Exploring Bias in Clinical Notes using Open-Source Large Language Model

Anonymous ACL submission

Abstract

Clinical documentation, available in many forms, plays a critical role in patient care, billing, and medical research. Bias in clinical documentation adversely affects these processes. We explore ways to use open-source Large Language Models (LLMs) for bias detection in medical notes drafted by providers. We considered implicit and explicit bias in clinical documentation, i.e., prejudiced or biased judgment that impacts medical decisions, clinical, cultural/racial and gender identity bias, and the use of stigmatizing language. Through this investigation, we study three core capabilities of LLMs in clinical contexts: reasoning (the ability to make logical inferences or structured thinking to solve problems), reflection (the ability to revise or critique their own outputs), and consistency (the ability to produce stable and reproducible responses across same prompts). To assess these capabilities in practical and clinically relevant contexts, we conducted experiments on both clinical notes that correspond to patient summaries from the PMC-Patients dataset and clinical notes from the MIMIC-IV dataset. Human evaluation and experiments across different frameworks demonstrated that our proposed frameworks exhibit measurable performance advantages compared to the baseline framework. Our finding suggests that open-source LLMs can support bias detection and mitigation in the healthcare domain.

1 Introduction

Clinical notes, which come in many forms including history and physical examination notes, daily progress notes, special consults, surgery or procedure reports, discharge summaries, and nurses' notes, are an important category of electronic health records (EHRs) that narrate patient care. Holistically, these various forms of clinical notes provide continuity of care, keeping health care teams on the same page and helping support activities such as billing and medical research (Wrenn et al., 2017). They may be supplemented by other free-text forms of healthcare documentation such as interview transcripts, patient-provider communication logs, and survey responses, especially in research or mental health settings (Obermeyer et al., 2019; Chen and et al., 2021).

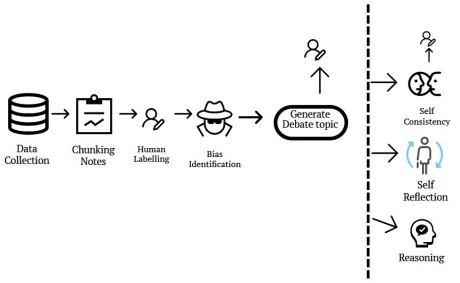


Figure 1: Methodological framework for bias detection.

However, clinical notes and other free-text healthcare documentation may vary widely in tone, structure, and completeness. Bias 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 and et al., 2022). Stigmatizing language and assumptions in documentation can affect downstream AI and human system decisions (Chen and et al., 2021) and produce inaccurate results (ROZIER et al., 2022). Bias in data 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 (Wong and Jackson, 2023). This illustrates the need for more standardized and fair documentation practices in clinical settings.

Clinical documentation of bias has traditionally been uncovered through manual chart review, qualitative coding, and formal interviewing. While dense with subtle observation, these are time-consuming, expert opinion-reliant, and generally

not scalable; they also hold less promise for detecting subtle or implicit bias dispersed across enormous volumes of text. As artificial intelligence rapidly advances, it is increasingly being leveraged to uncover patterns of bias in clinical notes at scale (Zhang et al., 2020; Chen and et al., 2021; Obermeyer et al., 2019), for example by detecting stigmatizing language, classifying note sentiment (Valentine et al., 2024), or predicting bias-prone words. However, automatically detecting such bias is challenging due to its implicit nature and continually evolving language in healthcare settings. Approaches to date often rely on fixed labeled data and miss subtle, context-dependent bias.

We propose a novel approach that leverages self-consistency, self-reflection, and a multi-agent debate framework to identify bias using open-source LLMs. This allows models to reason through ambiguity and critically analyze output, filling an existing gap. Our approach illustrates the potential of these models to uncover hidden patterns in clinical language in the form of bias and offers a more adaptive strategy for identifying and analyzing bias in medical documentation. We aim to answer the following research questions:

- How do common forms of bias manifest in clinical notes?
- How effectively can open-source LLMs detect bias in clinical notes in terms of its confident, reflective, reasoning and zero-shot abilities?
- Does bias present in real clinical notes propagate in the same way in clinical summary notes?
- Can multi-agent debate frameworks enhance bias detection in clinical notes over self-consistency and self-reflection approaches?

Experiments show that our proposed methods outperform standard prompting in detecting biases.

