Spurious Correlations and Beyond: Understanding and Mitigating Shortcut Learning in SDOH Extraction with Large Language Models

Anonymous ACL submission

Abstract

Social determinants of health (SDOH) extraction from clinical text is critical for downstream healthcare analytics. Although large language models (LLMs) have shown promise, they may rely on superficial cues leading to spurious predictions. Using the MIMIC portion of the SHAC (Social History Annotation Corpus) dataset and focusing on drug status extraction as a case study, we demonstrate that mentions of alcohol or smoking can falsely induce models to predict current/past drug use where none is present, while also uncovering concerning gender disparities in model performance. We further evaluate mitigation strategies—such as prompt engineering and chain-of-thought reasoning—to reduce these false positives, providing insights into enhancing LLM reliability in health domains.

1 Introduction

SDOH—including substance use, employment, and living conditions—strongly influence patient outcomes and clinical decision-making (Daniel et al., 2018; Himmelstein and Woolhandler, 2018; Armour et al., 2005). Extracting SDOH information from unstructured clinical text is increasingly important for enabling downstream healthcare applications and analysis (Jensen et al., 2012; Demner-Fushman et al., 2009). Although LLMs have shown promise in clinical natural language processing (NLP) tasks (Hu et al., 2024; Liu et al., 2023; Singhal et al., 2023), they often rely on superficial cues (Tang et al., 2023; Zhao et al., 2017), potentially leading to incorrect predictions undermining trust and utility in clinical settings.

Recent work has highlighted how LLMs can exhibit "shortcut learning" behaviors (Tu et al., 2020; Ribeiro et al., 2020; Zhao et al., 2018), where they exploit spurious patterns in training data rather than learning causal, generalizable features. This phenomenon spans various NLP tasks, from natural language inference (McCoy et al., 2019) to question-answering (Jia and Liang, 2017), and in clinical domains can lead to incorrect assumptions about patient conditions (Brown et al., 2023; Jabbour et al., 2020), threatening the utility of automated systems.

We investigate how LLMs produce spurious correlations in SDOH extraction through using drug status time classification (current, past, or none/unknown) as a case study. Using the MIMIC (Johnson et al., 2016) portion of the SHAC (Lybarger et al., 2021) dataset, we examine zero-shot and in-context learning scenarios across multiple LLMs (Llama (AI, 2024), Qwen (Yang et al., 2024), Llama3-Med42-70B (Christophe et al., 2024)).We explore multiple mitigation strategies to address these spurious correlations: examining the causal role of triggers through controlled removal experiments, implementing targeted prompt engineering approaches like chain-of-thought (CoT) reasoning (Wei et al., 2022), incorporating warning-based prompts, and augmenting with additional examples. While these interventions show promise—significant false positive rates persist, highlighting the deep-rooted nature of these biases and the need for more sophisticated solutions.

Contributions:

1. We present the first comprehensive analysis of spurious correlations in SDOH extraction across multiple LLM architectures, including domain-specialized models. Through extensive experiments in zero-shot and ICL settings, we demonstrate how models rely on superficial cues and verify their causal influence through controlled ablation studies.

2. We uncover systematic gender disparities in model performance, demonstrating another form of spurious correlation where models inappropriately leverage patient gender for drug

status time classification predictions.

3. We evaluate multiple prompt-based mitigation strategies (CoT, warnings, more examples) and analyze their limitations, demonstrating that while they reduce incorrect drug status time predictions, more robust solutions are needed for reliable clinical NLP deployments.

2 Methodology

2.1 Dataset and Task

We use the MIMIC-III portion of the SHAC dataset (Lybarger et al., 2021), which comprises 4405 deidentified social history note sections derived from MIMIC-III (Johnson et al., 2016) and the University of Washington clinical notes. SHAC is annotated using the BRAT tool (Stenetorp et al., 2012), capturing a variety of SDOH event types (e.g., Alcohol, Drug, Tobacco) as triggers along with associated arguments, including temporal status. To enable demographic analysis, we augmented the SHAC data by linking it with patient demographic information available in the original MIMIC-III dataset.

In this work, we examine spurious correlations in SDOH extraction through temporal drug status classification (current, past, or none/unknown). We adopt a two-step pipeline (Ma et al., 2022, 2023):

1. **Trigger Identification:** Given a social history note, the model identifies spans corresponding to the target event type (e.g., drug use).

