Analyzing the Capabilities of Large Language Models in Annotating Substance Use Behavior from Clinical Notes

Anonymous ACL submission

Abstract

Large language models (LLMs) have been trialed to annotate complex medical information. In this paper, we explore the capabilities of LLMs in annotating patient substance use behavior from clinical notes. We used MIMIC-SBDH data, which is based on MIMIC-3 discharge summaries, and annotated alcohol use, tobacco use, and drug use behavior into five instances(labels): *Past*, Present, *Never*, *Unsure*, and *nan*, using the Llama3 model. The model achieved high match scores for the *Past* category annotation, ranging from 83.26% to 90.62%. Overall, the model accurately predicted alcohol, drug, and tobacco behaviors with respective overall accuracies of 51.70%, 31.37%, and 72.62%. However, the model performed poorly in annotating the *Unsure* category, with match scores ranging from 2.25% to 3.47%. Our experimentation provides information regarding performance patterns and challenges with use of LLMs for annotating complex healthcare data.

1 Introduction

Substance abuse (alcohol, tobacco, drugs) is associated with multifaceted impacts on human health (McLellan, 2017; Lo et al., 2020; Amaro et al., 2021). According to the 2022 National Survey on Drug Use and Health (NSDUH) [1], 48.7 million people aged 12 or older (17.3%) had a Substance Use Disorder (SUD) in the past year. This staggering figure includes 29.5 million individuals with an Alcohol Use Disorder (AUD), 27.2 million with a Drug Use Disorder (DUD), and 8.0 million people with both an AUD and a DUD. Substance use affects not only adults but also younger populations. The NSDUH survey indicates that 7.3% of adolescents aged 12 to 17, approximately 1.9 million, used tobacco products or vaped nicotine in the past month.

[1]https://www.samhsa.gov/data/release/2022-national-survey-drug-use-and-health-nsduh-releases

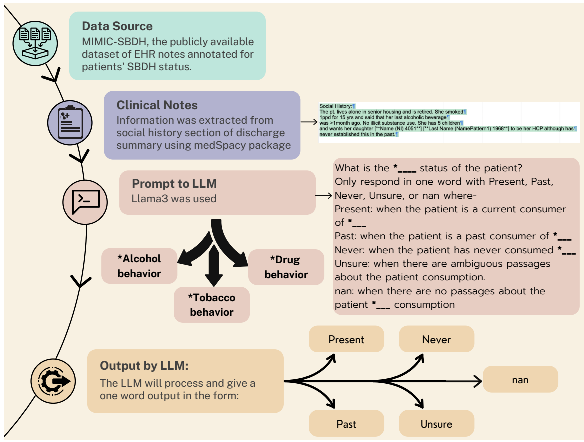

Figure 1: Process map of our study for annotating substance use behaviour, using Llama3 as the large language model.

Beyond individual health, the economic burden of SUD is significant; one study estimated the annual medical cost associated with SUD in US emergency departments and inpatient settings exceeded \$13 billion (Li et al., 2023). Substance use behavior is also strongly associated with developing chronic diseases (Wu et al., 2018), cardiovascular complications (Nishimura et al., 2020; Snow et al., 2019; Keloth et al., 2024), and cancer (Rumgay et al., 2021; Jayadevappa and Chhatre, 2016; Yusufov et al., 2019), highlighting the critical importance of accurate information on substance use for patient care. The digitization of clinical records presents a new opportunity to integrate information on indicators such as substance use into Electronic Health Records (EHRs) (Tai and McLellan, 2012; Chen et al., 2020; Frimpong et al., 2023). EHRs contain patients' demographics, medical history, social history, vital signs, laboratory tests, and medication orders. Information about substance use is typically included in the social history section of clinical notes. Manually extracting information from clinical notes is challenging and burdensome due to their richness in information and considerable length (Moy et al., 2021; Walsh, 2004). Recent advances in natural language processing algorithms, including the success of large language models (LLMs), offer hope in addressing this challenge (Denecke et al., 2024; Yang et al., 2023). Research shows that LLMs can effectively extract information from clinical data, including identifying social determinants of health (SDOH) such as employment, housing, transportation, relationships, and social support, achieving high performance across various tasks (Guevara et al., 2024; Ralevski et al., 2024; Keloth et al., 2024; Singhal et al., 2023).

