Investigating Stigmatizing Language in Clinical Documentation with Open-Source Large Language Models

Rajashree Dahal,[1] **Pardis Hossein Pour,**[2] **Pranathi Kamisetty,**[1]
Satwik Pamulaparthy,[1] **Saeid Tizpaz-Niari,**[1] and **Natalie Parde**[1]
[1]Department of Computer Science, University of Illinois Chicago, IL, USA
[2]College of Applied Health Sciences, University of Illinois Chicago, IL, USA
{rdaha1, phoss, pkami, vpamu, saeid, parde}@uic.edu

Abstract

Clinical documentation is essential for patient care, billing, and medical research, but it is subject to entrenched bias. While manual chart reviews can identify such bias, they are labor-intensive and expert-dependent. We introduce and evaluate StigMAD, a Multi-Agent Debate framework leveraging open-source Large Language Models (LLMs) to detect stigmatizing language in clinical documentation. We investigate reasoning (multi-agent debate), self-reflection, and self-consistency within this framework. Experiments on clinical notes and patient summaries demonstrate that our framework provides significant advantages over rule-based and supervised baselines. A domain-specific LLM (MedGemma) achieved its highest performance using the StigMAD reasoning framework, while a general-purpose LLM (Llama) showed superior results with the self-consistency framework. These findings suggest that open-source LLMs, steered by structured prompting and reflective reasoning, can effectively support the auditing of stigmatizing language, marking a critical step toward more equitable clinical NLP systems.

1 Introduction

A wide variety of clinical documentation, including history and physical examination notes, daily progress notes, special consults, surgery or procedure reports, discharge summaries, and nurses' notes, can contribute to narrating patient care. Holistically, these narratives provide continuity of care across diverse medical teams and support activities such as billing and medical research (Rizvi et al., 2016). They may be supplemented by other free-text healthcare documentation such as interview transcripts, patient-provider communication logs, case reports, and survey responses, especially in research or mental health settings (Park et al., 2019). However, all of this documentation may vary in tone, structure, and completeness, and bias

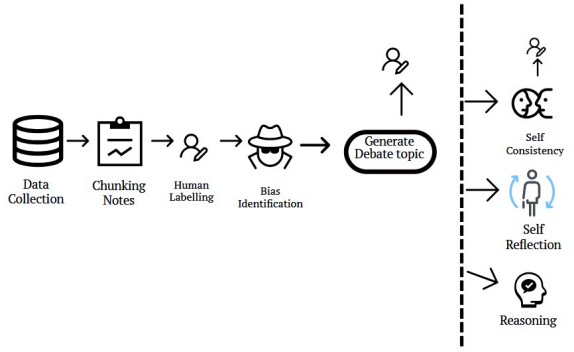


Figure 1: Methodological framework for bias detection.

can easily enter the process. For example, Black patients are more than twice as likely as white patients to have negative descriptors applied in their notes, which can influence diagnosis, treatment, and referral decisions (Sun et al., 2022). Stigmatizing language and assumptions in documentation can affect downstream AI and human decisions (Liu et al., 2023) and can produce disparities in clinical, policy, and organizational outcomes (Rozier et al., 2022). Bias can also lead to inappropriate provider decision support, lower reimbursements, and disparities in treatment recommendations, affecting clinical decision-making and contributing to health disparities in patient care (Obermeyer et al., 2019).

Bias in clinical documentation has traditionally been uncovered through manual chart review, qualitative coding, and formal interviews. All of these processes are time-consuming and expert-reliant, making them difficult to scale. To address this, researchers have started to harness artificial intelligence to uncover patterns of bias in clinical documentation (Zhang et al., 2020), for example by detecting stigmatizing language, classifying note sentiment (Valentine et al., 2024), or predicting bias-prone words. However, automatically detecting implicit forms of bias in continually evolving healthcare settings is challenging. Approaches

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 490–501

often rely on fixed labeled data and miss subtle, context-dependent bias (Himmelstein et al., 2022). Frameworks relying on lexicon-based detection of these phenomena may have limited performance without further expansion and more sophisticated classification methods (Walker et al., 2025).

In this work, we introduce the STIGMAD framework, which harnesses self-consistency, self-reflection, and multi-agent debate to identify stigmatizing language using open-source LLMs, hypothesizing that the framework will enable models to reason through ambiguity and more critically analyze output. We illustrate its potential to uncover hidden stigmatizing language patterns and offer an adaptive strategy for identifying and analyzing bias in clinical documentation. Our hope is that the framework can be replicated for use with other forms of bias in the future, and we lay the groundwork for doing so in this work. Our core contributions involve (1) evaluating how effectively and confidently open-source LLMs can detect stigmatizing language in clinical documentation using the STIGMAD framework; (2) deriving how stigmatizing language observed in clinical documentation propagates into clinical summary notes; and (3) ablating individual components of the STIGMAD framework, including self-consistency, self-reasoning, and self-reflection.

2 Literature Review

2.1 Bias in Clinical Documentation

Previously, Chen et al. (2024a) systematically reviewed AI-based bias mitigation and detection strategies for electronic health records (EHRs), identifying implicit bias, selection bias, algorithmic bias, and temporal bias. FitzGerald and Hurst (2017) discussed the presence and impact of implicit biases in healthcare professionals, performing a meta-analysis of findings from 42 studies and establishing that healthcare professionals have extensive implicit biases across numerous areas, including race, ethnicity, and gender. Burton et al. (2022)'s study of 401 patient safety reports found unequal reporting by physicians' gender and race/ethnicity. They also regularly found subthemes such as "inappropriate communication," "verbal abuse," "ignoring or omission of procedure," and "physical intimidation."

