ready to view. The perspectives of patients must therefore be considered when flagging potentially harmful language, if we aim to holistically address its downstream effects.

Sentiment analysis in NLP is the process of determining whether the tone conveyed by the written text is positive, neutral, or negative, and it has been leveraged to identify harmful language such as hate speech (Subramanian et al., 2023). As such, sentiment analysis can be used as a proxy in the clinical domain to identify discriminatory and stigmatizing language use. Yet, existing sentiment analysis methods have not been optimized for use in the clinical domain, particularly in psychiatry. Recent approaches to implementing sentiment analysis rely on sentiment lexicons of negative and positive words or sentences to label publicly available clinical note data (Bittar et al., 2021; McCoy et al., 2015). These existing methods exhibit low validity and high variability and do not generalize to psychiatry (Denecke and Reichenpfader, 2023; Holderness et al., 2019; Weissman et al., 2019). Furthermore, there is a lack of psychiatric clinical note datasets with sentiment annotation, and no existing annotations report including the perspectives of non-physicians. Finally, existing sentiment analysis datasets consist of single consensus labels amongst annotators, wherein there is only one correct sentiment per data point. This limits the ability to explore subjectivity and reflect real world scenarios when more than one label is correct to describe a patient (Basile, 2020).

3. Methods

3.1. Data

3.1.1. Patient Descriptor Lexicon

A lexicon of psychiatric patient descriptors was created by the two authors. One is a psychiatrist with experience in writing clinical notes and the second is a computational scientist with experience in fairness, justice, and ethics in machine learning. A set of words (N=54) was initially obtained combining those listed in a highly cited paper about patient descriptors (Sun et al., 2022) and a list of "never words" (i.e., discriminatory words or phrases that should never be used to describe patients in Emergency Medicine). The "never words" were identified by the Institute for Health Equity Research at the affiliated medical center of our authors. Words were then filtered to include:

Word	Sentence
Adamant	Patient very adamantly against hospitalization; states she is not suicidal and needs to go home to care for her dog.
Adherent	States he is adherent with his outpatient XXX and HIV meds.
Agitated	She has been awake now for several hours and has become active and hyperverbal, not overly agitated but mood labile.
Aggressive	Pt remains aggressive and very threatening upon arrival, tried to hit security officer with his head while still in handcuff.
Angry	He remained irritable with an angry edge but was able to respond to redirections.
Compliant	She states that pt has been compliant with meds.
Cooperative	Calm and cooperative, agrees with plan to stay overnight and went back to sleep.
Malingering	Pt had been evaluated in XXX ED earlier today, and was felt to be malingering re: XXX complaints leading her to request evaluation/admission.
Non-Adherent	Given recent non-adherence, will restart pt on VPA 500mg BID, fluphenazine 5mg PO BID and benztropine 1mg BID.
Not Agitated	On reassessment this AM, pt remained calm, not agitated, and again is without SI/HI/AH/VH/PI or delusional content.
Non-Compliant	His worsening psychotic symptom secondary to medication noncompliance and substance use (utox (+) cocaine/cannabis), will admit for safety.
Pleasant	MSE: pleasant, cooperative, euthymic, speech wnl, affect full and appropriate to content.
Uncooperative	He is not cooperative with questions and starts screaming incoherently "which is it, which is it, which is it" unable to re-direct after this.

Table 1: Lexicon and sample sentences.

- Words used exclusively to describe patients (e.g., *"claims"* was not included because it is often used for insurance information).
- Words often used to describe patient behavior in psychiatric settings (e.g., *"agitated"* is more often used to describe patients in psychiatric settings versus other medical settings).
- Words with subjective meaning, as determined by the aforementioned members of the research team (e.g., *"compliant"* carries different meaning depending on someone's clinical experience such that both authors could not agree on the meaning of the word but could agree it would be important to include for downstream assessment of differing perspectives).

After filtering, 13 words, including negations, were retained (see Table 1).

3.1.2. Sentence Selection

Three sentences matching each word from the lexicon of descriptors were extracted from an initial query of 1,000 random clinical notes from the the medical system's electronic health record database. Notes were limited to Progress Reports from clinical encounters where the billing diagnosis was a psychiatric diagnosis. Psychiatric diagnoses were defined as F01-99 International Classification of Disease, Tenth Revision (ICD-10) codes. A total of 39 sentences of length $>$ 30 characters was obtained and protected health information (PHI) was manually masked (see examples in Table 1 in 1).