Previous studies have leveraged LLMs to assess the severity of SUD through the analysis of clinical notes (Mahbub et al., 2024). One research has applied classical natural language processing approaches to annotate elements like the amount and frequency of substance use in clinical notes (Ganoe et al., 2021). Despite ongoing efforts, there remains limited research on deriving complex patterns of substance use behavior. In this study, we utilized patient clinical notes to evaluate LLMs for annotating substance use behavior patterns across different annotation instances *Present*, *Past* , *Never*, *Unsure*, *nan*, comparing their performance with human annotation. We provided instance-wise performance metrics of the LLMs, offering a detailed analysis of their effectiveness in handling specific types of information. This approach not only highlights the strengths and weaknesses of LLMs in this context but also emphasizes the need for comprehensive evaluations strategies.

2 Methods

2.1 Datasets and Model

We used MIMIC-SBDH (Ahsan et al., 2021), the publicly available dataset of EHR notes annotated for patients' SBDH (social and behavioral determinants of health) status. This dataset was generated using 7,025 discharge summaries randomly selected from the MIMIC-III (Johnson et al., 2016) dataset for the following SBDHs: community, economics, education, environment, alcohol use, tobacco use, and drug use. For our analysis, we selected substance use behavior determinants like alcohol use, tobacco use, and drug use. All this information was extracted from the Social History section of the discharge summaries using the Medspacy package (Eyre et al., 2021) to extract the social history section from the discharge summaries. For our analysis, we used the 8 billion parameter model from the Llama3 [2] model family, 'meta-llama/Meta-Llama-3-8B-Instruct', available on Hugging Face. We obtained access to the model by agreeing to the 'META LLAMA 3 COMMUNITY LICENSE AGREEMENT'.

2.2 Prompt strategy

We built the zero shot prompts to annotate Alcohol, Tobacco, and Drug behavior use in order to generate model outputs into 5 labels: *Present*, *Past*, *Never*, *Unsure*, and *nan*. The explanation of all the labels is provided in **Table 1 & Appendix A**. In the prompt, we included the explanation of all the labels verbatim from the original MIMIC-SBDH (Ahsan et al., 2021) paper to avoid any generation bias.

2.3 Labels generation and evaluation

From the HuggingFace "meta-llama/Meta-Llama-3-8B-Instruct" model, we generated labels for all 7025 discharge summaries by setting the temperature hyperparameter of the model to 0.6 (within the range of 0 to 1) to obtain more deterministic output from the model (Peeperkorn et al., 2024). We have also set the top-p hyperparameter value to 0.9, so that the LLM will only generate words that have a probability of at least 0.9. We have performed our experiments in Google Colab with advance subscription and used A100 gpu for our experiments. After obtaining the generated labels from the model, we compared them with the original human-annotated labels and calculated the matching scores for all three scenarios: alcohol use, tobacco use, and drug use. The match score (MS) measures the alignment between the original labels and the generated values, quantifying the proportion of cases where the generated labels match the original ones.

$$MS_{i,j} = \frac{Correct\ Generated\ Labels(N)_{i,j}}{Actual\ Labels(M)_j} \tag{1}$$

Where:

- i varies for alcohol, drug, and tobacco categories.
- j varies for 'never', '*Present*', '*nan*', '*Unsure*', and '*nan*' categories.

[2] https://llama.meta.com/llama3/

Behavior	Labels	Explanation	Prompt
Alcohol use	***Present***	Patient is a current consumer of alcohol.	What is the alcohol consumption status of the patient? Only respond in one word with Present, Past, Never, Unsure, or nan where : Present: when the patient is a current consumer of alcohol. Past: when the patient is a past consumer of alcohol. Never: when the patient has never consumed alcohol. Unsure: when there are ambiguous passages about the patient consumption. nan: when there are no passages about the patient alcohol consumption.
	Past	Past consumer of alcohol	
	Never	Has never consumed alcohol .	
	Unsure	Ambiguous passages about patient's consumption.	
	nan	No passages about the patient's alcohol consumption.	

Table 1: Target labels, explanations for alcohol use and corresponding prompts.