Wong and Jackson (2023) discussed positive bias, negative bias, stigmatizing language, implicit bias, explicit bias, gender bias, and racial bias, and highlighted common healthcare domains in which bias may arise. Similarly, Vela et al. highlighted stigma, racial bias, explicit bias, implicit bias, and gender discrimination in their study. We focus on stigmatizing language specifically as our test case, using the subset of MIMIC-IV clinical notes annotated for "Stigmatizing Language" in Harrigian et al. (2023)'s study. Harrigian et al. (2023) identified and classified three types of stigmatizing language in medical records: *credibility and obstinacy*, *compliance*, and *descriptors*.

2.2 LLM-Mediated Bias Analyses

Several studies have examined the potential of LLMs to aid in bias detection (Qiao et al., 2024; Hayes, 2025; Eschrich and Sterman, 2024), by performing zero-shot detection and categorization of bias in emergency department and discharge notes (Apakama et al., 2024). Xavier et al. (2026) evaluated the performance accuracy of LLMs in detecting stigmatizing language across diffent temperature settings and model size. We build on these studies to explore the potential of open-source general-domain (i.e., `Llama3.3-70b-instruct` (Dubey et al., 2024)) and domain-specific (i.e., `medgemma-27b-text-it` (Sellergren et al., 2025)) LLMs to detect stigmatizing language in clinical documentation. In more general domains, counterfactual techniques (e.g., Counter-GAP (Xie et al., 2023)) have emerged as strong performers in this area by enabling precise measurement of model biases through minimally altered, attribute-specific text pairs. Approaches like FACTUAL (Li et al., 2025) further leverage counterfactual augmentation to detect and effectively mitigate bias in LLMs, particularly for stance detection tasks.

With respect to approaches leveraged in our STIGMAD framework, Madaan et al. (2023) introduced *self-refine*, which refines initial LLM output through iterative feedback and refinement. Similarly, Chen et al. (2024b) introduced *self-debugging*, where the model identifies its mistakes in the output it generates. Although several studies have revealed that self-reflection may deteriorate performance (Chen et al., 2024b; Huang et al., 2024; Valmeekam et al., 2023), Xu et al. (2024) found that LLM size can reduce bias in the self-refinement pipeline. We use models with larger parameter sizes to implement *self-reflection* in our study. Finally, *self-consistency* (prompting the LLM with the same prompt multiple times while using a higher temperature value) has produced

promising results in many domains (Ahmed and Devanbu, 2023; Wang et al., 2023).

2.3 Agentic AI for Stigmatizing Language Detection

Acharya et al. (2025) comprehensively studied the use of agentic AI across many fields, including healthcare, motivating our exploration of an agentic AI framework for stigmatizing language detection. While Du et al. (2023) explored LLMs' reasoning ability through a multi-agent debate (MAD) framework, recent studies suggest that although MAD improves reasoning, it can inadvertently reinforce societal biases. For example, Borah and Mihalcea (2024) show that implicit gender biases persist and often escalate during multi-agent LLM interactions. Oh et al. (2026) correspondingly identified bias reinforcement as a key limitation of MAD. These limitations may be attributed to a lack of context and shared reasoning patterns (Borah and Mihalcea, 2024; Oh et al., 2026). Our setup in STIGMAD addresses these limitations fundamentally. By providing each agent with a clearly defined debate paragraph, a quote highlighting the bias, and explicit information about the stigmatizing language bias and its definition, we reduce ambiguity and encourage deliberate, reflective reasoning. We hypothesize that this makes bias reinforcement less likely in our setting, since agents are tasked with detecting and debating bias explicitly.

3 Dataset

3.1 Data Sources

We downloaded the PMC-Patients summary dataset (Zhao et al., 2023) from HuggingFace and also obtained access to the MIMIC-IV Dataset (Johnson et al., 2024, 2023) of real-world clinical documents.[1] Accessing MIMIC-IV notes required credentialed access via PhysioNet (Goldberger et al., 2000) and acceptance of a Data Use Agreement. All uses are within the scope of non-commercial research, in accordance with their respective licenses and access agreements. PMC-Patients summary notes correspond to the 167,000 patient summaries from PMC-Patients including patient visit, medical history, symptoms, treatments, discharge summary, and intervention based on case reports in PubMed Central (PMC). The stigmatization subset of MIMIC-IV is a collection of 4,710 de-identified free-text clinical notes for 4,259 patients and contains 5,043 annotations.

Notes in the MIMIC-IV stigmatization subset had 1802.71±479.59 tokens on average, and summaries in PMC-Patients had 288.58±113.28 tokens. Given the length disparity, we selected seven MIMIC-IV notes and 50 PMC-Patients summaries for further annotation of stigmatizing bias. These document quantities provided us with equal numbers (n=71 each) of context length-limited chunks. The design of our reasoning framework motivated our focus on chunk-level analysis rather than full-file analysis, and the inclusion of both MIMIC-IV and PMC-Patients allowed us to validate the generalizability of our framework.