2 Literature Review

Chen et al. (2024) systematically reviewed AI-based bias mitigation and detection strategies for EHR data, identifying six common types of bias, including implicit bias, selection bias, algorithmic bias, and temporal bias. Similarly, Ma et al. (2022) investigated bias in AI models for surgical applications, showing that competency assessment models for surgeons over- or underestimates some subgroups of surgeons. FitzGerald and Hurst (2017) discusses the presence and impact of implicit biases in healthcare professionals, performing a meta-analysis of findings from 42 studies. They establish that healthcare professionals have extensive implicit biases across numerous areas, including race, ethnicity, and gender. Another qualitative study by (Burton et al., 2022) of 401 patient safety reports found unequal reporting by physicians' gender and race/ethnicity. Subthemes such as "inappropriate communication," "verbal abuse," "ignoring or omission of procedure," and "physical intimidation" were regularly found.

Wong and Jackson (2023) discussed positive bias, negative bias, stigmatizing language, implicit bias, explicit bias, gender bias, and racial bias, and highlighted common domains in which bias may arise in healthcare domains. Similarly, Vela et al. (2022) highlighted stigmatizing language, racial/cultural bias, explicit bias, implicit bias, and gender discrimination in their study. Ultimately, we identified that six forms of bias (explicit bias, implicit bias, clinical bias, stigmatizing language, cultural/racial bias, and gender identity bias) hold fairly constant across the reviewed literature. We focus on those forms of bias in our study.

Because of the labor-, resource-, and time-intensive nature of qualitative analysis, several studies have looked into the potential of LLMs to aid such processes (Qiao et al., 2024), (Hayes, 2025), (Eschrich and Sterman, 2024). Apakama et al. (2024) evaluated GPT-4's zero-shot ability to detect and categorize bias in clinical notes, specifically in emergency department and discharge notes. However, since healthcare data is highly sensitive and may contain identifiable information, closed-source models like ChatGPT should not be considered a general solution for health-related bias analyses. In light of this, we explore the potential of open-source LLMs like Llama3.1-8b-instruct (Grattafiori et al., 2024) and their capability to detect bias in clinical notes.

Finally, Acharya et al. (2025) conducted a comprehensive study on the use of agentic AI. Based on the study, we aim to explore Multi-Agent Systems (MAS) as an agentic AI framework for bias detection, as a potential alternative to more conventional zero-shot and few-shot prompting approaches. Similarly, (Du et al., 2023) explored the reasoning ability of LLMs through a multi-agent debate framework. This work inspired us to experiment with reasoning (Liang et al., 2024),

self-reflection Renze and Guven (2024), and self-consistency (Ahmed and Devanbu, 2023) in detecting bias in medical notes using open-source LLMs.

3 Methodology

3.1 Datasets

We downloaded the PMC-Patients summary dataset (Zhao et al., 2023) from HuggingFace and obtained access to the MIMIC IV Dataset (Johnson et al., 2020) as our testbed of real-world clinical notes.[1] Accessing MIMIC-IV notes required credentialed access via PhysioNet and agreement to a Data Use Agreement (DUA). 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 the PMC-Patients dataset including patient visit, medical history, symptoms, treatments, discharge summary and intervention based on case reports in PubMed Central (PMC). Similarly, the MIMIC-IV dataset is a collection of de-identified free-text clinical notes for patients and contains 331,794 de-identified discharge summaries from 145,915 patients admitted to the hospital and emergency department at the Beth Israel Deaconess Medical Center in Boston, Massachusetts, USA. It also contains 2,321,355 Radiology reports.

The average word count and standard deviation for the MIMIC-IV dataset was 1501.70 and 465.06, respectively, and 288.58 and 113.28 for the PMC-Patients summary dataset. Considering the difference in average count of words between the datasets, we selected 16 MIMIC-IV and 50 PMC-Patients summary notes for annotation, as these quantities provide us with approximately equal numbers of context length-limited chunks to be processed by our downstream LLM. Three English-speaking MS and PhD-level computer science students manually added verified bias labels at the chunk level for both datasets, based on the established definitions of bias type.

3.2 Types of Bias

As noted earlier, we selected six commonly occurring and well-documented types of bias in clinical text, focusing on those that were repeatedly observed across studies and are known to impact healthcare disparities. These bias types include:

- **Explicit Bias:** Includes the use of judgmental or overly negative phrases in a patient's medical notes. Examples include calling a patient "non-compliant" or "irresponsible." Terms like "non-compliant" can reflect bias even without group disparities, they imply blame and can damage the patient-provider relationship.

- **Implicit Bias:** Includes stereotypes or labels associated with someone's ethnicity/age/gender/class without articulating it verbally. An example would be referring to a patient as "well-spoken for their background."

- **Clinical Bias:** Consists of any prejudice or biased judgment that impacts medical decisions, interactions, or results, typically unconsciously (e.g., "Chest pain likely stress-related; no further workup needed.").