3 Results

The social history section of all 7025 discharge summaries has an average word count of 30.02 words with a standard deviation of 23.01. The average processing time for one social history was 0.10 seconds (SD=0.02) for alcohol, 0.10 seconds (SD=0.13) for tobacco, and 0.09 seconds (SD=0.03) for drug use. Since the prompt was written in a way that the generated output should consist of 1 word with options: Never, *Present*, *nan*, *Unsure*, or *nan*, however, we found some outputs other than these 5 labels. In the case of Alcohol, there were 255 instances (3.62%), for tobacco 113 instances (1.60%), and for drug use 41 instances (0.58%). We referred to all those outputs as 'Random'.

3.1 Model performance for generating alcohol labels

In the case of alcohol behavior annotation, we found that the overall model correctly generated 3632 (51.70%) labels **Figure 2A**. After analyzing the match score of all labels, we found the maximum match score for the *Past* class to be 88.74%, and the minimum match score for the *Unsure* class at 2.71%. Additionally, match scores were observed for the *nan* class at 37.84%, the *Present* class at 56.91%, and the Never class at 55.52% **Figure 2D**.

3.2 Model accuracy for generating tobacco labels

In the case of tobacco behavior annotation, we found that the overall model correctly generated 5102 (72.62%) labels **Figure 2B**. After analyzing the match score of all labels, we found the maximum match score for the *Past* class to be 90.62%, and the minimum match score for the *Unsure* class at 2.25%. Additionally, match scores were observed for the *nan* class at 33.77%, the *Present* class at 74.75%, and the Never class at 88.10% **Figure 2E**.

3.3 Model performance for generating drug labels

In the case of drug behavior annotation, we found that the overall model correctly generated 2204 (31.37%) labels **Figure 2C**. After analyzing the match score of all labels, we found the maximum match score for the *Past* class to be 83.26%, and the minimum match score for the *Unsure* class at 3.47%. Additionally, match scores were observed for the *nan* class at 16.61%, the *Present* class at 53.62%, and the Never class at 55.57% **Figure 2F**.

4 Limitations

Our study is subject to certain limitations that warrant consideration. Firstly, we have performed analysis on only one data source, and our findings need to be confirmed with other data sources. Secondly, we are presenting results only using the Llama3 model. The reason is that Llama3 is an open-source

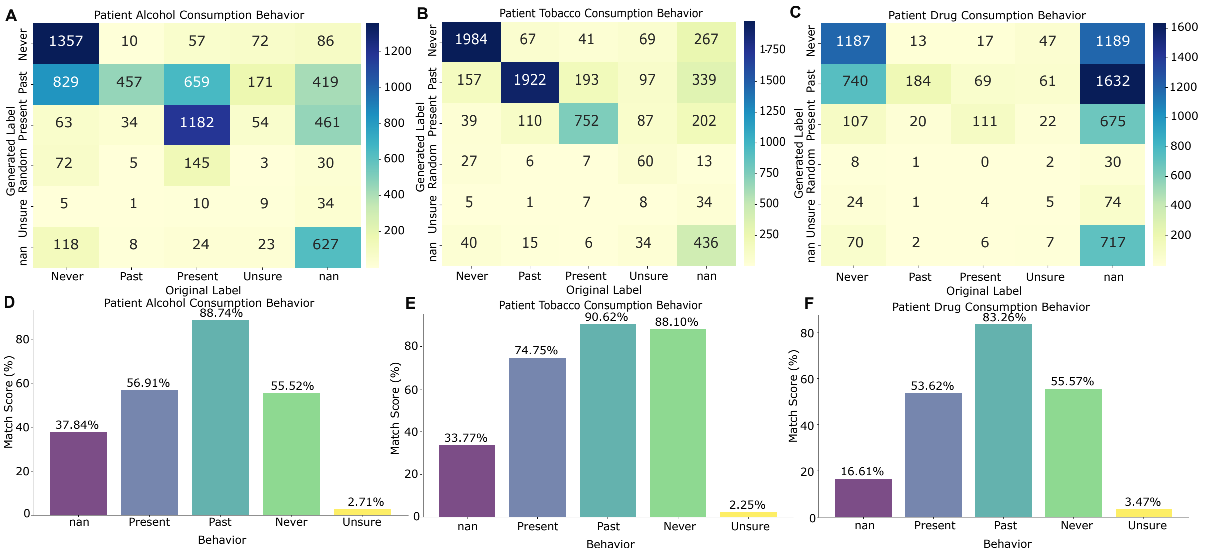

Figure 2: The heatmap shows the original and generated levels of all behavior instances, highlighting patterns of correct and incorrect generated instances (A-C). The bar plot Presents the instance-wise match scores of all substance use behaviors (D-E).