3.1.3. Physician and Non-Physician Annotations

Members of our research institute, not involved in the study, labeled the sentences identified in Section 3.1.2. Physician raters (N=10) have medical degrees and extensive experience writing and reading clinical notes. Non-physician raters (N=10) have no clinical experience, nor medical degree, received no training on clinical note writing, and had no experience reading clinical notes except those from their own health record. Annotators were asked to label the 39 sentences as neutral, negative, or positive. Physicians were given the direction *"If you're the physician who wrote this sentence: what is your attitude towards the patient?"*. Whereas non-physicians were given the direction: *"If you're the patient: how do you feel reading this description of you?"* The goal of these prompts were to address the downstream harms of negative patient descriptions when they are read by patients, and when they are read by physicians. However, it should be noted that doing so involves the anthropomorphism of LLM behavior, which has its own limitations discussed in section 6. This dataset can be accessed on GitHub here.

3.1.4. Unified Labels

For each sentence, separately for physicians and non-physicians, the sentiment with the most agreement among the annotators was assigned as the unified label. In one instance, the 10 non-physicians were evenly split between neutral and negative labels for the sentence: *"However this morning he is adamant that he wants to go to XXX, does not want to go to program and does not want to go home."* The lead author decided to label the sentence as neutral.

3.1.5. Sentiment Analysis Datasets

To investigate how the language models adapt to the physician/non-physician point of views at different levels of subjectivity, five subsets of labeled sentences were created and split into train and validation with 30/70% ratio for the prompt-based approach:

1. Baseline: All sentences (N=39).
2. 70% agreement: Sentences with $\geq$ 70% agreement within physician/non-physician labels (N=33).
3. 80% agreement: Sentences with $\geq$ 80% agreement within physician/non-physician labels (N=23).
4. 90% agreement: Sentences with $\geq$ 90% agreement within physician/non-physician labels (N=14).
5. No agreement: Sentences with no agreement between physician/non-physician labels (i.e., the physician label is negative, and the non-physician label is positive; N=8).

It is worth noting, that the subset of sentences with 70% agreement refers to sentences for which at least 7 out of 10 physicians or non-physicians assigned the same label. In other words, a sentence with 70% agreement from non-physicians but 60% agreement from physicians would still be included in the 70% agreement dataset.

The test set for this task consisted of 15 newly selected sentences from the dataset of 1,000 random psychiatric clinical notes that were not already used in this experiment. Each sentence contained at least one word from the patient descriptor lexicon.

3.2. LLMs with Prompt-Based Approach

We utilized three LLMs: GPT-3.5-turbo, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.2. These models were chosen based on their state-of-the-art performance as of Spring 2024 when this project was completed. The temperature of the models was set to 0 or 0.001 (i.e., the minimum for each model) and seed to 42. First, the LLMs labeled the sentiment of sentences in the validation datasets using both a zero-shot and ICL approach. We used the validation set to determine which combination of ICL sentences in the training set best improved model performance, sampling sentences from the training set for the physician and non-physician labels of each Baseline, 70%, 80%, and 90% agreement dataset. The test set evaluated each LLM with a zero-shot approach and the best performing ICL prompts during validation.

3.2.1. Prompt Engineering

Our prompt-based approach utilized two prompts, one for the physician task and one for the non-physician task, for each model. We aimed to replicate the prompts that the annotators were given when completing their annotations as described in Section 3.1.3.

The prompt for the physician task: *"You are a doctor familiar with medical jargon that writes many clinical notes about patients. Your task is to analyze the sentiment of a series of sentences you wrote about patients. For each sentence, what is your attitude towards the patient you wrote about?"*

The prompt for the non-physician task: *"Doctors write lots of clinical notes about you. Your task is to analyze the sentiment of a series of sentences your doctor wrote about you. For each sentence, how do you feel reading this description of you?"*

In each prompt, we utilized the subsets of training sentences as contextual examples for the model. To determine which training sentences led to the best model's performance, we engineered our prompt using different combinations of negative, neutral, and positive sentences. For example, the training sentences with 80% agreement contained 2 negative, 3 neutral, and 2 positive sentences, so we created a matrix to loop through every combination: $[negative : [0, 1, 2]; neutral : [0, 1, 2, 3]; positive : [0, 1, 2]]$.

Due to the small number of examples in the subset of sentences with no agreement between physician and non-physician labels, we only applied a zero-shot prompt-based approach to that subset. The composition of the best ICL sentences from the training datasets can be found in Figure 4. The ICL sentences that led to the best model performance on the validation task of Baseline and 70/80/90% agreement subsets were saved and used with the corresponding prompt for evaluating the model on the same test set. Test results were manually validated by the psychiatrist and computational scientist that created the psychiatric patient descriptor lexicon.