3.2 Stigmatizing Language

Stigmatizing language has been defined in numerous ways in prior work (Harrigian et al., 2023; Himmelstein et al., 2022; Barcelona et al., 2024). While some studies treat the terms "stigma," "stigmatizing language," and "bias" interchangeably (Barcelona et al., 2024), others emphasize that stigmatizing language can serve as a mechanism for transmitting bias (Himmelstein et al., 2022). Harrigian et al. (2023) positioned the detection of stigmatizing language as a bias-related task in relation to other bias detection tasks in NLP, and we adopted their definition in our work:

> ***Stigmatizing Language:*** *Any language that portrays the patient in a negative or judgmental way—by expressing doubt about their credibility, implying they are difficult, noncompliant, or deceptive, or by using language that casts their behavior, attitude, or emotional state in a negative light. This includes both overt and subtle wording that reflects blame, shame, or disrespect.*[2]

In Prompt A.1, we also included definitions of other bias types from existing literature alongside stigmatizing language to reinforce that in real-world scenarios, multiple biases may coexist.

[1] This study was reviewed by the Institutional Review Board (IRB) at our institution and determined not to involve human subjects as defined by the Department of Health & Human Services (DHHS) and/or U.S. Food and Drug Administration (FDA) regulations.

[2] The final single definition is curated using class-based definitions from three stigma types for Stigmatizing Language.

3.3 Annotation

For MIMIC-IV, we implemented the existing stigmatizing language annotations developed by Harrigian et al. (2023). For PMC-patients summary dataset, we developed annotation guidelines collaboratively with a licensed physical therapist who had extensive experience writing and analyzing clinical notes. Additional guidance was provided by experts specializing in (1) fair and responsible AI, and (2) the intersection of computer science and healthcare. The physical therapist and three computer science graduate students carried out the annotations, following the guidelines provided in §3.2 and Prompt A.1. The graduate students followed the established rules to extend coverage and ensure consistency across the dataset, while the physical therapist oversaw the work to maintain clinical accuracy. All annotators underwent a training session using real clinical notes (not included in the final dataset) to practice annotations together and ensure consistency. They separately annotated every chunk (n=71) and held follow-up meetings to discuss discrepancies and align their understanding. The experts were also involved in these meetings to ensure that clinical and methodological perspectives were incorporated.

We calculated inter-annotator agreement for PMC-Patients using Cohen's cappa (Cohen, 1960) and obtained k=0.70 which is a substantial agreement (Landis and Koch, 1977). A clinician adjudicated the final label for cases of disagreement. Stigmatizing language was found in n=28 chunks and n=19 chunks from the MIMIC-IV and PMC-Patients datasets, respectively.

4 Methodology

4.1 Baseline Framework

Given an input clinical document, our base framework *bias-identifier* (see Prompt A.1) identifies instances of stigmatizing language and uses these as debate topics. We leverage two open-source LLMs, `Llama-3.3-70B-Instruct` and `medgemma-27b-text-it`, to generate the debate topic using a multi-step process. We first ask the LLM to identify stigmatizing language, based on a relevant quote from the input. We set the LLM temperature to 0.1 to encourage more deterministic detection, maintaining logical consistency.

Since medical notes vary in size depending on note-taking context, we chunk the note into 300-token windows, with a sliding window chunk overlap of 10 tokens to preserve coherence across segments.[3] Each chunked paragraph is fed to the LLM to extract the bias (stigmatizing language) itself and quotes. We conducted this process in a zero-shot manner, experimenting with multiple prompting strategies and styles to identify those that performed best.

4.2 Self-Reflection Framework

Our self-reflection framework revisits previously identified quotes and stigmatizing language from clinical documents, along with the definition of stigmatizing language. Its role is to re-evaluate whether the quote within the context of the note truly reflects the specified stigmatizing language. The model provides a boolean output regarding its decision, without providing any explanation. Since justification and reasoning are already addressed elsewhere within the StigMAD framework, the binary reflection provides complementary validation of the correctness of the extracted quote and the predicted stigmatizing language. We used few-shot prompting to implement this framework. The prompt contains few-shot examples of other bias type from existing literature to reinforce real-word scenarios (see Prompt A.2). This framework is conceptually similar to stigma detection guided prompt by (Chen et al., 2025). However, Chen et al. (2025) rely on fixed set of predefined stigmatizing keywords whereas self-reflection uses contextually extracted quotes identified dynamically by bias identifier.

4.3 Self-Consistency Framework

Our self-consistency framework is analogous to *bias-identifier*, but we run this model eight times while increasing the temperature to 0.5 to promote response diversity. The added variability makes it easier to identify consistent outputs through voting across runs. We consider only those outputs for stigmatizing language that occur at least three times out of eight runs for a given input chunk. We keep this threshold low since the *bias-identifier* also extracts quotes, and the exact quote boundary may vary across different runs. The prompt used to implement the self-consistency framework is the same as that of *bias-identifier*.

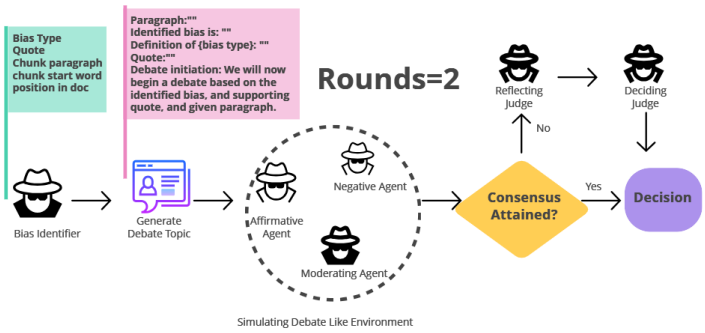


Figure 2: Proposed STIGMAD framework.