- **Stigmatizing Language:** Includes blameful or shame-inducing language, especially regarding addiction, obesity, or mental illness (e.g.,"drug abuser," "refused care," or "alcohol abuser").

- **Cultural/Racial Bias:** Mentions of a patient's race, ethnicity, or culture in a way that is unnecessary for care or based on stereotypes (e.g., "Typical behavior for his community").

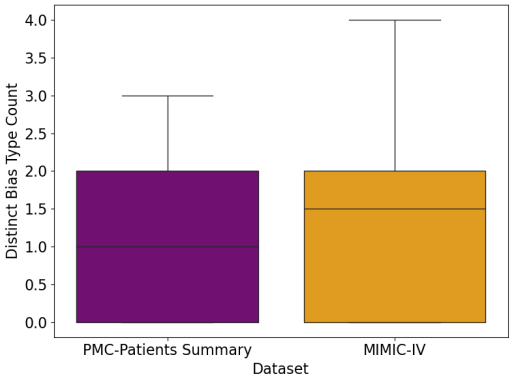


Figure 2: Figure: Box plot comparing the distribution of distinct bias types identified per clinical note in the PMC-Patients Summary and MIMIC-IV datasets. (Note: Considering 16 MIMIC notes and 50 PMC-Patients Summary notes only).

[1]This study was reviewed by the Institutional Review Board 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.

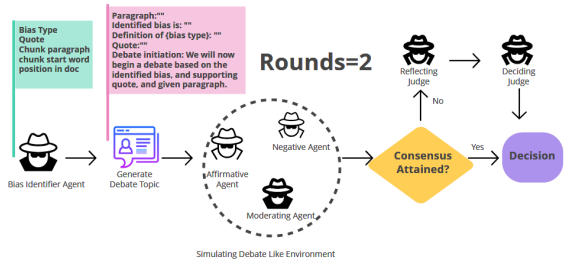


Figure 3: Agentic AI based Reasoning Framework for Bias Detection.

- **Gender Identity Bias:** Includes misgendering or failure to respect a patient's gender identity, for example through dismissive or inaccurate documentation.

While the addressed bias types are conceptually distinct, we acknowledge that overlap can occur. For example, stigmatizing language may also express implicit or clinical bias. We allow notes to be tagged with multiple bias types if applicable.

3.3 Baseline Framework

Since existing work toward analyzing bias in clinical notes using LLMs is limited, we treat the extracted debate topic as a baseline (*bias-identifier framework*) compared to our proposed frameworks. Given a clinical note as input, the role of the *bias-identifier* agent is to generate debate topic-worthy information. We leverage an open-source LLM, the `Llama-3.1-8B-Instruct` model, to generate the debate topic using a three-step process. We first ask the LLM to identify bias type, based on a relevant bias quote from the input paragraph. We set the LLM temperature to 0.1 to reduce randomness and encourage more deterministic bias detection to maintain logical consistency. Since medical notes can vary greatly in size depending on notetaking context, we chunk the note into a window size of 80 words, with a chunk overlap of 10 words. Each chunked paragraph is fed to the LLM to extract bias types and quotes; the prompt used for this process is shown in Prompt. A.1. We conducted this process in a zero-shot manner, defining bias types as part of the prompt.

3.4 Proposed Frameworks

Our reasoning, self-consistency, and self-reflection frameworks (see Figure 1) are described below.

3.4.1 Self-reflection Framework

Our self-reflection framework builds on prior predictions, revisiting a previously identified quote and bias type from a given clinical note, along with the definition of the corresponding bias type. It then re-evaluates whether the quote within the context of the input note truly reflects the specified type of bias. Based on this comparison, the model provides a boolean output without providing any explanation. We used few-shot prompting to implement the self-reflection framework; the prompt used for this process is shown in Prompt. A.2.

3.4.2 Self-Consistency Framework

Our self-consistency framework is analogous to the baseline model; however, for self-consistency we run this model eight times while increasing the temperature to 0.5 to promote creative and diverse responses. The added variability makes it easier to identify consistent outputs through voting across runs. We consider only those outputs that occur at least three times out of eight runs for a given input chunk. We keep this threshold low since the *bias-identifier* agent also extracts quotes, and this quote extraction may vary across different runs. The prompt used to implement the self-consistency framework is the same as that of the baseline framework (see Prompt. A.1).