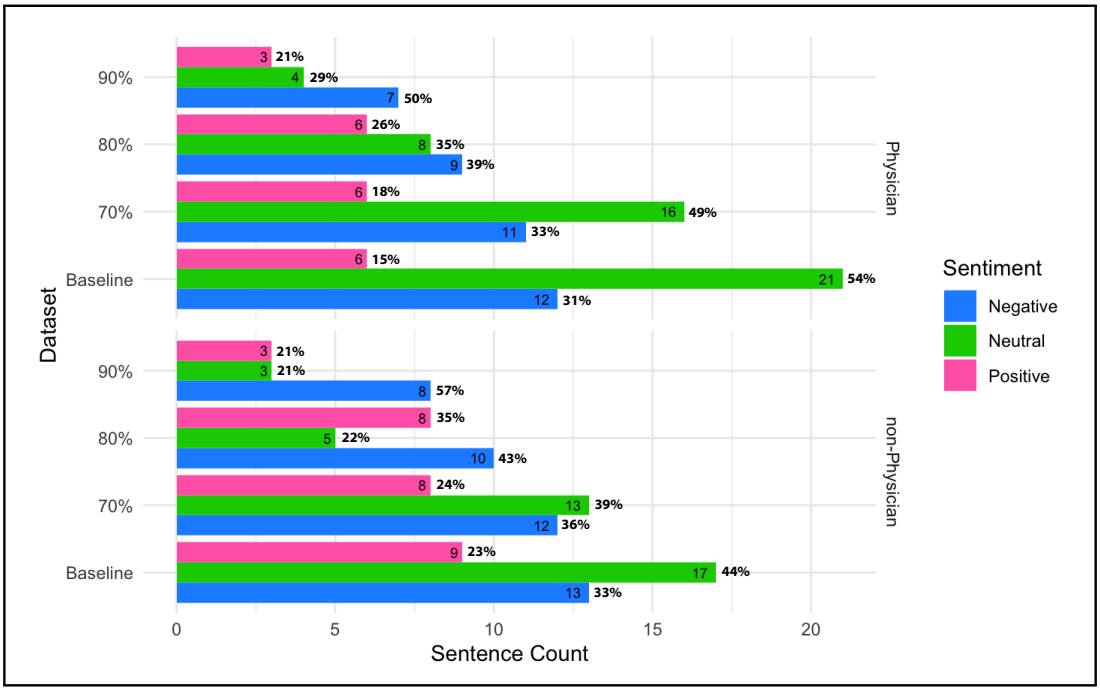

Figure 1: Counts and percents of negative, neutral, and positive sentences within each dataset for physician and non-physician labels. Datasets: Baseline, 70% Agreement, 80% Agreement, and 90% Agreement.

4. Results

We investigated whether physician's and non-physician's points of view could be captured by LLMs via sentiment analysis by asking physicians (N=10) and non-physicians (N=10) to label 39 rule-based extracted real-world sentences describing psychiatric patients. We also explored how subjectivity impacts model performance on the sentiment analysis task by evaluating the models on four datasets with different agreement levels, thus leveraging the datasets described in Section 3.1.5. We also evaluated rater agreement between and within non-physician and physician groups using statistical and non-statistical methods which we first discuss below.

4.1. Rater Agreement

Our results demonstrate a nuanced subjectivity in how physicians and non-physicians perceive the sentiment of the same sentence such that non-physician labels were more polarized than physicians (Figure 1). Namely, physicians labeled more sentences as neutral compared to non-physicians (54% vs 44% in the Baseline dataset) and had higher average agreement on neutral labels than non-physicians (Table 2 in 4.1). Non-physicians demonstrated higher agreement on negative labels than physicians and lower agreement on sentences containing clinical words, such as *"compliant"* (Table 3). However, average percent agreement was lowest on neutral sentences for both physician and non-physician raters (66%). This resulted in a drop of neutral sentences included in the 80% agreement datasets, compared to the baseline and the 70% agreement datasets (Figure 1).

When examining label agreement between all 20 raters, we found fair agreement ($k = 0.323, p < 0.001$) using Fleiss classification (Fleiss, 1971). A fair level of agreement was also found within physician raters ($k = 0.321, p < 0.001$) and non-physician raters ($k = 0.345, p < 0.001$). Good agreement was found when comparing the unified physician and non-physician labels ($k = 0.673, p < 0.001$) with Cohen's Kappa coefficient (McHugh, 2012), however Pearson's Chi-squared test suggests physicians and non-physicians label sentiment differently ($\chi^2 = 8.9, p < 0.05$).

	Physician Raters	non-Physician Raters
Negative	75%	79%
Neutral	69%	63%
Positive	80%	77%

Table 2: Average percent agreement of unified label per sentence for sentiment label and rater group.