4.4 Reasoning Framework

The quotes for the stigmatizing language generated by the *bias-identifier* are converted into a debate topic as shown in Figure 2. The debate topic includes the input clinical document chunk, identified stigmatizing language, definition of stigmatizing language, quote that supports the identified bias, and debate initiation process. Overall, STIGMAD comprises four types of agents:

Affirmative Agent. Can support or deny the debate topic. It justifies its reflection on that topic, arguing affirmatively in brief structured statements (see Prompt A.4).

Negative Agent. Challenges or critiques the affirmative agent's perspective, providing arguments or support for its position (i.e., whether the stigmatizing language exists) (see Prompt A.5).

Moderator Agent. Oversees debate rounds and analyzes responses from affirmative and negative agents, determines whether consensus on bias classification has been reached, and summarizes reasons supporting its final decision. It ensures consistency across debates, reassesses contentious cases, and validates that the consensus aligns with bias definitions. The agent moderates the debate process for two rounds. If consensus is reached at the final round, its decision is the final decision (see Prompt A.3).

Judge Agent. Triggered when no consensus is reached after two moderation rounds. The judge agent takes two roles. The *reflecting judge* (see Prompt A.6) compiles upto five final arguments from affirmative and negative agents's results of the final debate rounds conclude, by listing bias classifications and reasoning without initial evaluation. The *deciding judge* (see Prompt A.7) makes the ultimate decision regarding bias presence based on its initial reflection. This agent is fed with debate topic information, and uses compiled pointers as context to produce a single JSON decision in the form `{"Presence":"True/False", "Reason":""}` which constrains output to one binary label and one justification.

4.5 External Baselines

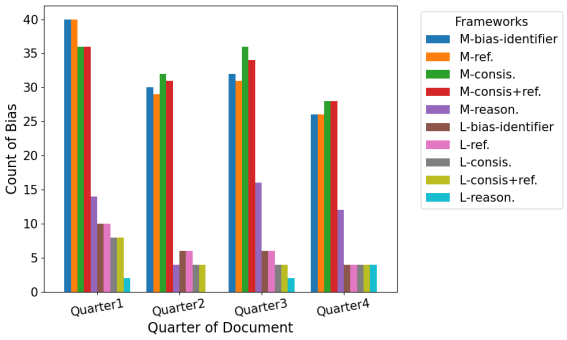
For the stigma-annotated MIMIC-IV dataset, we also compared to the rule-based baseline established by Himmelstein et al. (2022) and the supervised stigmatizing language detection model created by Harrigian et al. (2023) as an external point of comparison on the same data [4].

5 Evaluation

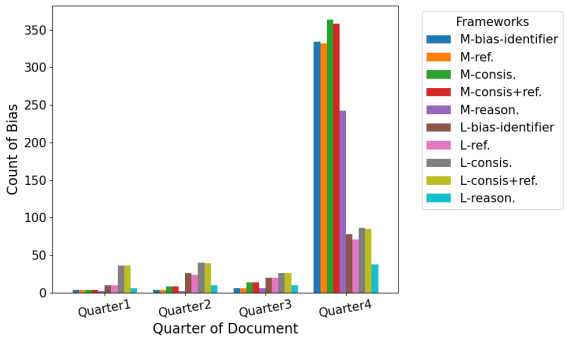
We evaluated our frameworks across PMC-Patients and MIMIC-IV. Experiments were run using four NVIDIA A100 80GB PCIe GPUs and runtime for the full suite of experiments was approximately four days. Models were loaded in native precision float (bfloat16) across four A100 GPUs. Decoding used the temperature 0.1 for bias-identifier and self-reflection frameworks, and temperature 0.5 for self-consistency and reasoning. For bias-identifier and self-consistency framework, max_length was set to 4096. Similarly, max_new_tokens was set to 1 and 250 for self-reflection and reasoning framework respectively. In the following subsections, we discuss stigmatizing language across datasets and documents (§5.1), the framework-level evaluation (§5.2), and comparison to external baselines (§5.3). For brevity, we

[3]We observed hallucination and context drift when longer passages were provided. The reasoning framework also benefited from shorter, focused spans to directly interrogate.

[4]Code is available at https://github.com/RajashreeDahal4/Stigmatizating_Language_identification.



a. PMC-Patients Summary Dataset.



b. MIMIC-IV Dataset.

Figure 3: Positional stigmatizing language distribution across datasets. M=*MedGemma*, L=*Llama*, ref.=bias-identifier+reflection framework, consis.=consistency framework, reason.=bias-identifier+reasoning framework, consis+ref.=consistency followed by reflection framework.

refer to `Llama-3.3-70B-Instruct` as *Llama* and `medgemma-27b-text-it` as *MedGemma* throughout §5.1–§5.3.