2. **Argument Resolution:** For each identified trigger, the model applies a multiple-choice QA prompt to determine the temporal status (current/past/none). See Appendix C for detailed examples of the task and annotation schema.

2.2 Experimental Setup

Model Configurations We evaluate multiple model configurations:

- **Zero-Shot:** Models receive only the task instructions and input text, with no examples.

- **In-Context Learning (ICL):** Models are provided three example demonstrations before making predictions on a new instance. Examples are selected to maintain balanced representation across substance use patterns (none/single/multiple) and drug use outcomes (positive/negative).

- **Fine-Tuning (SFT):** We also fine-tune a Llama-3.1-8B model on the MIMIC portion of the SHAC dataset to assess whether domain adaptation reduces spurious correlations.

See Appendix B for more details on prompting strategies.

We consider Llama-3.1-70B (zero-shot, ICL), Llama-3.1-8B (fine-tuned on MIMIC), Qwen-72B (ICL), and Llama3-Med42-70B (ICL). These models span various parameter sizes and domain specializations. The fine-tuned Llama-8B model provides insights into whether in-domain adaptation mitigates the observed shortcut learning.

Evaluation Framework Our primary evaluation metric is the false positive rate (FPR), defined as: $FPR = FP/(FP + TN)$ where FP represents false positives (predicted current/past use when ground truth was none/unknown) and TN represents true negatives (correctly predicted none/unknown). We prioritize FPR given the clinical risks of incorrect positive drug use predictions. A higher FPR indicates more frequent erroneous predictions that could lead to patient stigmatization. See appendix D for detailed discussion.

To analyze potential spurious correlations, we categorize notes based on their ground truth substance use status:

- **Substance-positive**: Notes documenting current/past use of the respective substance (alcohol or smoking)

- **Substance-negative**: Notes where the ground truth indicates no use or unknown status

Experimental Settings

- **Original:** Evaluate models on the original notes.

- **Without Alcohol/Smoking Triggers:** Remove mentions of alcohol/smoking to test their causal role in inducing false positives.

3 Results

3.1 RQ1: Do Large Language Models Exhibit Spurious Correlations in SDOH Extraction?

As shown in Table 1, our analysis in a zero-shot setting with Llama-70B reveals high false positive rates for drug status time classification in alcohol-positive (66.21%) and smoking-positive (61.11%)

Table 1: False Positive Rates (%) Across Different Models and Approaches. *Smoking+Alcohol* refers to cases where both *Smoking-positive* and *Alcohol-positive* are true.

Cases	Llama-70B					Llama-8B		Llama3-Med42-70B	Qwen-72B
	Zero-shot	ICL	CoT	Warning	Increased-Examples	Vanilla	Fine-tuned	ICL	ICL
Alcohol-positive	66.21	48.28	33.79	40.69	45.52	73.10	32.41	66.90	62.76
Smoking-positive	61.11	36.42	25.93	29.63	30.25	74.07	36.42	57.41	53.09
Alcohol-negative	28.83	11.71	6.76	5.41	10.81	37.39	12.16	16.22	46.85
Smoking-negative	29.76	18.05	10.73	11.22	20.00	33.66	7.32	19.51	53.17
Smoking+Alcohol	73.26	51.16	34.88	45.35	39.53	81.40	40.70	76.74	56.98

notes. In contrast, alcohol-negative and smoking-negative notes show substantially lower false positive rates (28.83% and 29.76%, respectively). This stark contrast suggests that the mere presence of alcohol or smoking triggers biases the model towards inferring nonexistent drug use. These biases likely stem from the pre-training phase, potentially reinforcing societal assumptions about correlations between different types of substance use.

3.2 RQ2: Do In-Context Learning and Fine-Tuning Reduce These Spurious Correlations?

Providing three in-context examples reduces false positives significantly. For Llama-70B, ICL lowers alcohol-positive mismatches from 66.21% to 48.28%, though a gap remains relative to alcohol-negative notes (11.71%). Similarly, smoking-positive mismatches decrease from 61.11% to 36.42% versus 18.05% for smoking-negative. The effectiveness of ICL suggests that explicit examples help the model focus on relevant features, though the persistence of some bias indicates deep-rooted associations from pre-training. Fine-tuning Llama-8B on the MIMIC subset (SFT) yields further improvements: alcohol-positive mismatches drop to 32.41% and smoking-positive to 36.42%, with corresponding negatives at 12% and 7% respectively, indicating that domain adaptation helps override some pre-trained biases.