4.2. LLM Performance

Mistral and Llama-3.1 generally outperformed GPT-3.5 on both the physician and non-physician tasks with a zero-shot or ICL approach during validation (Figure 2). However, during evaluation on the test

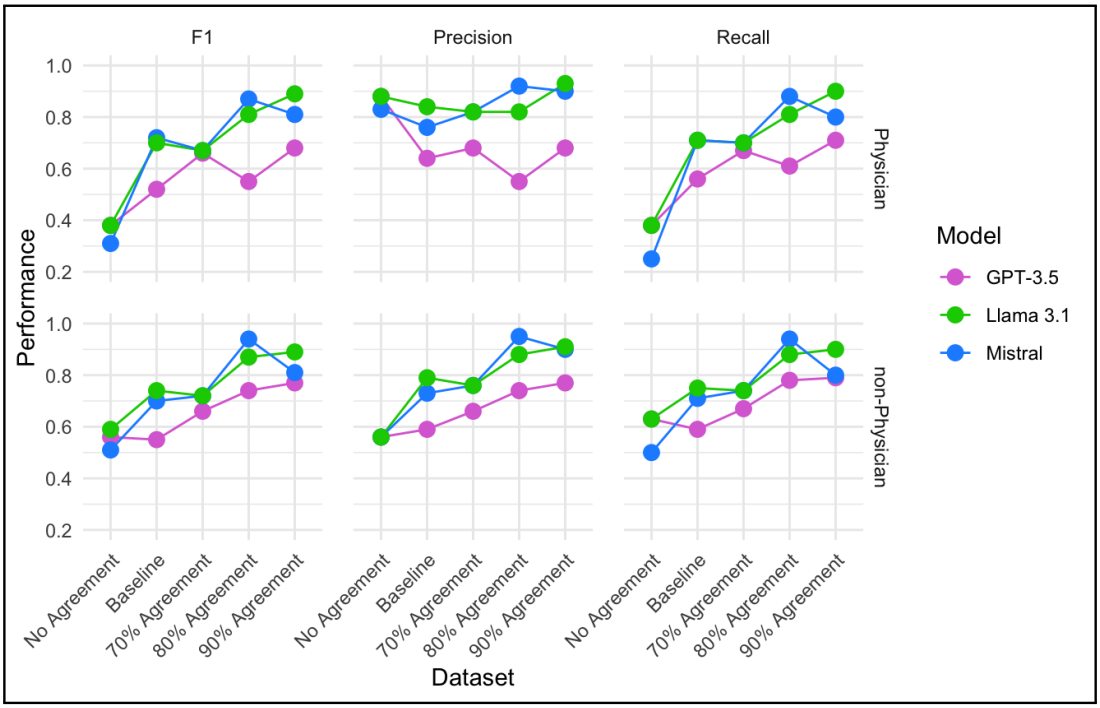

Figure 2: The macro F1 score, precision, and recall for each LLM on the validation data for each dataset (No Agreement/Baseline/60/70/80/90%) using a zero-shot approach.

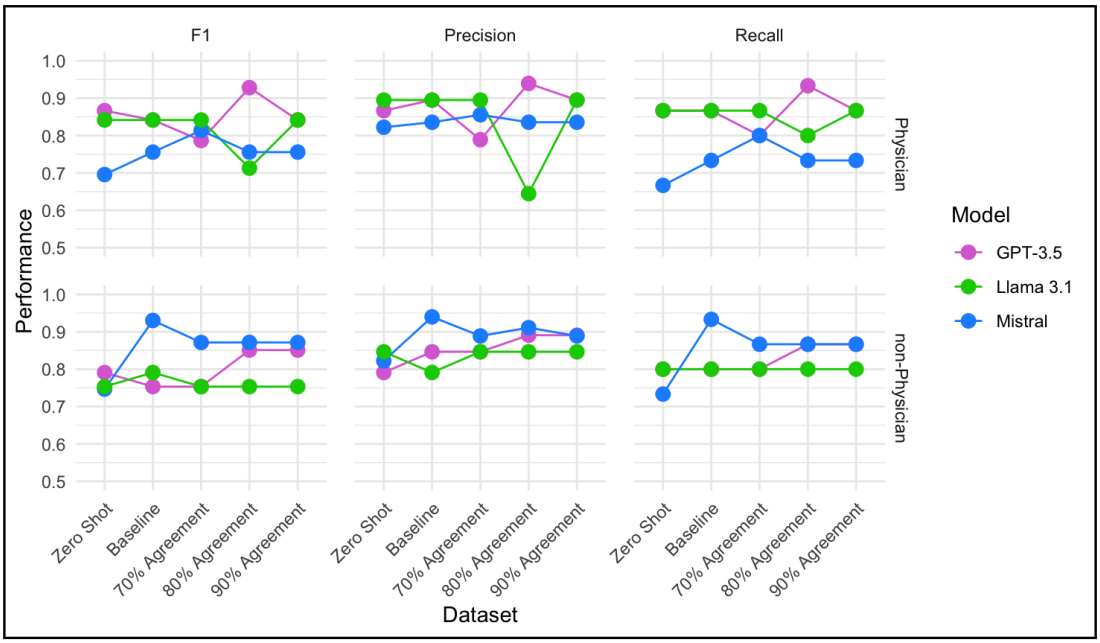

Figure 3: The macro F1 score, precision, and recall for each LLM on the test data using a zero-shot and ICL approach. The ICL approach used the best prompt with ICL sentences from each corresponding training dataset (Baseline/60/70/80/90%).

set, GPT-3.5 performed best on the physician task ($F1 = 0.93$) when using ICL with sentences with a high level of agreement ($\geq$ 80%). Mistral performed best on the non-physician task ($F1 = 0.93$) when using ICL with sentences with the lowest level of agreement (the Baseline dataset; Figure 3). This suggests that ICL approaches may be able to inject some subjectivity into LLM behavior to improve alignment towards physician or non-physician perspectives.