3.4.3 Reasoning Framework

The quotes and bias types generated by the *bias-identifier* agent are converted into a debate topic as shown in Fig. 3. The debate topic includes the input clinical note chunk, identified bias type, definition of identified bias type, quote that supports the identified bias, and debate initiation process. The player prompt used for the debate is shown in Prompt. A.3. Our reasoning framework is inspired by Liang et al. (2024), who showed that multi-agent debate encourages diverse perspectives and helps overcome limitations of self-reflection, such as overconfidence and repetitive thinking. This design aligns with our goal of promoting deeper, more reliable reasoning about bias in clinical notes. Our debate-based reasoning framework comprises four types of agents, defined below.

Affirmative Agent. This agent has the right to choose whether it wants to support or deny the debate topic. It supports and justifies its reflection on that topic, arguing affirmatively in brief and structured statements. The prompt used for the affirmative agent is shown in Prompt. A.5.

Negative Agent. This agent challenges or critiques the point of view of the affirmative agent. It

provides arguments or reasons supporting its position (i.e., whether the bias type exists). The prompt used for the negative agent is shown in Prompt. A.6.

Moderator Agent. This agent oversees debate rounds and analyzes responses from affirmative and negative agents. It determines whether consensus on bias classification has been reached and explicitly summarizes reasons supporting its final decision. It ensures consistency across debates, reassesses contentious cases, and validates that the consensus reached aligns with established bias definitions. Prompt. A.4 highlights how the moderator agent should behave, and Prompt. A.7 shows the prompt based on which it should respond. The moderator agent moderates the debate process for a total of two rounds. If consensus is reached at the final round, the decision of the moderator agent is the final decision. The decision across the first round is also tracked for further analysis.

Judge Agent. In the reasoning framework, the judge agent is only triggered when no consensus is reached at the end of two rounds. To mimic a real-life judge scenario, the judge agent takes two roles. The *reflecting judge* agent compiles final arguments from both affirmative and negative agents after the debate rounds conclude, by listing bias classifications and reasoning without initial evaluation. The prompt used for the reflective judge agent is shown in Prompt. A.8. Similarly, the other agent, which plays the role of *deciding judge*, is responsible for making the ultimate decision regarding the presence of bias based on its initial reflection. The prompt used for this agent is shown in Prompt. A.9.

4 Evaluation

We discuss how common forms of bias manifest in clinical notes in Section 4.1.1. Bias propagation across two datasets are studied in Section 4.1.2. Llama 3.1's effectiveness in detecting bias in clinical notes in terms of its confident, reflective, reasoning and zero-shot abilities is highlighted in Sections 4.1.3 and 4.2 We evaluated our proposed framework relative to the baseline using qualitative and quantitative methods. The experiment took around 5 days to run and was carried out on an NVIDIA A100 80GB PCIe GPU.

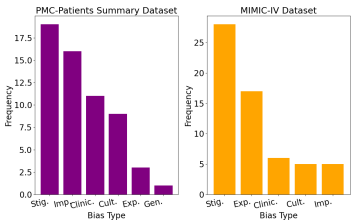


Figure 4: Distribution of bias types in human-provided labels across datasets. *Left:* PMC-patients summary dataset. *Right:* MIMIC-IV dataset. *Stig.*=Stigmatizing Language, *Imp.*=Implicit Bias, *Clinic.*=Clinical Bias, *Cult.*=Cultural/Racial Bias, *Exp.*=Explicit Bias, *Gen.*=Gender Identity Bias.

4.1 Qualitative Results

4.1.1 Bias Distribution Across Datasets

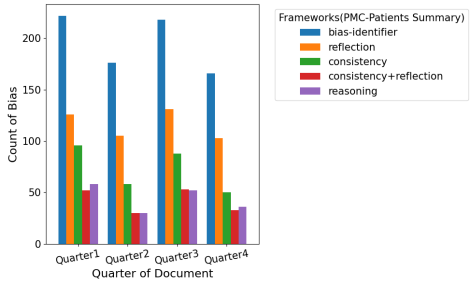
The *bias-identifier* framework, which serves as the basis of all frameworks, extracted 409 debate topics from the PMC-Patients Summary dataset (which had 57 true potential debate topics, or human-labeled types of bias). It also extracted 398 debate topics from MIMIC-IV (which had 64 true biases). Figure 4 summarizes the frequency of each of our six bias types in these notes, confirming the existence of bias in both MIMIC-IV and PMC-Patient summary notes and revealing more detailed bias patterns. Stigmatizing language was the most commonly used form of bias across both notes. Interestingly (and encouragingly), gender identity bias was not observed in MIMIC-IV.