3.3 RQ3: Are These Superficial Mentions Causally Driving the Model's Predictions?

To confirm the causal role of alcohol and smoking mentions, we remove these triggers from the notes. Across models, this consistently lowers false positives. For instance, Llama-70B zero-shot sees alcohol-positive mismatches fall from 66.21% to 55.17% after removing alcohol triggers. Similarly, Llama-8B-SFT reduces alcohol-positive errors from 32.41% to 26.9%. Similar trends are observed across other architectures including domain-specific models (see appendix G), confirming that alcohol and smoking cues spuriously bias the models' drug-use predictions.

3.4 RQ4: Are there systematic demographic variations in these spurious correlations?

Beyond substance-related triggers, our analysis (Table 2) uncovers another concerning form of spurious correlation: systematic performance differences based on patient gender. Just as models incorrectly rely on mere mentions of alcohol or smoking to infer substance use, they appear to leverage patient gender as an inappropriate predictive signal. For the base Llama-70B model in zero-shot settings, false positive rates show stark gender disparities - male patients consistently face higher misclassification rates compared to female patients (71.15% vs 53.66% for alcohol-positive cases, and 66.67% vs 50.88% for smoking-positive cases). This pattern persists with in-context learning, with the gender gap remaining substantial (alcohol-positive: 52.88% male vs 36.59% female). Fine-tuned models showed similar disparities, with Llama-8B-SFT maintaining a performance gap of approximately 15 percentage points between genders for alcohol-positive cases.

Notably, these gender-based differences exhibit complex interactions with substance-related triggers. Cases involving positive substances mentions show the most pronounced disparities, with male patients seeing up to 20 percentage point higher false positive rates. This suggests that the model's shortcut learning compounds across different dimensions - gender biases amplify substance-related biases and vice versa. The persistence of these interacting biases across model architectures, sizes, and prompting strategies suggests they arise from deeply embedded patterns in both pre-training data and medical documentation practices.

Table 2: Gender-Based Analysis of False Positive Rates (%) Across Models

	Llama-70B Zero-shot		Llama-70B ICL		Llama-8B SFT		Qwen-72B	
Cases	Female	Male	Female	Male	Female	Male	Female	Male
Alcohol-positive	53.66	71.15	36.59	52.88	21.95	36.54	68.29	60.58
Smoking-positive	50.88	66.67	28.07	40.95	24.56	42.86	49.12	55.24
Alcohol-negative	29.13	28.42	9.45	14.74	9.45	15.79	47.24	46.32
Smoking-negative	27.03	32.98	9.91	27.66	6.31	8.51	54.05	52.13
Smoking+Alcohol	81.82	84.62	54.55	58.97	27.27	53.85	27.27	30.77

4 Mitigation Strategies and Results

We explore several mitigation techniques to address the spurious correlations identified in our analysis:

Chain-of-Thought (CoT) As shown in Table 1, instructing the model to reason step-by-step before producing an answer leads to substantial reductions across all architectures. For Llama-70B, CoT reduces alcohol-positive mismatches from 66.21% (zero-shot) to 33.79%, with smoking-positive cases decreasing from 61.11% to 25.93%. Similar improvements are observed in other models (see appendix H), with Qwen-72B showing particularly strong response to CoT. This suggests CoT helps models avoid superficial cues and focus on explicit information.

Warning-Based Instructions We prepend explicit instructions cautioning the model not to assume drug use without evidence and to treat each factor independently. With Llama-70B, these warnings lower alcohol-positive mismatches from 66.21% to approximately 40.69%, and also benefit smoking-positive scenarios. While not as strong as CoT, these warnings yield meaningful improvements across different architectures.

Increased Number of Examples Providing more than three examples—up to eight—further stabilizes predictions. For Llama-70B, increasing the number of examples reduces false positive rates considerably, with alcohol-positive mismatches falling to 45.52% (compared to 66.21% zero-shot). Similar trends are observed in other models, though the magnitude of improvement varies (see appendix H). While not as dramatic as CoT, additional examples help guide models away from faulty heuristics.