The results of the zero-shot approach demonstrate that models perform best on sentences

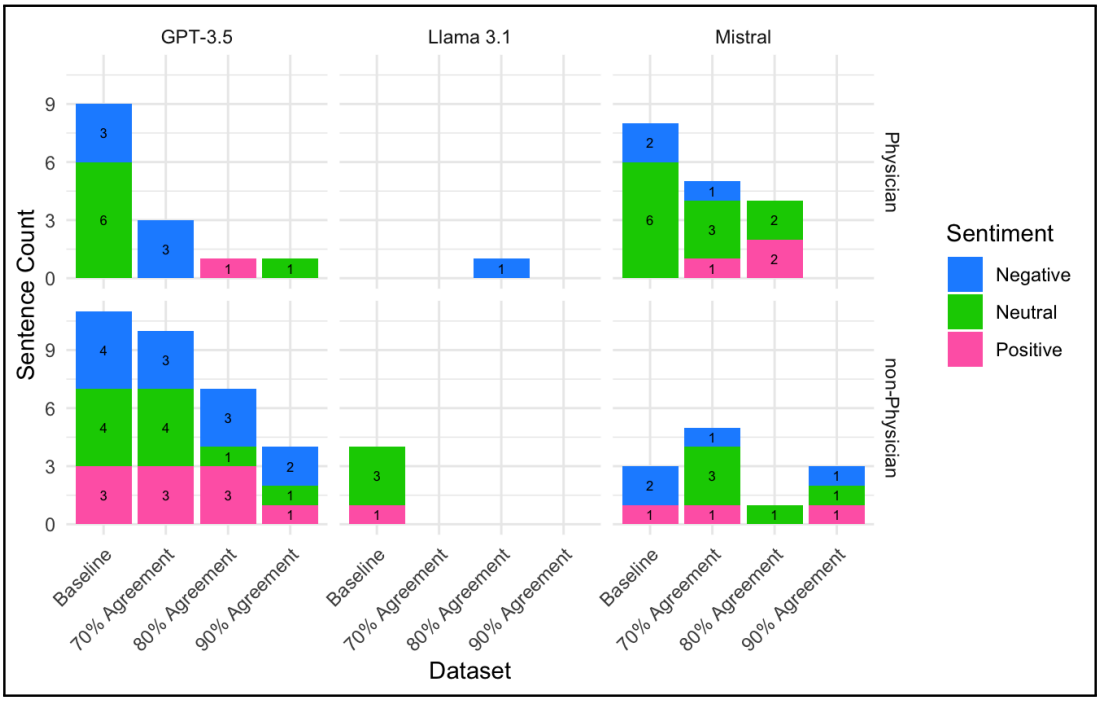

Figure 4: Sentence count per sentiment label, model, and dataset. This figure describes the number of sentences across each sentiment label included in the prompt for each dataset and LLM that led to the best performance on the validation data for each dataset.

Word	Physician (%)	non-Physician (%)	Δ (%)
adamant	70	57	13
adherent	63	63	0
agitated	77	80	3
not agitated	70	66	4
aggression	83	73	10
angry	80	70	10
compliant	70	53	17
cooperative	70	77	7
malingering	77	70	7
non-adherent	63	77	14
non-compliant	67	77	10
pleasant	87	80	7
uncooperative	63	83	20

Table 3: Percentage agreement by word and rater type. Rows highlighted in yellow indicate a disagreement of $>$ percentage points.

with high level of agreement, as seen in Figure 2. When evaluating the models on the sentences with no agreement between physicians and non-physicians, they all perform best on the non-physician labels. Interestingly, all three models demonstrate high precision (> 0.80) when evaluated on the physician labeled no-agreement dataset with a zero-shot approach, but poor recall (< 0.40). This implies that the sentences that only physicians perceive as negative or positive are often mislabeled as neutral.

Utilizing ICL improved the performance of Mistral and GPT-3.5, but not Llama-3.1 (Figure 3). As seen in Figure 4, Mistral and GPT-3.5 perform better when using more ICL sentences in their prompt such that both models typically used at least one ICL sentence. In contrast, Llama-3.1 doesn't perform better during the validation stage with ICL sentences apart from two scenarios. Therefore, when evaluated on the test set, Llama-3.1's performance plateaus and remains the same as zero-shot.