4.1.2 Bias Position Across Documents

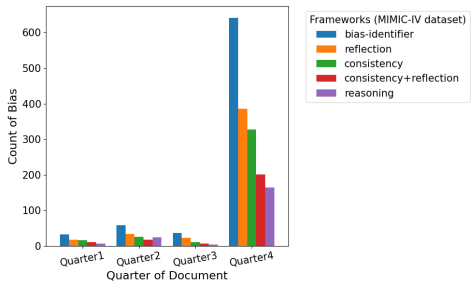
To examine the positional distribution of biases across documents, each note was divided into equal-length quartiles. Fig. 5 shows that bias was distributed relatively evenly across quartiles for the PMC-Patients summary dataset. However, MIMIC-IV had a heavy concentration of bias in the last quartile. This suggests that subjective or evaluative language, which may imply bias, is usually found towards the end of healthcare provider notes.

4.1.3 Bias Type Distribution across Frameworks

We compared the biases identified by all baseline and proposed frameworks (*bias-identifier*, *self-consistency*, *self-reflection*, *self-consistency-reflection*, and *self-reasoning*) across both datasets. We observed that the proposed *self-consistency*, *self-reflection*, *self-consistency-reflection*, and *self-reasoning* frameworks penalized documents with

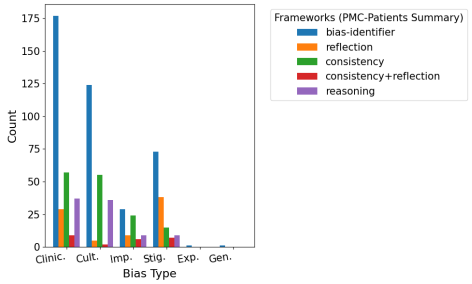


(a) PMC-Patients Summary Dataset.

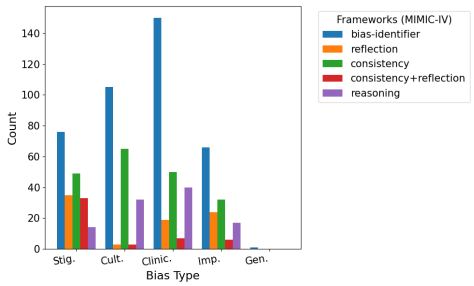


(b) MIMIC-IV Dataset.

Figure 5: Positional bias distribution across the MIMIC-IV and PMC-Patients Summary datasets.



(a) PMC-Patients Summary Dataset.



(b) MIMIC-IV Dataset.

Figure 6: Bias types identified by different frameworks across the MIMIC-IV and PMC-Patients Summary datasets.

lower bias counts (see Fig. 6). The main effect of self-reflection and self-reasoning was to prune spurious predictions when applied on top of the baseline *bias-identifier*, reducing ambiguity in detected bias locations. Similarly, self-consistency filtered spurious predictions based on its confidence (i.e., number of votes). Adding the self-reflection framework on top of the self-consistency framework further pruned the number of predictions.

The *bias-identifier* framework identified more clinical biases, followed by cultural/racial biases, in both datasets. Adding self-reflection to *bias-identifier* reduced this disparity by a large margin, highlighting stigmatizing language more across both datasets, which more closely mirrored the human label distribution. Gender identity bias was detected the least frequently, which also mirrored the true label distribution (see Fig. 4). Overall, the proposed frameworks were considerably less likely to produce false positive bias identifications.

4.2 Quantitative Results

Since our goal is to extract bias topics at the bias-type level and evaluate how accurately these types are identified, we focus on computing precision, recall, and F1 score for class 1, representing correctly classified bias types at the input chunk level. As the analysis is carried out in the finalized debate topics across frameworks, the recall score would always be 1, and the precision score would be the same as the accuracy, as there are no true negatives and false negatives for debate topics. These metrics reflect the model's ability to detect the presence and type of bias. Frameworks like self-reflection, self-consistency, and reasoning may reject certain outputs (labeling them as "0"), but these do not correspond to ground-truth class 0 examples.[2] Quantitative results across datasets and frameworks are reported in Tables 1 and 2.

4.2.1 PMC-Patients Summary Dataset

The 409 debate topics extracted by *bias-identifier* include 42 correct debate topics out of 57 true potential debate topics in the overall 50 notes. The self-reflection framework reduced this number from 409 to 81 debate topics (preserving 19 true topics) (see Table. 1). Separately, the self-consistency framework reduced the total number of debate topics to 151 (a 63.08% reduction in debate topics) while maintaining an accuracy of 17%

[2]In fact, the baseline extraction is constructed such that nearly all examples contain at least one bias type. Therefore, the task is essentially one of detecting which bias type is present, not whether bias exists at all. As a result, we do not report metrics for class 0, since this would only indicate the absence of a specific bias type, rather than confirming the absence of all bias.