5 Discussion

Our findings highlight a key challenge in applying large language models to clinical information extraction: even when models achieve strong performance on average, they rely on superficial cues rather than genuine understanding of the underlying concepts. The presence of alcohol- or smoking-related mentions biases models to infer drug use incorrectly, and these shortcuts persist across Llama variants, Qwen, and Llama3-Med42-70B. The effectiveness of mitigation strategies like chain-of-thought reasoning, warning-based instructions, and additional examples underscores the importance of careful prompt design. While these interventions help guide models to focus on explicit evidence, their partial success suggests the need for more robust approaches - integrating domain-specific knowledge, implementing adversarial training, or curating more balanced datasets. Our demographic analysis reveals that these spurious correlations are not uniformly distributed across patient groups, raising fairness concerns for clinical deployment. Addressing such disparities requires both algorithmic improvements and careful consideration of deployment strategies. Clinicians and stakeholders must be aware of these limitations before deploying LLMs in clinical decision-support systems. Understanding these systematic biases in automated analysis can inform improvements not only in model development but also in clinical documentation practices and standards (see appendix F).

6 Conclusion

This work presents the first systematic exploration of spurious correlations in SDOH extraction, revealing how contextual cues can lead to incorrect and potentially harmful predictions in clinical settings. Beyond demonstrating the problem, we've evaluated several mitigation approaches that, while promising, indicate the need for more sophisticated solutions. Future work should focus on developing robust debiasing techniques, leveraging domain expertise, and establishing comprehensive evaluation frameworks to ensure reliable deployment across diverse populations.

7 Limitations

Dataset limitations Our analysis relied exclusively on the MIMIC portion of the SHAC dataset, which constrains the generalizability of our findings. While we observe consistent gender-based performance disparities, a more diverse dataset could help establish the breadth of these biases.

Model coverage We focused solely on open-source large language models (e.g., Llama, Qwen). Extending the evaluation to additional data sources, closed-source models, and other domain-specific architectures would help verify the robustness of our conclusions.

Causal understanding While we established the causality of triggers through removal experiments, understanding why specific triggers affect certain models or scenarios would require deeper analysis using model interpretability techniques.

Methodology scope Our study focused exclusively on generative methods; results may not generalize to traditional pipeline-based approaches that combine sequence labeling and relation classification.

Mitigation effectiveness While we identified various spurious correlations, our mitigation strategies could not completely address the problem, leaving room for future work on addressing these issues.

8 Ethics Statement

All experiments used de-identified social history data from the SHAC corpus, with LLMs deployed on a secure university server. We followed all data use agreements and institutional IRB protocols. Although the dataset is fully de-identified, biases within the models could raise ethical concerns in real-world applications. Further validation and safeguards are recommended before clinical deployment.

9 Acknowledgments

We thank our collaborators for their valuable feedback and support. Generative AI assistants were used for grammar checking and LaTeX formatting; the authors retain full responsibility for the final content and analysis.

References

Meta AI. 2024. Llama 3.1 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md. Accessed: 2024-12-13.

BS Armour, T Woollery, A Malarcher, TF Pechacek, and C Husten. 2005. Annual smoking-attributable mortality, years of potential life lost, and productivity losses—united states, 1997-2001. *JAMA: Journal of the American Medical Association*, 294(7).

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. 2023. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature communications*, 14(1):4314.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *Preprint*, arXiv:2408.06142.

Rachel A Dahl, J Priyanka Vakkalanka, Karisa K Harland, and Joshua Radke. 2022. Investigating healthcare provider bias toward patients who use drugs using a survey-based implicit association test: Pilot study. *Journal of addiction medicine*, 16(5):557–562.

Hilary Daniel, Sue S Bornstein, Gregory C Kane, Health, and Public Policy Committee of the American College of Physicians*. 2018. Addressing social determinants to improve patient care and promote health equity: an american college of physicians position paper. *Annals of internal medicine*, 168(8):577–578.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Sifei Han, Robert F Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of biomedical informatics*, 127:103984.

Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, Elyse Lasser, Felicia Hill-Briggs, Jill Marsteller, Hadi Kharrazi, et al. 2019. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR medical informatics*, 7(3):e13802.