From the MD perspective, a Friedman test indicated a significant effect of model on F1 score ($\chi^2 = 6.4, p < 0.05$), likely driven by the significant effect of model on Recall ($\chi^2 = 8.0, p < 0.05$). However, no significant differences were observed between models for F1 from the non-MD perspective.

5. Discussion

In these experiments, we demonstrated that:

1. There are nuances between how physicians and non-physicians perceive the sentiment of sentences describing psychiatric patients in clinical notes.

2. LLMs are sensitive to this subjectivity, thus sometimes showing differential performance on physician and non-physician sentiment analysis tasks.

3. The use of ICL sentences can improve model alignment on sentiment analysis tasks with subjective points of view, but does not benefit all models.

Our results suggest that there are differences in how physicians and non-physicians label the sentiment of sentences that go undetected by agreement statistics. Namely, physicians perceive more sentences as neutral than non-physicians, whose labels have more variation and more likely to be negative or positive (Figure 1). These differences could be explained by the training physicians receive, where they learn established clinical language and the medicolegal approach to documenting in the clinical notes, all influencing their intuitive clinical interpretation of the language.

This work's findings suggest that neutrality is the hardest sentiment for humans to agree on, and consequently for language models to label. We found that increased agreement is associated with fewer neutral sentences within our annotated datasets. Moreover, LLMs struggled with labeling neutrality. GPT-3.5, and Mistral were "too neutral", i.e., negative and positive sentences were often mislabeled as neutral. Conversely, Llama-3.1 was more polarized, wherein neutral sentences were more likely to be mislabeled as negative or positive. These findings underline the importance of carefully investigating the sentiment composition in annotated text and determining whether to prioritize reducing false positives or false negatives in LLM outputs, particularly when deploying models for harmful language detection.

Lastly, we see that subjectivity might be possible to inject into off-the-shelf LLMs within sentiment analysis tasks. For instance, Mistral shows performs best on the non-physician labels in test set if using ICL sentences with the lowest level of agreement (i.e., the most subjective) from the Baseline dataset. Similarly, GPT-3.5 performs best on the physician labels in the test set if using ICL sentences with higher agreement (i.e., 80%).

In conclusion, this work advocates for sentiment analysis methods in psychiatry to reflect real-world scenarios. This can be achieved by recognizing the point of view and subjectivity inherent to describing and perceiving the descriptions of psychiatric patients. Our work is the first to demonstrate how to leverage LLMs for such tasks and our contributions include:

- The generation of a manually curated lexicon of words and annotated dataset of sentences describing patients in psychiatry.
- The investigation of the differences between physicians' and non-physicians' sentiment towards sentences describing psychiatric patients.
- The implementation of an NLP framework to assess the alignment of LLMs to physicians and non-physicians point of views.

These contributions bring us closer to utilizing methods that take into account the downstream harms of clinical language, and understanding how bias or harmful patient descriptions are represented in clinical text.

6. Limitations

One crucial limitation of this work is the size of our annotated dataset. When we started this project, we were motivated to see if there are differences between how doctors and patients perceive language in psychiatric clinical notes. The logistics of doing such a project meant to find physician annotators that would donate their time, therefore we prioritized creating a small dataset to make it easier to obtain more annotators. As it was, we noticed large rater disagreement levels when we started with three annotators per group. Therefore, although we have a small dataset, we have very robust annotations. Furthermore, we argue that exploring the disagreement within annotations is perhaps the most rewarding aspect of such work, in opposition to the typical standards/motivations of defining a 'gold label'. We hope that by making this dataset publicly available (GitHub link here), others can build upon our work meanwhile appreciate our first attempt.

Due to a lack of transparency about pretraining data, it is difficult to speculate why certain LLMs align better to physician or non-physician points of view. Since Mistral performed best on the non-physician task, one weak explanation could be that Mistral's pretraining corpus does not include clinical text. Since GPT-3.5 and Llama-3.1 perform better on the physician task, someone could hypothesize that their pretraining data includes clinical text.

Our prompt-based approach was reliant on asking LLMs *"how do you feel?"* This can be referred to as an anthropomorphism of LLM behavior. Anthropomorphism has many functions across NLP, human-AI interaction, and even robotics (Xiao et al., 2025). In this project, the sentiment analysis task used human-like perceptual cues dependent on the assumption that LLMs can mimic human behavior (i.e., *"feeling"*). Recent work has shown that there are significant benefits to anthropomorphic cues, especially when using LLMs in a mental health context (Xiao et al., 2025). However, LLMs demonstrate varied psychological capabilities, with distinct differences between GPT, Llama, and Mistral models' emotional abilities (Dong et al., 2025).