Framework	C	P	R	F_1	Acc. (%)	$\frac{n}{N}$
bias-identifier	1	0.1	1	0.19	10	$\frac{42}{409}$
+self-reflection	1	0.23	1	0.38	23	$\frac{19}{81}$
+reasoning	1	0.09	1	0.16	9	$\frac{9}{104}$
self-consistency	1	0.17	1	0.29	17	$\frac{26}{151}$
self-consistency +self-reflection	1	0.29	1	0.45	29	$\frac{7}{24}$

Table 1: Performance across frameworks for the PMC-Patients Summary dataset. *P*=precision, *R*=recall, *Acc.*=accuracy, and *n*=number of samples predicted for the specified class. *N*=Total number debate topics. *C*=1 represents correctly classified bias type. The *bias-identifier* framework is used as the base for all conditions (e.g., *+self-reflection* indicates *bias-identifier+self-reflection*).

Framework	C	P	R	F_1	Acc. (%)	$\frac{n}{N}$
bias-identifier	1	0.05	1	0.09	5	$\frac{19}{398}$
+self-reflection	1	0.15	1	0.26	15	$\frac{12}{81}$
+reasoning	1	0.04	1	0.08	4	$\frac{5}{124}$
self-consistency	1	0.09	1	0.17	9	$\frac{18}{196}$
self-consistency +self-reflection	1	0.22	1	0.37	22	$\frac{11}{49}$

Table 2: Performance across frameworks for the MIMIC-IV dataset. Specifications are similar to those indicated for Table 1.

in the generated debate topics.[3] Adding a layer of self-reflection on top of self-consistency pruned the number of debate topics further to 24 from 151, but 29% of the debate topics were actually true, highlighting the increased proficiency of a joint framework in identifying genuine potential biases.

4.2.2 MIMIC-IV Dataset

We observe a similar pattern in the MIMIC-IV dataset (see Table 2). Out of 398 debate topics identified by *bias-identifier*, only 19 were among the 64 potential true debate topics. However, when the 398 topics were passed to the self-reflection framework, they were reduced to 81 (a 79.64% reduction in debate topics) while preserving 63.17% of the true debate topics. Although the reasoning framework reduced the debate topics by 68.84%, it could only preserve 26.31% of the true debate topics, confirming the poorer performance also observed with the PMC-Patient Summary dataset at detecting bias in clinical notes. In contrast, the self-consistency framework reduced the topics to 196 (a 50.75% reduction) with only a 5.26% reduction in the number of identified true debate topics. Adding a self-reflection layer on top of the self-consistency framework resulted in a further 75% reduction in debate topics compared to self-consistency while preserving 61.11% of true debate topics generated by the self-consistency framework.

Considering all results across frameworks and datasets, we observed that in general the reasoning framework did not perform well, but the self-reflection and self-consistency frameworks outperformed *bias-identifier* by a wide margin. This shows that encouraging the model to reflect on its decision helps it produce more coherent and reliable outputs. Likewise, adding a self-consistency layer helps ensure that output is more stable and dependable, especially when the model is uncertain. Leveraging both frameworks jointly leads to more accurate and bias-aligned predictions.

4.2.3 Overall Bias Detection

Additionally, we compared the ability of the proposed frameworks to accurately capture the presence of bias at the document level. We consid-

Dataset	Framework	P	R	Acc. (%)
PMC-Patients	*bias-identifier*	0.30	1.00	31
	+self-reflection	0.39	0.80	57
	+reasoning	0.32	0.87	41
	self-consistency	0.45	1.00	48
	+self-consistency +self-reflection	0.67	0.48	68
MIMIC-IV	*bias-identifier*	0.38	1.00	38
	+self-reflection	0.36	0.83	38
	+reasoning	0.38	1.00	38
	self-consistency	0.50	1.00	50
	+self-consistency +self-reflection	0.67	1.00	75

Table 3: Document-level bias classification performance across frameworks on both datasets.