5.1 Stigmatizing Language across Datasets and Documents

We found that *bias-identifier* extracted 13 (Llama) and 64 (MedGemma) debate topics from PMC-Patients. It extracted 68 (Llama) and 178 (MedGemma) debate topics from MIMIC-IV. When dividing each note into equal-length quartiles, stigmatizing language was observed across all quartiles for PMC-Patients (Figure 3), with a slight elevation in the first quartile, whereas it was heavily concentrated in the final quartile for MIMIC-IV. MedGemma extracted more stigmatizing language across all quartiles for PMC-Patients, and in the last quartile for MIMIC-IV notes. The more frequent presence of stigmatizing language near the end of MIMIC-IV notes suggests that it may feature more prominently in evaluative language, which is usually found towards the end of healthcare provider notes. Except for reasoning, stigmatizing language concentrations were consistent regardless of which framework (see §4) was implemented using a particular LLM.

5.2 Framework Performance

We compared stigmatizing language identified by all proposed frameworks (*bias-identifier*, *self-consistency*, *self-reflection*, and *reasoning*) across both datasets using both models. We observed that the *self-reflection* and *reasoning* frameworks penalized documents or remained consistent with lower stigmatizing language counts in most cases (see Figure 3). *Self-consistency-reflection* did not show a similar trend. This may due to the voting strategy involved in making the final prediction. The main effect of self-reflection and reasoning was to prune spurious predictions when applied on top of *bias-identifier*, reducing ambiguity in detected stigmatizing language locations. Similarly, the number of debate topics identified by *bias-identifier* and *self-consistency* were comparable, perhaps due to the LLM's confidence at generating similar debate topics across different runs. Adding self-reflection to *bias-identifier* reduced this disparity by a small margin, highlighting stigmatizing language more across both datasets.

MedGemma and Llama exhibited different trends across frameworks (see Table 1). MedGemma's reasoning framework outperformed the baseline across both datasets, whereas Llama demonstrated its lowest performance with the reasoning framework. This was interesting given that MedGemma is less than 2.5 times the size of Llama, and suggests that MedGemma's pretraining data (including a diverse set of medical text) may enable it to leverage the reasoning framework to a much fuller extent in the structured clinical setting than a larger model pretrained on general-purpose data. Also, Llama's reasoning drop might be from sparse initial extractions (13 debate topics vs MedGemma's 64, where overpruning by debate agents devastates recall). The self-consistency framework in turn demonstrated superior performance for Llama. Finally, we observed that while reasoning and consistency-based prompts sometimes increased performance, the gains were not consistent across datasets (see, e.g., Llama on PMC-Patients).

From a qualitative perspective, some examples of false positive quotes included "he did not comply with during his admission," "homeless male with long history of etoh and heroin abuse," and

Table 1: Performance across models*.

Model	MIMIC(m)			PMC(p)		
	Framework	F_1	A	Framework	F_1	A
MedGemma	m-bias-identifier	0.36	0.34	p-bias-identifier	0.54	0.66
MedGemma	m-refl.	0.36	0.35	p-refl.	0.55	0.68
MedGemma	m-reason.	0.39	0.42	p-reason.	0.59	0.80
MedGemma	m-consis.	0.37	0.28	p-consis.	0.49	0.62
MedGemma	m-consis+refl.	0.37	0.28	p-consis+refl.	0.51	0.65
Llama	m-bias-identifier	0.34	0.73	p-bias-identifier	0.43	0.82
Llama	m-refl.	0.34	0.73	p-refl.	0.43	0.82
Llama	m-reason.	0.34	0.73	p-reason.	0.20	0.77
Llama	m-consis.	0.51	0.73	p-consis.	0.43	0.82
Llama	m-consis+refl.	0.51	0.73	p-consis+refl.	0.43	0.82

* F_1= F_1 score and A = accuracy. A total of 71 chunks were human-labeled, each treated as either stigmatizing or not.

"pt's paranoia seemed to be based in reality as the other resident in past had 'picked on' a gay resident." Some of these flagged examples are qualitatively ambiguous and may require closer examination in future work.

5.3 Comparison to External Baselines

We aligned Harrigian et al. (2023)'s fine-grained work on the MIMIC-IV stigmatizing language subset by mapping their "Exclude," "Neutral," and "Positive" model predictions to *non-stigmatizing* and the rest ("Disbelief," "Difficult," "Negative") to *stigmatizing*. We also fine-tuned the `emilyalsentzer/Bio_ClinicalBERT`[5] (Alsentzer et al., 2019) model on the binarized MIMIC-IV stigmatizing language data (*baseline-stigma*) using a 0.7/0.2/0.1 train-test-validation split. The 7 MIMIC-IV evaluation notes were part of test split. All reasoning frameworks outperformed these external alternatives (Figure 4) by a wide margin across models and datasets, with the exception of MedGemma with MIMIC-IV (`M-m`). This exception might be due to the poor *bias-identifier* score upon which the reasoning framework is dependent. The *bias-identifier* framework outperforms the alternatives in general though, justifying the additional computational cost of LLM-based approaches for improved performance at identifying stigmatizing language. While *baseline-stigma* slightly outperformed *L-bias-identifier*, we observed that *L-consistency* and *L-consistency + reflection* outperformed by a wider margin (13.5% improvement in precision). The *baseline-stigma* method achieved a perfect recall (1.0) for stigmatizing language but low precision (0.37).

[5]We implemented Bio_ClinicalBERT because the original work by Harrigian et al. (2023) implemented this model in their work.