state-of-the-art [3] language model that aligns with the MIMIC-III data use guidelines [4], which prevent data sharing with third parties to avoid privacy breaches.

5 Conclusion

Despite the importance of substance use behavior in clinical decision-making (Stokes, 2019; Mejía et al., 2022) there is very limited research on automated information extraction of substance use behavior. In this study, we have evaluated the use of a LLM on annotating different instances of substance use behavior from the clinical notes. Our results explain the pre-trained LLM's ability to annotate complex substance use behavior using clinical notes. In cases of ambiguous text (*Unsure* class) and absence of text (*nan* class), the models perform poorly, highlighting the limitations of LLMs. In the case of the *Past* class, the models performed well, highlighting the strength of the model. This suggests the need for more stratified strategies and a robust evaluation methodology to adapt LLMs for real-time clinical applications.

References

Hiba Ahsan, Emmie Ohnuki, Avijit Mitra, and Hong You. 2021. Mimic-sbdh: a dataset for social and behavioral determinants of health. In *Machine Learning for Healthcare Conference*, pages 391–413. PMLR.

Hortensia Amaro, Mariana Sanchez, Tara Bautista, and Robynn Cox. 2021. Social vulnerabilities for substance use: Stressors, socially toxic environments, and discrimination and racism. *Neuropharmacology*, 188:108518.

Min Chen, Xuan Tan, and Rema Padman. 2020. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *Journal of the American Medical Informatics Association*, 27(11):1764–1773.

Kerstin Denecke, Richard May, LLMHealthGroup, and Octavio Rivera Romero. 2024. Potential of large language models in health care: Delphi study. *Journal of Medical Internet Research*, 26:e52399.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438. American Medical Informatics Association.

Jemima A Frimpong, Xun Liu, Lingrui Liu, and Ruoqiuyan Zhang. 2023. Adoption of electronic health record among substance use disorder treatment programs: Nationwide cross-sectional survey study. *Journal of medical Internet research*, 25:e45238.

Craig H Ganoe, Weiyi Wu, Paul J Barr, William Haslett, Michelle D Dannenberg, Kyra L Bonasia, James C Finora, Jesse A Schoonmaker, Wambui M Onsando, James Ryan, et al. 2021. Natural language processing for automated annotation of medication mentions

[3] https://ai.meta.com/blog/meta-llama-3/

[4] https://physionet.org/news/post/gpt-responsible-use

in primary care visit conversations. *JAMIA open*, 4(3):ooab071.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Ravishankar Jayadevappa and Sumedha Chhatre. 2016. Association between age, substance use, and outcomes in medicare enrollees with prostate cancer. *Journal of geriatric oncology*, 7(6):444–452.

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.

Vipina K Keloth, Salih Selek, Qingyu Chen, Christopher Gilman, Sunyang Fu, Yifang Dang, Xinghan Chen, Xinyue Hu, Yujia Zhou, Huan He, et al. 2024. Large language models for social determinants of health information extraction from clinical notes-a generalizable approach across institutions. *medRxiv*, pages 2024–05.

Mengyao Li, Cora Peterson, Likang Xu, Christina A Mikosz, and Feijun Luo. 2023. Medical costs of substance use disorders in the us employer-sponsored insurance population. *JAMA network open*, 6(1):e2252378–e2252378.

T Wing Lo, Jerf WK Yeung, and Cherry HL Tam. 2020. Substance abuse and public health: A multilevel perspective and multiple responses.

Maria Mahbub, Gregory M Dams, Sudarshan Srinivasan, Caitlin Rizy, Ioana Danciu, Jodie Trafton, and Kathryn Knight. 2024. Leveraging large language models to extract information on substance use disorder severity from clinical notes: A zero-shot learning approach. *arXiv preprint arXiv:2403.12297*.