David U Himmelstein and Steffie Woolhandler. 2018. Determined action needed on social determinants. *Annals of internal medicine*, 168(8):596–597.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Zalaya K Ivy, Sharon Hwee, Brittany C Kimball, Michael D Evans, Nicholas Marka, Catherine Bendel, and Alexander A Boucher. 2024. Disparities in documentation: evidence of race-based biases in the electronic medical record. *Journal of Racial and Ethnic Health Disparities*, pages 1–7.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W Sjoding, and Jenna Wiens. 2020. Deep learning applied to chest x-rays: exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR.

Peter B Jensen, Lars J Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Min Kyung Kim, Joy Noel Baumgartner, Jennifer Headley, Julius Kirya, James Kaggwa, and Joseph R Egger. 2021. Medical record bias in documentation of obstetric and neonatal clinical quality of care indicators in uganda. *Journal of Clinical Epidemiology*, 136:10–19.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 30(8):1389–1397.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631.

Mingyu Derek Ma, Alexander K Taylor, Wei Wang, and Nanyun Peng. 2022. Dice: data-efficient clinical event extraction with generative models. *arXiv preprint arXiv:2208.07989*.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

David M Markowitz. 2022. Gender and ethnicity bias in medicine: A text analysis of 1.8 million critical care records. *PNAS nexus*, 1(4):pgac157.

CM Mateo and DR Williams. Addressing bias and reducing discrimination. *The professional responsibility of health care providers*, 2020:95.

RT McCoy, E Pavlick, and T Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arxiv preprint arxiv: 190201007.

SA Meyers, VA Earnshaw, Brittany D'Ambrosio, Natasia Courchesne, Dan Werb, and LR Smith. 2021. The intersection of gender and drug use-related stigma: A mixed methods systematic review and synthesis of the literature. *Drug and alcohol dependence*, 223:108706.

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, et al. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12):2716–2727.

Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nicholas J Dobbins, Özlem Uzuner, and Meliha Yetisgen. 2023. Prompt-based extraction of social determinants of health using few-shot learning. *Preprint*, arXiv:2306.07170.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Brendan Saloner, Wenshu Li, Michael Flores, Ana M Progovac, and Benjamin Lê Cook. 2023. A widening divide: Cigarette smoking trends among people with substance use disorder and criminal legal involvement: Study examines cigarette smoking trends among people with substance use disorders and people with criminal legal involvement. *Health Affairs*, 42(2):187–196.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. 2021. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA open*, 4(3):ooaa069.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. *arXiv preprint arXiv:2305.17256*.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Leonieke C Van Boekel, Evelien PM Brouwers, Jaap Van Weeghel, and Henk FL Garretsen. 2013. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. *Drug and alcohol dependence*, 131(1-2):23–35.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zehao Yu, Xi Yang, Chong Dang, Songzi Wu, Prakash Adekkanattu, Jyotishman Pathak, Thomas J George, William R Hogan, Yi Guo, Jiang Bian, et al. 2022. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. In *AMIA Annual Symposium Proceedings*, volume 2021, page 1225.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

A Related Work

Previous work on extracting SDOH from clinical text spans a progression from rule-based methods to fine-tuned neural models, leveraging annotated corpora for tasks like substance use and employment status extraction (Hatef et al., 2019; Patra et al., 2021; Yu et al., 2022; Han et al., 2022; Uzuner et al., 2008; Stemerman et al., 2021; Lybarger et al., 2023). More recent efforts have explored prompt-based approaches with LLMs, including GPT-4, to reduce reliance on extensive annotations (Ramachandran et al., 2023). While these approaches achieve competitive performance, studies across NLP tasks have shown that both fine-tuned and prompting-based methods often exploit spurious correlations or superficial cues (Ribeiro et al., 2020; Geirhos et al., 2020; Tu et al., 2020). Prior investigations have focused largely on spurious correlations in standard NLP tasks and supervised scenarios (McCoy et al., 2019; Zhao et al., 2018). In contrast, our work examines how these issues manifest in zero-shot and in-context SDOH extraction settings, and we propose prompt-level strategies to mitigate these correlations.