6 Discussion and Conclusion

Through the study of self-reflection, self-consistency, and self-reasoning frameworks, our work sheds light on the use of LLMs to detect stigmatizing language across two popular clinical datasets. We found that *bias-identifier* tended to over-predict stigmatizing language with low precision. Self-reflection improved (or remained constant) over this outcome, and self-consistency introduced stability across multiple runs. MedGemma generally had higher performance using the STIGMAD framework across both datasets, although Llama achieved stronger performance using the self-consistency framework specifically. Our analysis shows that self-reflection and self-consistency can provide comparable results to more complex self-reasoning frameworks. Self-reasoning outperforms alternative frameworks when used with MedGemma, although not when used with the general-domain Llama. This might be because Llama, as a general-purpose model lacks clinical domain grounding to correctly adjudicate borderline cases during debate.

We focused our investigation on building comprehensive understanding of the capacity of STIGMAD to recognize stigmatizing language in clinical settings. The self-consistency framework itself validates performance stability, through its inherent aggregation of multiple sampled outputs (repeated runs) of *bias-identifier*. While STIGMAD incurs higher computational cost compared to simpler approaches (e.g., *baseline-stigma*), our study reveals what appears to be a clear performance tradeoff. For those wishing to study stigmatizing language in their own clinical environments as a downstream

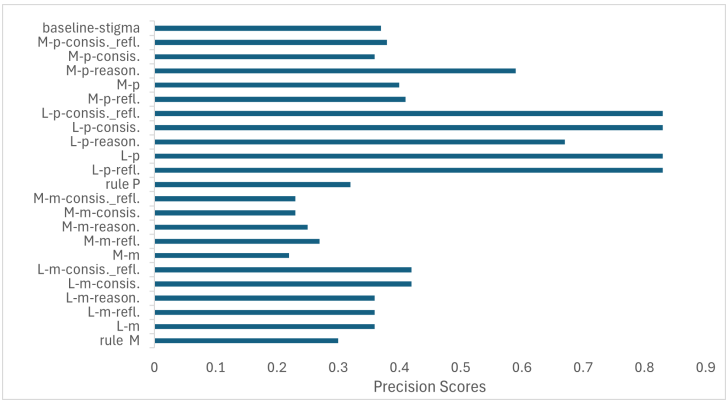


Figure 4: We compare precision scores of the rule-based method (*rule*) proposed by Himmelstein et al. (2022) and the baseline stigma detection method (*baseline-stigma*) from Harrigian et al. (2023) on the MIMIC-IV dataset, alongside our proposed frameworks evaluated on both MIMIC-IV and PMC using two LLMs. Here, `reason.` denotes reasoning, `refl.` denotes self-reflection, `consis.` denotes consistency, `p` refers to PMC, `m` refers to MIMIC-IV, `L` denotes Llama, and `M` denotes MedGemma.

application, we recommend weighing the relative performance advantages of STIGMAD against the available computational infrastructure and/or associated costs to build such an infrastructure.

Applying reasoning frameworks to stigmatizing language detection in clinical documents is an emerging topic; thus, a large part of the innovation for the work reported here lies in the task formulation and the discovery of domain-specific debate protocols. The resulting performance advantages compared to internal and external baselines highlights the strength of the implementation. Overall, our results confirm that open-source LLMs, steered by structured prompting and agent roles through STIGMAD, are well-positioned as a tool to detect stigmatizing language in clinical text. Our hope is that this work acts as a stepping stone to the development of responsible clinical NLP systems for stigmatizing language detection, as well as for other forms of bias detection and auditing.

Limitations

We aim to scale this research to larger, more heterogeneous clinical datasets to evaluate performance across more diverse clinical settings. We will also experiment with smaller, resource-light models for self-consistency and self-reflection to allow for faster and more scalable clinical deployment.

Our research was carried out on 57 patient files from MIMIC-IV (n=7) and PMC-Patients (n=50), which may not be representative of the overall datasets. We also acknowledge that annotating stigmatizing language requires interpreting tone and writer intent from text alone, which naturally opens the door to ambiguity. To mitigate the potential for this, annotators were trained to closely follow stigmatizing language definitions grounded in contemporary literature. Moreover, our annotation process was guided by experts in both the clinical and computational domains.

While individual outputs from LLMs are inherently non-deterministic, our framework improves internal consistency by structuring the generative process, supporting reproducible evaluation pipelines. Temperature values were set according to the conventions of their respective frameworks with low temperature for deterministic single-pass extraction and moderate temperature for diversity-dependent voting. Here, a formal sensitivity analysis across temperature remains an important direction for future work. The future work should also include sensitivity analysis over the self-consistency voting threshold and number of rounds for debate, which were set based on computational constraint and task-specific reasoning rather than formal hyperparameter search.

Finally, analyzing the influence of de-identification on stigmatizing language annotation and automated identification is beyond the scope of our research; however, it presents an intriguing avenue for future work. Likewise, broadening the analysis to include HIPAA-compliant closed-source models (e.g., ChatGPT for Healthcare) was infeasible for the current study due to cost and access constraints, but would present a valuable next step for external validation.

Ethical Considerations

This study was reviewed by the Institutional Review Board at the University of Illinois Chicago and determined not to be human subjects research according to Department of Health & Human Services (DHHS) or US Food and Drug Administration (FDA) regulations. While no direct patient interaction was involved, this research touches on ethically sensitive areas, including bias in healthcare. Our research includes a publicly available, de-identified clinical dataset (MIMIC-IV notes) and the PMC-Patients Summary dataset. While our goal is to identify bias in clinical notes, our findings are intended for academic analysis and model improvement, not direct clinical use. We have used open-source LLMs in our local machine to maintain privacy. However, we acknowledge the limitations of such models in reliably detecting nuanced or systemic bias, and emphasize that automated outputs should not replace human judgment in clinical settings.