A Thomas McLellan. 2017. Substance misuse and substance use disorders: why do they matter in healthcare? *Transactions of the American Clinical and Climatological Association*, 128:112.

Diana Mejía, Laurent Avila-Chauvet, and Aldebarán Toledo-Fernández. 2022. Decision-making under risk and uncertainty by substance abusers and healthy controls. *Frontiers in psychiatry*, 12:788280.

Amanda J Moy, Jessica M Schwartz, RuiJun Chen, Shirin Sadri, Eugene Lucas, Kenrick D Cato, and Sarah Collins Rossetti. 2021. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *Journal of the American Medical Informatics Association*, 28(5):998–1008.

Marin Nishimura, Harpreet Bhatia, Janet Ma, Stephen D Dickson, Laith Alshawabkeh, Eric Adler, Alan Maisel, Michael H Criqui, Barry Greenberg, and Isac C Thomas. 2020. The impact of substance abuse on heart failure hospitalizations. *The American journal of medicine*, 133(2):207–213.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Alexandra Ralevski, Nadaa Taiyab, Michael Nossal, Lindsay Mico, Samantha N Piekos, and Jennifer Hadlock. 2024. Using large language models to annotate complex cases of social determinants of health in longitudinal clinical records. *medRxiv*.

Harriet Rumgay, Kevin Shield, Hadrien Charvat, Pietro Ferrari, Bundit Sornpaisarn, Isidore Obot, Farhad Islami, Valery EPP Lemmens, Jürgen Rehm, and Isabelle Soerjomataram. 2021. Global burden of cancer in 2020 attributable to alcohol consumption: a population-based study. *The Lancet Oncology*, 22(8):1071–1080.

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.

Sarah C Snow, Gregg C Fonarow, Joseph A Ladapo, Donna L Washington, Katherine J Hoggatt, and Boback Ziaeian. 2019. National rate of tobacco and substance use disorders among hospitalized heart failure patients. *The American journal of medicine*, 132(4):478–488.

Jackie Stokes. 2019. Substance use decision-making–are clinicians using the evidence? *Journal of Social Service Research*, 45(1):16–33.

Betty Tai and A Thomas McLellan. 2012. Integrating information on substance use disorders into electronic health record systems. *Journal of substance abuse treatment*, 43(1):12–19.

Stephen H Walsh. 2004. The clinician's perspective on electronic health records and how they can affect patient care. *Bmj*, 328(7449):1184–1187.

Li-Tzy Wu, He Zhu, and Udi E Ghitza. 2018. Multicomorbidity of chronic diseases and substance use disorders and their association with hospitalization: Results from electronic health records data. *Drug and alcohol dependence*, 192:316–323.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Miryam Yusufov, Ilana M Braun, and William F Pirl. 2019. A systematic review of substance use and substance use disorders in patients with cancer. *General Hospital Psychiatry*, 60:128–136.

A Appendix

A.1 Annotation examples

In this section, we provide examples of alcohol behavior labels from clinical notes.

A.1.1 Present

*The patient quit smoking 20 years ago; **ethanol one glass of wine a day**. He is a retired elementary school principal and now works in management.*

A.1.2 Past

*The patient is a significant smoker who requires home oxygen and does have a **history of alcohol** in the past but quit 20 years ago.*

A.1.3 Never

*Patient lives alone but sons visit and a neighbor checks on her. There is a restraining order against her eldest son. Occupation: She is retired but previously worked as an American Airlines interpreter. She speaks five languages. Mobility: Unaided per family. Smoking: Never. **Alcohol: Never**. Illicits: Denies.*

A.1.4 Unsure

*Patient lives with a partner. Currently on disability. Prior prison sentence for assault many decades ago. **ETOH history in past, current use unknown**. Smokes 1-22 PPD. History of intravenous drug use, none in 8 years. His partner does not think he is taking additional non-prescription opiate meds that she knows of. Had a recent admission for narcotics overdose.*

A.1.5 nan

Patient is a non-smoker, worked at GE. According to his wife, he had never been sick before this. He is an avid golfer. In the last few weeks, he has been using his arms to climb stairs and experiencing some shortness of breath.