B Prompting Strategies

All prompting approaches share a base system message identifying the model's role as "an AI assistant specialized in extracting and analyzing social history information from medical notes." Each strategy then builds upon this foundation with specific modifications:

Zero-Shot

The baseline approach uses a minimal prompt structure: System: AI assistant specialized in social history extraction User: For the following social history note: [Clinical note text] [Task instruction] [Options if applicable] This setup evaluates the model's ability to perform extraction tasks using only its pre-trained knowledge, without additional guidance or examples.

In-Context Learning (ICL)

This approach augments the base prompt with three carefully selected demonstration examples. Each example follows a structured JSON format: json "id": "example-id", "instruction": "Extract all Drug text spans...", "input": "Social History: Patient denies drug use...", "options": "[Multiple choice options if applicable]", "output": "Expected extraction or classification"

Chain-of-Thought (CoT)

Building upon ICL, this method explicitly guides the model through a structured reasoning process: Please approach this task step-by-step: 1. Carefully read the social history note 2. Identify all relevant information related to the question 3. Consider the examples provided 4. Explain your reasoning process 5. Provide your final answer This approach aims to reduce spurious correlations and shortcut learning by encouraging explicit articulation of the reasoning process before arriving at the final extraction or classification.

Warning-Based

This specialized approach incorporates explicit rules and warnings in the system message: Important Guidelines: 1. Evaluate each factor independently - never assume one behavior implies another 2. Extract only explicitly stated information - don't make assumptions based on demographics or other factors 3. If information isn't mentioned, use [none] or select "not mentioned" option These guidelines specifically address the challenge of false positives in substance use detection by discouraging inference-based conclusions without explicit textual evidence. The warnings are designed to counteract the model's tendency to make assumptions based on superficial cues or demographic factors.

C Dataset Details

C.1 Data Format and Annotation Process

The SHAC dataset originally consists of paired text files (.txt) containing social history notes and annotation files (.ann) capturing SDOH information. We convert these into a question-answering format to evaluate LLMs. Below we demonstrate this process with a synthetic example:

Raw Note (.txt)

```
SOCIAL HISTORY:
Patient occasionally uses alcohol.
Denies any illicit drug use.
```

BRAT Annotations (.ann)

```
T1 Alcohol 24 31 alcohol
T2 Drug 47 50 drug
T3 StatusTime 8 19 occasionally
T4 StatusTime 32 37 denies

E1 Alcohol:T1 Status:T3
E2 Drug:T2 Status:T4

A1 StatusTimeVal T3 current
A2 StatusTimeVal T4 none
```

Here, `T1` and `T2` are triggers - spans of text that indicate the presence of SDOH events (e.g., "alcohol" for substance use). The annotations also capture arguments - additional information about these events, such as their temporal status represented by `T3` and `T4`. For example, `T3` ("occasionally") indicates a temporal status of *current* for alcohol use.

We transform these structured annotations into two types of questions:

Trigger Identification Questions about identifying relevant event spans:

```
{"id": "0001-Alcohol",
 "instruction": "Extract all Alcohol
  text spans as it is from the note.
  If multiple spans present, separate
  them by [SEP]. If none, output
  [none].",
 "input": "SOCIAL HISTORY: Patient
  occasionally uses alcohol. Denies
  any illicit drug use.",
 "output": "alcohol"}
```

Argument-Resolution Questions about determining event properties:

```
{"id": "0001-Alcohol_StatusTime",
 "instruction": "Choose the best
  StatusTime value for the <alcohol>
  (Alcohol) from the note:",
 "input": "SOCIAL HISTORY: Patient
  occasionally uses alcohol. Denies
  any illicit drug use.",
 "options": "Options: (a) none.
  (b) current. (c) past.
  (d) Not Applicable.",
 "output": "(b) current."}
```

D Metric Selection and Justification

Our focus on False Positive Rate (FPR) is motivated by the unique risks associated with incorrect substance use predictions in clinical settings (Van Boekel et al., 2013; Dahl et al., 2022). While traditional metrics like accuracy or F1-score treat all errors equally, FPR specifically captures the rate of unwarranted "positive" classifications—a critical concern when dealing with sensitive patient information. High FPR values indicate that models frequently make unjustified drug use predictions, which could lead to:

- Patient stigmatization and potential discrimination
- Reduced quality of care due to biased provider perceptions
- Diminished trust in automated clinical decision support systems

Conversely, lower FPR values suggest better model reliability in avoiding these harmful misclassifications. While comprehensive evaluation would benefit from additional metrics, FPR serves as a particularly relevant indicator for assessing model safety and reliability in clinical applications.