Acknowledgments

We thank the anonymous reviewers for their helpful suggestions. N. Parde was partially supported during this work by the National Science Foundation under Grant No. 2125411 and the National Institutes of Health under Grants R41NR020667, 1R61DA057629-01A1, and 1R01AG091762-01. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

Deepak Bhaskar Acharya, Karthigeyan Kuppan, and Divya Bhaskaracharya. 2025. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*, 13:18912–18936.

Toufique Ahmed and Premkumar Devanbu. 2023. Better patching using llm prompting, via self-consistency. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1742–1746.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Donald U. Apakama, Kim-Anh-Nhi Nguyen, Daphnee Hyppolite, Shelly Soffer, Aya Mudrik, Emilia Ling, Akini Moses, Ivanka Temnycky, Allison Glasser, Rebecca Anderson, Prathamesh Parchure, Evajoyce Woullard, Masoud Edalati, Lili Chan, Clair Kronk, Robert Freeman, Arash Kia, Prem Timsina, Matthew A. Levin, and 8 others. 2024. Identifying and characterizing bias at scale in clinical notes using large language models. *medRxiv*.

Veronica Barcelona, Danielle Scharp, Betina R. Idnay, Hans Moen, Kenrick Cato, and Maxim Topaz. 2024. Identifying stigmatizing language in clinical documentation: A scoping review of emerging literature. *PLOS ONE*, 19(6):e0303653.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, Miami, Florida, USA. Association for Computational Linguistics.

Élan Burton, Brenda Flores, Barbara Jerome, Michael Baiocchi, Yan Min, Yvonne A Maldonado, and Magali Fassiotto. 2022. Assessment of bias in patient safety reporting systems categorized by physician gender, race and ethnicity, and faculty rank: a qualitative study. *JAMA Network Open*, 5(5):e2213234–e2213234.

Feng Chen, Liqin Wang, Julie Hong, Jiaqi Jiang, and Li Zhou. 2024a. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, 31(5):1172–1183.

Hongbo Chen, Myrtede Alfred, and Eldan Cohen. 2025. Efficient detection of stigmatizing language in electronic health records via in-context learning: comparative analysis and validation study. *JMIR Medical Informatics*, 13(1):e68955.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2024b. Teaching large language models to self-debug. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and

510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

James Eschrich and Sarah Sterman. 2024. A framework for discussing llms as tools for qualitative analysis. *arXiv preprint arXiv:2407.11198*.

Chloë FitzGerald and Samia Hurst. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC medical ethics*, 18(1):19.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Euee Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. 2023. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329.

Adam S Hayes. 2025. "conversing" with qualitative data: Enhancing qualitative research through large language models (llms). *International Journal of Qualitative Methods*, 24:16094069251322346.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). RRID:SCR_007345.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Adam Gayles, Ahmad Shammout, Steven Horng, Tom J. Pollard, Sixiang Hao, Benjamin Moody, Benjamin Gow, Li-wei H. Lehman, Leo Anthony Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2025. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7075–7092, Albuquerque, New Mexico. Association for Computational Linguistics.

Yizhi Liu, Weiguang Wang, Guodong Gao, and Ritu Agarwal. 2023. The impact of stigmatizing language in ehr notes on ai performance and fairness. In *Proceedings of the International Conference on Information Systems (ICIS)*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. 2026. From belief entrenchment to robust reasoning in llm agents. *Preprint*, arXiv:2503.16814.

J. Park, D. Kotzias, P. Kuo, R.L. LoganIV, K. Merced, S. Singh, M. Tanana, E. KarraTaniskidou, J.E. Lafata, D.C. Atkins, M. Tai-Seale, Z.E. Imel, and P. Smyth. 2019. Detecting conversation topics in primary care office visits from transcripts of patient–provider interactions. *Journal of the American Medical Informatics Association*, 26(12):1493–1504.

Shan Qiao, Xingyu Fang, Camryn Garrett, Ran Zhang, Xiaoming Li, and Yuhao Kang. 2024. Generative ai for qualitative analysis in a maternal health study: Coding in-depth interviews using large language models (llms). *medRxiv*.

Rubina F. Rizvi, Kathleen A. Harder, Gretchen M. Hultman, Terrence J. Adam, Michael Kim, Serguei V.S. Pakhomov, and Genevieve B. Melton. 2016. A comparative observational study of inpatient clinical note-entry and reading/retrieval styles adopted by physicians. *International Journal of Medical Informatics*, 90:1–11.

Michael D. Rozier, Kavita K. Patel, and Dori A. Cross. 2022. Electronic health records as biased tools or tools against bias: A conceptual model. *The Milbank Quarterly*, 100(1):134–150.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall,

and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record. *Health Affairs*, 41(2):203–211.

Alissa A. Valentine, Lauren A. Lepow, Alexander W. Charney, and Isotta Landi. 2024. The point of view of a sentiment: Towards clinician bias detection in psychiatric notes. *CoRR*, abs/2405.20582.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43(1):477–501.

