A Preliminary Study of Identifying Housing Outcomes from Casenotes Using Large Language Models

Anonymous submission

Abstract

In collaboration with a nonprofit organization providing homelessness services in New York, we explore the use of machine learning and statistical analysis to evaluate the impact of street outreach efforts. Assessing causal effects presents significant challenges, particularly when outcomes are missing. The first step of this work is to obtain outcome labels. While the ideal gold standard would be to have expert annotations for the entire dataset, that can be very expensive. In this preliminary study, we investigate using large language models (LLMs) to obtain these labels from unstructured case notes from street outreach teams. We compare the accuracy of LLMs when it comes to predicting human labels for four critical outcomes in street outreach. We aim for this study to serve as a proof of concept. In future work, we would like to expand on this evaluation effort and demonstrate how expert labels and LLM annotations can be combined strategically and used for causal effect estimation and evidence-based policy-making with limited data.

Introduction

Homelessness is a persistent challenge faced globally. In particular, the United States continues to experience a rise in the number of people experiencing homelessness across the nation. The National Alliance to End Homelessness reports seeing about a 12% increase in homelessness in 2023. With outreach service work being the primary form of intervention, frontline workers are tasked with the challenge of not only attempting to help such vulnerable communities get the services they need but also with documenting and measuring their progress. It is vital for social workers and nonprofit organizations engaged in outreach efforts to understand the trajectories of their clients, the challenges they may face, track intake, and measure progress and impact of their outreach efforts. Much of this information is documented as unstructured case notes in their case management system.

As natural language processing (NLP) methods and large language models (LLMs) become increasingly popular and powerful, there have been recent advances in the social and health sciences to leverage these models for extracting information efficiently from unstructured text data. This can take the form of social media posts to understand public attitudes (Ranjit et al. 2024) or clinical notes to assess patient behaviors, with a vast majority of the literature focused on the later and in healthcare domains (Ahsan et al. 2024). These tools have been proposed to help improve data quality, synthesis, and analysis, through tasks such as summarization, information extraction, and auto-complete. In this preliminary study, we focus on the information extraction task. We are interested in evaluating the ability of LLMs in the nuanced analysis of social work casenotes. This work is done in collaboration with [BLINDED FOR REVIEW], a New York based nonprofit focused on supporting people experiencing homelessness through street outreach and housing programs. For the purposes of this preliminary study, we are interested to see how human labels compare to LLM labels. Ultimately, in future work, we would like to combine these model-generated labels efficiently and responsibly with human labels for valid downstream statistical analysis of the causal effect of street outreach.

Methods

Cohort data selection. We begin by selecting a cohort of clients with consistent measurements in the engagements dataset, which contains deidentified structured data and casenotes from $35,699$ unique clients and $782,183$ client engagement notes from 2007 to 2021. We identified clients who had a documented engagement with the outreach team within a 4 month window at least 70% of the time from 2019 to 2021. We did this by a simple counting procedure for every unique client in the dataset. We removed clients who did not match this criteria. This was to ensure that clients in the final cohort have a substantive number of engagements and to avoid making comparisons between clients with systematic differences. The final cohort contained 809 unique clients, which was 2.26% of the total client database, and $272,427$ casenotes, which was roughly 34% of total engagements. Due to our teams data labeling capacity, for this initial study, we only obtained human labels and LLM annotations on a random sample (n=200) of the final dataset. For future directions, we will conduct our experiments on the full corpus of casenotes.

Keyword matching. We first consider a keyword-based approach. The approach requires a list of predefined keywords and then matches these terms in the text. We implement a simple substring matching approach to detect the presence of the keyword in each casenote. We constructed

lists of keywords for four outcomes by first conducting a manual review of a sample of notes and conversations with outreach teams for commonly used terms, acronyms, and misspellings. We identified four major keywords to use in our keyword search: *important documents and/or appointments, housing application, refusal of services, and government benefits.* Table 2 lists all the terms included under each major keyword.

Human annotations and LLM annotations. While keyword extraction is computationally inexpensive and fast, it is still inherently limited by a lack of contextual and nuanced understanding of casenotes. For this preliminary study, we specify a set of themes, two that align with the keyword search and two that are much too nuanced. The performance of the keyword match is dependent on the quality of the term list. For instance, the term "apply" used in the theme *housing application* can refer to many things including the other themes. But for a nuanced theme like *progress* or *challenges*, it becomes a challenge to enumerate a list of terms and thus LLM annotations become especially useful.