We hope that future work will consider cultural differences in the perspectives of sentiment towards psychiatric patients, as well as implement community-based approaches for recruiting active psychiatric patients to aid in defining negative patient descriptions. Throughout this project, we at-

tempted to involve as many perspectives as possible when finding sentence annotators. However, due to privacy reasons, we could not collect personal data from raters, including their mental health or sociodemographic factors. It is possible that some of the non-physician raters are living with psychiatric diagnoses, but we are unable to make that assumption or claim that our work describes the point of view of active psychiatric patients. We consider this to be a large limitation, as psychiatric patients would be reading their own notes containing descriptions using psychiatric lexicon.

7. Ethical Considerations

Work such as ours has considerable ethical dilemmas. Namely, should LLMs be used in high risk domains like psychiatry? Our intuition is "No!" As discussed earlier, LLMs clearly perpetuate societal biases and introduce additional risks to patients. In reality, LLMs are already being used in clinical settings and, outside of the medical system, people leverage them for social support and pseudotherapy everyday. Given that LLMs are already being utilized in these scenarios, it is our duty as scientists to evaluate models, develop recommendations on how to safely use them, and push for more transparency in how these models are developed. This project aims to take a step in that direction, by evaluating which consensus opinions LLMs reflect in the psychiatric domain.

The long-term goal of our work is to explore approaches to debiasing language models and clinical note datasets that consider the point of view of the physician and non-physician. Once biased and harmful text is detected in clinical notes, the next task is deciding what to do about it. One solution is to “neutralize” language use, wherein harmful words are either removed or replaced with words conveying a neutral point of view (Pryzant et al.). However, not all sentences with negative sentiment are harmful to the patient, and some are necessary for describing the patient’s experience and clinical needs. Furthermore, no consensus exists on whose point of view, the patient or physician, should be used in such practices. As such, we must work towards debiasing approaches that consider the context of the bias due to the reality that removing or replacing biased language is not always the most ethical approach (Holm et al., 2023).

8. Data Availability

The dataset referenced in this paper with annotations is available on GitHub at this address: github.com/valentinealissa/sentiment_POV/blob-/main/sentiment_sentences_annotated.csv

9. Acknowledgements

We thank Ipek Ensari PhD, Ashwin Sawant MD PhD, and Matthew O’Connell PhD for their invaluable contributions to this research as members of the first author’s advisory committee. We also thank the annotators who made this work possible. The physician annotators include Alexander Charney, Ali Soroush, Ashwin Sawant, Caroline Massarelli, Donald Apakama, Emma Holmes, Ethan Abbott, Girish Nadkarni, Jihan Ryu, and Lili Chan. The non-physician annotators include Alisha Aristel, Brian Fennessy, Darielle Lewis-Sanders, Eric Vornholt, Eugenia Alessandra Enrica Alleva Bonomi, Maria Koromina, Renata Gonzalez Chong, Rozalyn Wood, Simon Lee, and Tom Kaszemacher. Special appreciation goes to the patients of Mount Sinai – may their data always be used for their benefit.

This work was supported in part through the computational and data resources and staff expertise provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai and supported by the Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences. Research reported in this publication was also supported by the Office of Research Infrastructure of the National Institutes of Health under award number S10OD026880 and S10OD030463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We also acknowledge the funding and resources provided by the Mount Sinai Hospital System and Institute for Personalized Medicine. This study was funded by the US National Institutes of Health grants R01MH121923, R01-PAR-18-896, and IMPACT-MH U01. Work on this project was partially funded by the Lundbeck foundation (Lundbeck Collaborative TRUSTMIND, grant number R453-2024-356). The funding agencies had no influence on the writing, submission, or publication of this manuscript.

10. Bibliographical References

Valerio Basile. 2020. It’s the end of the gold standard as we know it. on the impact of preaggregation on the evaluation of highly subjective tasks. In *CEUR workshop proceedings*, volume 2776, pages 31–40. CEUR-WS.

Isabel Bilotta, Scott Tonidandel, Winston R Liaw, Eden King, Diana N Carvajal, Ayana Taylor, Julie Thamby, Yang Xiang, Cui Tao, and Michael Hansen. 2024. Examining linguistic differences in electronic health records for diverse patients

with diabetes: Natural language processing analysis. *JMIR Medical Informatics*, 12(1):e50428.

André Bittar, Sumithra Velupillai, Angus Roberts, and Rina Dutta. 2021. Using general-purpose sentiment lexicons for suicide risk assessment in electronic health records: Corpus-based analysis. *JMIR medical informatics*, 9(4):e22397.

Sean Boley, Abbey Sidebottom, Marc Vacquier, David Watson, Bailey Van Eyll, Sara Friedman, and Scott Friedman. 2024. Racial differences in stigmatizing and positive language in emergency medicine notes. *Journal of Racial and Ethnic Health Disparities*, pages 1–11.

Kerstin Denecke and Daniel Reichenpfader. 2023. Sentiment analysis of clinical narratives: a scoping review. *Journal of Biomedical Informatics*, 140:104336.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, Shengmin Xu, Xinyi Huang, and Xinlei He. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications.