[3]The self-reasoning framework also reduced the debate topics by a large amount, but it filtered many true potential debate topics, suggesting inefficiency of the self-reasoning framework in debate topic identification. For instance, there were some instances where the quote generated for an assigned bias type did not belong to the input chunk. While self-reflection and self-consistency could easily filter such cases, the self-reasoning framework forced the model to iterate through multiple debate rounds and reach consensus in such cases, overcomplicating the process.

ered a document as "biased" if it contained at least one bias type; otherwise it was considered "non-biased." Table 3 shows that for the PMC-Patients Summary dataset, all proposed frameworks outperformed *bias-identifier* in terms of both accuracy and precision. Interestingly, for the MIMIC-IV dataset, we did not observe a noticeable change for the self-reflection and reasoning frameworks. This might be due to document length in the MIMIC-IV dataset (e.g., the presence of many chunks, or opportunities to capture bias, in the document might have naturally resulted in at least one instance of identified bias in many cases). However, we observe that when using the self-consistency framework with MIMIC-IV, its accuracy increases from 38% to 50%. When layering this with the self-reflection framework, accuracy further increases by 50% with a 76.32% increase in precision.

5 Discussion and Conclusion

Through the development of structured prompting techniques to implement self-reflection, self-consistency, and self-reasoning frameworks, our work sheds light on the use of these approaches to detect six distinct forms of bias in both MIMIC-IV and PMC-Patients Summary notes using open-source LLMs. We found that a baseline *bias-identifier* framework tended to over-predict bias with low precision and abundant false positives. Self-reflection improved upon this outcome by encouraging further consideration of the generated results. Self-consistency introduced stability through guaranteed constancy across multiple runs. Combined, self-reflection and self-consistency resulted in more accurate and reliable bias detection compared to *bias-identifier* across both PMC-Patients Summary notes and MIMIC-IV clinical notes.

Our analysis shows that simpler frameworks such as self-reflection and self-consistency can outperform more complex frameworks such as self-reasoning, especially when the reasoning chains are fragile or overly verbose. The fact that simpler prompting strategies outperformed the self-reasoning framework suggests that the added complexity may introduce noise or fragility in reasoning.

Overall, our results confirm that open-source LLMs, steered by structured prompting and agent roles, are well-positioned for use as a tool to detect bias in clinical text. Self-consistency followed by self-reflection outperformed alternative frameworks at the document level and facilitated bias-type level analysis. This provides a scalable, interpretable solution that integrates human guidance with machine intelligence. Our hope is that this work acts as a stepping stone to the development of responsible clinical NLP systems for implicit and explicit bias detection and auditing. It emphasizes the importance of meticulous prompt engineering, multi-agent alignment, and hybrid evaluation pipelines, suggesting that carefully structured prompting methods can improve internal validity through consistent model behavior.

While individual outputs from LLMs are inherently non-deterministic, our framework improves internal consistency by structuring the generative process. Although exact responses per framework may vary, our codebase and the design of these prompts allows the process to be reliably rerun and extended, supporting reproducible evaluation pipelines. External validity is limited by the use of a specific set of clinical notes and open-source LLMs, which may not capture the full variability of real-world healthcare documentation or institutional language norms. Further validation across diverse clinical settings, larger datasets, and different model families is needed to assess generalizability.

6 Future Works

As our approach demonstrates the utility of open-source LLMs to detect bias in clinical notes, it opens several directions for future exploration. As the results for identifying the type of bias were not that good, the first future direction would be to speculate on ways to improve the results for detecting the type of bias. We aim to scale this research to larger, more heterogeneous clinical datasets to evaluate performance in more realistic 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. Similarly, in the future we plan to involve medical professionals in the bias assessment loop to validate model predictions, refine labeling criteria, and increase trustworthiness. Finally, another intriguing direction lies in testing the extensibility of the approach to other types of medical text, such as radiology reports, discharge summaries, and nursing notes.

Limitations

The research was carried out on selected patient files. MIMIC-IV data analysis was restricted to 16 files and PMC-Patients Dataset was limited to 50 files, which might not be representative of the overall dataset. Even though human labeling was carried out based on established definitions of bias, human labeling of certain bias types, such as implicit bias or cultural bias, is inherently subjective. Although we did not conduct a formal inter-rater agreement study in this iteration, we recognize this as a limitation and an area for improvement in future work to further enhance reliability and transparency. This may introduce labeling inconsistencies due to human bias and variability. The whole experiment was carried out using one open-source LLM, `llama3.1-8b-instruct`, and output may differ for other models and scales.

Ethical Considerations

This study was reviewed by the Institutional Review Board at our institution and determined not to involve human subjects as defined by Department of Health & Human Services (DHHS) and/or U.S. Food and Drug Administration (FDA) regulations. While no direct patient interaction was involved, this research touches on ethically sensitive areas which includes bias in healthcare. Our research includes publicly available, de-identified clinical dataset (MIMIC-IV notes) and 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 large language model `llama3.1-8b-instruct` 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.

References

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

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. IEEE.