Andrew Walker and 1 others. 2025. Care-sd: classifier-based analysis for recognizing provider stigmatizing and doubt marker labels in electronic health records: model development and validation. *Journal of the American Medical Informatics Association*, 32(2):365–374.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Christopher J Wong and Sara L Jackson. 2023. *The Patient-Centered Approach to Medical Note-Writing*.

Teenu Xavier, Jane M Carrington, and Joshua Lambert W. 2026. Detecting stigmatizing language with large language models: mind the settings. *JAMIA Open*, 9(2):ooag037.

Zhongbin Xie, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. Counter-GAP: Counterfactual bias evaluation through gendered ambiguous pronouns. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3761–3773, Dubrovnik, Croatia. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

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*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10 1:909.

A Prompt Templates

Prompt A.1: Extracting Bias Type and Quote for Debate Topic

You are tasked with identifying if any of the following types of bias are present in the given medical note. If so, name the bias type and provide the quote that highlights the bias type in the given medical note. The definitions for bias type are:

{other bias type definitions. . . }

Stigmatizing Language – Stigmatizing language is any language in clinical documentation that portrays the patient in a negative or judgmental way—by expressing doubt about their credibility, implying they are difficult, noncompliant, or deceptive, or by using language that casts their behavior, attitude, or emotional state in a negative light. This includes both overt and subtle wording that reflects blame, shame, or disrespect. Example: "claims to see a therapist".

Note: You are supposed to answer the bias and the corresponding quote which highlights the bias in the form: `[{"bias_type":"", "quote":""},{"bias_type":"", "quote":""}]`. Please write as many bias type as you can.

Note: Your output should explicitly be in the above mentioned format and do not give any other justification. If no bias is found, just answer `[{"bias_type":"", "quote":""]}`. Do not repeat the same bias type with the same quote. Make sure your answer starts with only one "[" and ends with only one "]". You can add as many unique bias types and quote pairs within the brackets in the given format.

Medical Note:

`##medical_note##`

Answer:

Prompt A.2: Prompt used for self-reflection framework

You are a medical language model trained to identify bias in clinical notes. You previously predicted bias type in the input paragraph and the corresponding quote that highlights the bias. Now reflect on that decision. Based on the definition for bias type and input paragraph, was your prediction for bias type correct? Use the information to determine whether the quote, as it appears in the paragraph, matches the bias_type. You are supposed to answer a single digit 1 if there exists bias mentioned above, else answer single digit 0. Your answer should explicitly be a single digit 1 or 0. Do not explain or justify your answer.

###Note####: Ensure that you do not give any justification and reasoning for your answer.

###Example1:

****quote****: "The patient is non-compliant with their medication regimen."

****Paragraph****: "Despite multiple counseling sessions, the patient is non-compliant with their medication regimen and shows little interest in improving their condition."

****bias type****: Explicit Bias

**** The definition of Explicit Bias is ****: Explicit forms of bias include preferences, beliefs, and attitudes of which people are generally consciously aware, endorsed, and can be identified and communicated (e.g., labeling a patient as

'non-compliant' or 'irresponsible').)
Answer: 1
Example2:
****quote****: "feeling tired during the day."
****Paragraph****: "The patient reports feeling tired during the day but denies any issues with sleep or recent lifestyle changes. Vitals are within normal range."
****bias type****: Explicit Bias
**** The definition of Explicit Bias is ****: Overtly negative or judgmental language used in a medical note (e.g., labeling a patient as 'non-compliant' or 'irresponsible')
Answer: 0
Question:
****quote****: ##quote##
****Paragraph****: ##paragraph##
****bias type****: Stigmatizing Language
**** The definition of Stigmatizing Language is ****:
##definition##.
Answer:

Prompt A.3: Prompt for Moderator

You are a moderator overseeing a multi-agent debate involving medical bias detection. Two expert agents will analyze the same medical note and debate whether the given bias type is present, and if so, which type and why. After each round, you must assess both perspectives and decide if a consensus is reached.

Prompt A.4: Prompt for Affirmative Agent

The debate topic is ##debate_topic##
It was previously predicted that the given bias type in the input paragraph and the corresponding quote highlights the bias. Now reflect on that decision with your reasoning. Based on the definition for bias type and input paragraph, was the prediction for bias type correct? Please provide your single reasoning in less than 3 sentences. Your reasoning should be focused on the mentioned bias type. Do not introduce new type of bias. You are not supposed to ask any questions.

Prompt A.5: Prompt for Negative Agent

##aff_ans##
You are supposed to disagree with my answer. Please provide your single reasoning in less than 3 sentences. Your reasoning should be focused on the mentioned bias type. Do not introduce new type of bias. You are not supposed to ask any questions.

Prompt A.6: Prompt for Judge Reflection

Affirmative side's answer: ##aff_ans##
Negative side's answer: ##neg_ans##
Write down key pointers addressed by different experts that could be crucial to reflect on before making judgement. You are not supposed to introduce your personal pointers. Try to limit the pointers to as low as 5 sentences.

Prompt A.7: Prompt for Judge Decision

Now, based on the debate topic:
##debate_topic## Provide only one reasoning and the final decision on whether the given bias exists, and explanation. Respond strictly in JSON format: `{"Presence":"True/False","Reason": ""}`. You should provide the result in a single JSON form. Do not give more than one json answer. Your response should be explicitly in the given json format where json key and values are double quoted. You are not allowed to give any more clarification apart from the single json form.