E Model Fine-tuning and Computational Resources

We fine-tuned Llama-8B using LoRA with rank 64 and dropout 0.1. Key training parameters include a learning rate of 2e-4, batch size of 4, and 5 training epochs. Training was conducted on 2 NVIDIA A100 GPUs for approximately 3 hours using mixed precision (FP16). For our main experiments, we used several large language models: Llama-70B (70B parameters), Qwen-72B (72B parameters), Llama3-Med42-70B (70B parameters), and our fine-tuned Llama-8B (8B parameters). The inference experiments across all models required approximately 100 GPU hours on 2 NVIDIA A100 GPUs. This computational budget covered all experimental settings including zero-shot, in-context learning, and the evaluation of various mitigation strategies.

F Implications Beyond NLP: Clinical Documentation and Practice

The implications of this study extend beyond NLP methodologies. Our analysis reveals that these models not only learn but potentially amplify existing biases in clinical practice. The identified error patterns—particularly the tendency to infer substance use from smoking/alcohol mentions and gender-based performance disparities—mirror documented provider biases in clinical settings (Saloner et al., 2023; Meyers et al., 2021). Notably, these biases appear to originate partly from medical documentation practices themselves (Ivy et al., 2024; Kim et al., 2021; Markowitz, 2022). Our finding that explicit evidence-based reasoning (through CoT) reduces these biases aligns with established strategies for mitigating provider bias (Mateo and Williams). This parallel between computational and human biases suggests that systematic analysis of LLM behavior could inform broader efforts to identify and address biases in medical documentation and practice, potentially contributing to improved provider education and documentation standards.

G Trigger Removal Experiments

Table 3: Impact of Trigger Removal on Llama 3.1 Models False Positive Rates (%)

	Llama 3.1 70b Zero-shot			Llama 3.1 8b SFT		
Cases	Full	Without Alcohol	Without Smoking	Full	Without Alcohol	Without Smoking
Alcohol-positive	66.21	55.17	64.14	32.41	26.90	33.10
Smoking-positive	61.11	54.94	56.79	36.42	32.10	31.48
Alcohol-negative	28.83	25.23	23.87	12.16	12.16	8.11
Smoking-negative	29.76	22.93	26.34	7.32	6.83	7.32
Smoking+Alcohol	73.26	65.12	72.09	40.70	32.56	41.86

Table 4: Impact of Trigger Removal on Additional Models' False Positive Rates (%)

	Llama 3.1 70B ICL			Llama3-Med42-70B			Qwen-72B		
Cases	Full	Without Alcohol	Without Smoking	Full	Without Alcohol	Without Smoking	Full	Without Alcohol	Without Smoking
Alcohol-positive	48.28	38.62	47.59	66.90	53.10	64.83	62.76	51.72	54.48
Smoking-positive	36.42	32.72	32.09	57.41	51.85	52.47	53.09	45.68	51.23
Alcohol-negative	11.71	16.22	10.81	16.22	16.22	13.96	46.85	45.05	47.75
Smoking-negative	18.05	14.15	15.12	19.51	14.15	19.51	53.17	49.27	49.76
Smoking+Alcohol	51.16	44.19	46.51	76.74	66.28	73.26	56.98	43.02	50.00

H Mitigation Experiments

Table 5: Impact of Mitigation Strategies on Additional Models' False Positive Rates (%)

	Llama3-Med42-70B				Qwen-72B			
Cases	ICL	CoT	Warning	Increased Examples	ICL	CoT	Warning	Increased Examples
Alcohol-positive	66.90	48.28	62.76	63.45	62.76	28.97	34.38	36.55
Smoking-positive	57.41	35.19	53.09	50.62	53.09	23.46	32.09	33.33
Alcohol-negative	16.22	6.76	16.67	15.76	46.85	19.82	22.07	26.12
Smoking-negative	19.51	13.66	18.54	18.05	53.17	17.07	25.85	29.27
Smoking+Alcohol	76.74	53.49	72.09	68.60	56.98	32.56	37.21	41.86