Donald U Apakama, Kim-Anh-Nhi Nguyen, Daphnee Hyppolite, Shelly Soffer, Aya Mudrik, Emilia Ling, Akini Moses, Ivanka Temnycky, Allison Glasser, Rebecca Anderson, and 1 others. 2024. Identifying and characterizing bias at scale in clinical notes using large language models. *medRxiv*, pages 2024–10.

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. 2024. 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.

Ivy Chen and et al. 2021. Stigmatizing language in clinical notes and its effects on patient care. *JAMA Network Open*, 4(7):e2116713.

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*.

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*, 18:1–18.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904.

R. Ma and 1 others. 2022. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *NPJ Digital Med.*, 5:187.

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.

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*, pages 2024–09.

Matthew Renze and Erhan Guven. 2024. Self-reflection in large language model agents: Effects on problem-solving performance. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 516–525.

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.

Mandy Sun and et al. 2022. Negative patient descriptors: documenting bias in the medical record. *Health Affairs*, 41(2):203–211.

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

Monica B Vela, Amarachi I Erondu, Nichole A Smith, Monica E Peek, James N Woodruff, and Marshall H Chin. 2022. Eliminating explicit and implicit biases in health care: evidence and research needs. *Annual review of public health*, 43(1):477–501.

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

Jesse O Wrenn, Donna M Stein, Suzanne Bakken, and Peter D Stetson. 2017. Identifying and characterizing high-volume, high-variability clinical note types in ehrs. *Journal of biomedical informatics*, 69:134–141.

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.

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 templates used to implement our proposed framework are provided below, including:

- **Prompt A.1:** The prompt used to extract bias type and quote for a particular debate topic.
- **Prompt A.2:** The prompt used to implement self-reflection framework.
- **Prompt A.3:** The player prompt used for debate.
- **Prompt A.4:** The moderator prompt used for initiating debate.
- **Prompt A.5:** The prompt used for affirmative agent.
- **Prompt A.6:** The prompt used for negative agent.
- **Prompt A.7:** The prompt used for moderator agent for providing decision.
- **Prompt A.8:** The prompt used for reflecting judge.
- **Prompt A.9:** The prompt used for deciding judge.

Prompt templates were refined iteratively using GPT-4 to assist with clarity and overall fluency, before a final manual refinement step.

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:

1. **Explicit Bias** – Overtly negative or judgmental language used in a medical note (e.g., labeling a patient as 'non-compliant' or 'irresponsible').
2. **Implicit Bias** – Subtle assumptions or stereotypes based on race, gender, socioeconomic status, etc., even when not explicitly stated (e.g., noting that a patient is 'well-spoken for their background').
3. **Clinical Bias** – Clinical bias refers to any form of prejudice or skewed judgment that influences medical decisions, interactions, or outcomes, often unintentionally.
4. **Stigmatizing Language** – Shame-inducing or blaming language, especially regarding addiction, obesity, or mental illness (e.g., 'drug abuser', 'refused care').
5. **Cultural/Racial Bias** – Unnecessary mention of race, ethnicity, or culture when it is not clinically relevant or is used in a stereotypical manner.
6. **Gender Identity Bias** – Misgendering or failure to respect a patient's gender identity, including dismissive or incorrect documentation.

Note: You are supposed to answer 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 same quote. Make sure your answer starts with only one '[' and ends with only one ']'. You can add as many unique bias type and quoate pair 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 ****: Overtly negative or judgmental language used in a medical note (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****: ##bias type##
**** The definition of ##bias type## is ****:
##definition##.
Answer:

Prompt A.3: Player prompt for debate

You are a medical bias detection expert participating in a multi-agent (two) debate as an agent. Your role is to analyze the medical note, quote, and assigned bias category given by some expert and justify your reasoning if its true. The debate topic is:
##debate_topic##
It's not necessary to fully agree with the other agent, as our goal is to reach a well-reasoned consensus.

Prompt A.4: 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.5: 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.6: 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.7: Prompt for Moderator Agent2

Now the ##round## round of debate has ended.
Affirmative side's answer:
##aff_ans##
Negative side's answer:
##neg_ans## As the moderator, analyze both responses and determine if there's a clear preference. If so, summarize your reason based on the experts conversation and provide the final agreed-upon answer. Do not put your own reasoning on the debate topic. Respond strictly in JSON:
`{"Bias Presence": "True or False","Consensus":"True or False", "Reason": ""}`.
You should provide result in a single JSON form. 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 json form.

Prompt A.8: Prompt for Judge Reflection

Affirmative side's answer: ##aff_ans##
Negative side's answer: ##neg_ans##
Write down key pointers by 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.9: 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 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.