Leonor Fernández, Alan Fossa, Zhiyong Dong, Tom Delbanco, Joann Elmore, Patricia Fitzgerald, Kendall Harcourt, Jocelyn Perez, Jan Walker, and Catherine DesRoches. 2021. Words matter: what do patients find judgmental or offensive in outpatient notes? *Journal of general internal medicine*, pages 1–8.

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

Tabor E Flickinger, Somnath Saha, Richard D Moore, and Mary C Beach. 2013. Higher quality communication and relationships are associated with improved patient engagement in hiv care. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 63(3):362–366.

Christoph Flückiger, Aaron C Del Re, Bruce E Wampold, and Adam O Horvath. 2018. The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4):316.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Gracie Himmelstein, David Bates, and Li Zhou. 2022. Examination of stigmatizing language in the electronic health record. *JAMA Network Open*, 5(1):e2144967–e2144967.

Eben Holderness, Philip Cawkwell, Kirsten Bolton, James Pustejovsky, and Mei-Hua Hall. 2019. Distinguishing clinical sentiment: The importance of domain adaptation in psychiatric patient health records. *arXiv preprint arXiv:1904.03225*.

Sune Holm, Eike Petersen, Melanie Ganz, and Aasa Feragen. 2023. Bias in context: What to do when complete bias removal is not an option. *Proceedings of the National Academy of Sciences*, 120(23):e2304710120.

Adam O Horvath and B Dianne Symonds. 1991. Relation between working alliance and outcome in psychotherapy: A meta-analysis. *Journal of counseling psychology*, 38(2):139.

Marc S Karver, Alessandro S De Nadai, Maureen Monahan, and Stephen R Shirk. 2018. Meta-analysis of the prospective relation between alliance and outcome in child and adolescent psychotherapy. *Psychotherapy*, 55(4):341.

Patrick JA Kelly, Andrew M Snyder, Madina Agénor, Cassandra R Navalta, Chelsea Misquith, Josiah D Rich, and Jaclyn MW Hughto. 2023. A scoping review of methodological approaches to detect bias in the electronic health record. *Stigma and Health*.

Thomas H McCoy, Victor M Castro, Andrew Cagan, Ashlee M Roberson, Isaac S Kohane, and Roy H Perlis. 2015. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. *PloS one*, 10(8):e0136341.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *BIOCHEMIA MEDICA*, 22(3):276–282.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, and Benjamin S Glicksberg. 2025. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9.

Anna P Goddu, Katie J O'Conor, Sophie Lanzkron, Mustapha O Saheed, Somnath Saha, Monica E Peek, Carlton Haywood, and Mary Catherine

Beach. 2018. Do words matter? stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33:685–691.

Ridam Pal, Hardik Garg, Shashwat Patel, and Tavpritesh Sethi. 2023. Bias amplification in intersectional subpopulations for clinical phenotyping by large language models. *medRxiv*, page 2023.03. 22.23287585.

Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach. 2021. Physician use of stigmatizing language in patient medical records. *JAMA Network Open*, 4(7):e2117052–e2117052.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Jorge A Rodriguez, Cheryl R Clark, and David W Bates. 2020. Digital health equity as a necessity in the 21st century cures act era. *Jama*, 323(23):2381–2382.

Jennifer M Silva, T Elizabeth Durden, and Annemarie Hirsch. 2023. Erasing inequality: Examining discrepancies between electronic health records and patient narratives to uncover perceived stigma and dismissal in clinical encounters. *Social Science Medicine*, 323:115837.

Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathiskumar, G Deepalakshmi, Jaehyuk Cho, and G Manikandan. 2023. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. *Alexandria Engineering Journal*, 80:110–121.

M. Sun, T. Oliwa, M. E. Peek, and E. L. Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record. *Health Aff (Millwood)*, 41(2):203–211. Sun, Michael Oliwa, Tomasz Peek, Monica E Tung, Elizabeth L eng K23 HL145090/HL/NHLBI NIH HHS/ L30 HL148782/HL/NHLBI NIH HHS/ P30 DK092949/DK/NIDDK NIH HHS/ UL1 TR000430/TR/NCATS NIH HHS/ Research Support, N.I.H., Extramural 2022/01/20 Health Aff (Millwood). 2022 Feb;41(2):203-211. doi: 10.1377/hlthaff.2021.01423. Epub 2022 Jan 19.

Gary E Weissman, Lyle H Ungar, Michael O Harhay, Katherine R Courtright, and Scott D Halpern. 2019. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. *Journal of biomedical informatics*, 89:114–121.

Yunze Xiao, Lynnette Hui Xian Ng, Jiarui Liu, and Mona T. Diab. 2025. Humanizing machines: Rethinking LLM anthropomorphism through a multi-level framework of design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3331–3350, Suzhou, China. Association for Computational Linguistics.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, and Raja-Elie E Abdulnour. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.