Our team expertly annotated a random sample of 200 casenotes along the following four themes: *progress towards an appointment, client challenges or regressions faced, important document types, and appointment types*. Each text received only one annotation from a member of our team, to avoid inter-annotator disagreement for this initial study. Future work will include a more detailed evaluation of the labeling schema and contextual framing of the annotations. We prompted two local LLM-models, Llama 3.1 with 8B paramenters and Llama 3.2 with 3B parameters, to assign labels within each of these themes. Tables 3 to 6 in the Appendix denotes the numerical labels used for each theme.

Results

Keyword matching. To explore the association of these themes with housing outcomes, we ran logistic regression with the themes as the explanatory variable and the housing placement status as the dependent variable and bootstrapped confidence intervals in Table 1. We find that all of the themes, except *service refusal*, are positively associated with the positive outcomes of *permanent* and *temporary/-transitional* housing, and they are negatively associated with negative outcomes like *on the streets* or *other*. *Service refusal* is positively associated with being *on the streets* and is negatively associated with *other*. This motivated our study of more complex models (i.e. LLMS) and nuanced themes in this setting.

Human annotations and LLM annotations. Next, we explore the capabilities of local large language models on understanding street outreach team casenotes. We find that Llama 3.2 with 3B parameters performs better on tasks that are very similar to keyword matching such as *types of appointments* and *types of documents*. Additionally, Llama 3.1 with 8B parameters outperformed the smaller model on more nuanced tasks such as *progress towards an appointment* and *challenges faced by clients*. Figure 1 evaluates how accurate the llm-extracted labels are compared to the human labels.

Discussion

We conducted a preliminary study to evaluate the capabilities of Llama 3.1 with 8B parameters and Llama 3.2 with 3B parameters to accurately extract relevant themes (client progress, challenges, documents mentioned and appointments) discussed in associated case notes. These initial results are promising, and we aim to make further refinements to this study. First, we only tested our results on two models. There is also room for improvement in the prompts provided to the models and testing few shot learning models. Prompt engineering significantly impacts model performance (Liu et al. 2023), so a rigorous evaluation of various models, model parameters and output formats is necessary. Another area for improvement is the size of the model. The current models were selected because they are less than 5GB and run under one hour on a 32GB Macbook Pro. Additionally, fine-tuning these local models could lead to further enhancements in performance. Another challenge lies in the labels themselves, as model-generated labels are often biased. If these imperfect annotations are to be used for downstream analysis and decision-making, we need to correct for these biases and ensure valid statistical inference.

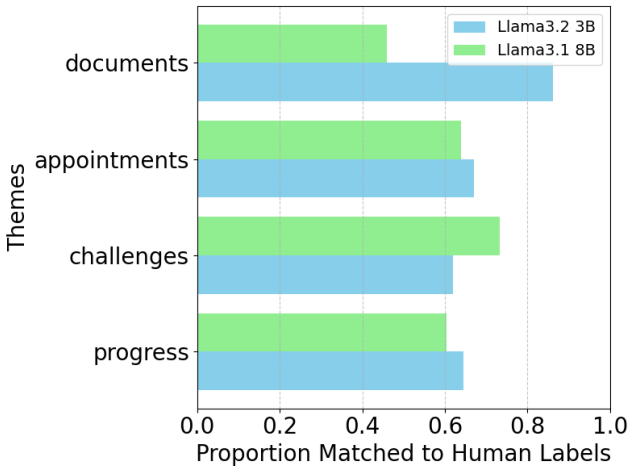

Figure 1: Accuracy comparison of Llama 3.1 and Llama 3.2 against human labels across the four themes.

Future Directions: Adaptive Human + Model Annotations

Our next step is to develop an adaptive method for obtaining model-generated labels to be used as outcomes in a downstream causal inference task. Previous work has addressed this question with M-estimation (Egami et al. 2023) and budget constraints (Zrnic and Candes 2024). We aim to apply this approach to the entire corpus of case notes to estimate the causal effect of street outreach on housing outcomes. To the best of our knowledge, this challenge has not been addressed in the causal inference and missing outcomes framework. The methodology developed in this work has broad applicability to various social datasets, particularly in scenarios characterized by missing data and unstructured text

Themes	Perm	Temp	Other	Streets
ImptDocs	0.07^*	0.088^*	-0.06^*	-0.09^*
	$(0.05, 0.09)$	$(0.065, 0.108)$	$(-0.12, -0.02)$	$(-0.11, -0.07)$
HousApp	0.13^*	0.15^*	-0.18^*	-0.09^*
	$(0.09, 0.16)$	$(0.11, 0.18)$	$(-0.26, -0.02)$	$(-0.11, -0.05)$
Benefits	0.14^*	0.083^*	0.08^*	-0.14^*
	$(0.12, 0.16)$	$(0.066, 0.107)$	$(-0.14, -0.03)$	$(-0.15, -0.11)$
Refusal	0.05	-0.03	-0.21^*	0.22^*
	$(-0.02, 0.03)$	$(-0.06, 0.005)$	$(-0.28, -0.11)$	$(0.18, 0.25)$

Table 1: Logistic regression coefficients with 95% confidence intervals for 4 features across 4 outcomes. Significance level at * $p < 0.05$.

(e.g., clinical patient notes). We also plan to enhance the usability of this approach by testing it on publicly available datasets. This work attempts to provide a method to inform evidence-based policy making with limited and unstructured data.

Ethics Statement

Our work deals with sensitive information about a vulnerable community so care must be taken when deploying our methods. The case notes are redacted and any sensitive information is removed from the notes. Furthermore, our use of local LLMs is emphasized to mitigate these privacy concerns. Our future work requires a robust evaluation to ensure responsible and fair use and to maintain the privacy of clients.

References

Ahsan, H.; McInerney, D. J.; Kim, J.; Potter, C.; Young, G.; Amir, S.; and Wallace, B. C. 2024. Retrieving Evidence from EHRs with LLMs: Possibilities and Challenges. arXiv:2309.04550.

Egami, N.; Hinck, M.; Stewart, B.; and Wei, H. 2023. Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 68589–68601. Curran Associates, Inc.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9).

Ranjit, J.; Joshi, B.; Dorn, R.; Petry, L.; Koumoundouros, O.; Bottarini, J.; Liu, P.; Rice, E.; and Swayamdipta, S. 2024. OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants. arXiv:2406.14883.

Zrnic, T.; and Candes, E. 2024. Active Statistical Inference. In *Forty-first International Conference on Machine Learning*.

Theme	Keywords
Documents/Appointments	green card, ss card/ss office, paperwork/papers, notarize/notary, start the process/process started, appointment/appt., new card/card/identification card/id, psy eval/psyc eval/psych eval/psychiatric evaluation/psycho- social/psychosocial, primary care physicians, pa appointment/pa paperwork/pa application, id copies, releases, signed/sign, photo id, tb test/tb results/tb, form/forms, b.c./bc, take his picture/take her picture, medical releases/medical records/med records, consent form/consent forms, ppd test, letter, inspection, insurance, interview, evr appointment, wecare appt., non drivers license, alien card, immigration car, state ID, eval, procured copies of, chest x-ray records
Housing Application	housing placement form, housing interview, housing application, housing pref/housing pref., progress, studio apt, waiting to hear, housing process, apply/reapply/applied, housing options, processes/processes that were involved with securing housing, securing housing
Gov't Benefits	snap, food stamps, ssi, ssdi, disability, benefits card/benefits/benefit.card, ss benefits, income, progress, medicaid/medicaid card, pa benefits, entitlement, public assistance, lost, reapply, unemployment benefits, ny benefit ids, das services
Service Refusal	refusal/refused, uncooperative

Table 2: Keywords used to search casenotes by housing theme.

Label	Progress Definition
0	No progress made
0.25	Client attempted to call or text, didn't go through
0.5	Conversation with outreach worker
1	Talked about appointment (fine-grained) ina neutral way
2	Talked about plans for completing an appointment (i.e. reminders, travel plans, status updates)
2.5	Signing documents, or completing paperwork
3	Record of completing appointment
-1	Record of a challenge or regressed on the progress later

Table 3: Labels assigned to casenotes by human annotators and LLMs for Progress theme.

Label	Challenges Definition
0	No challenges or regression mentioned
1	Client forgot or missed appointment
2	Lack of transportation
3	Client not able to be found
4	Client feeling discouraged
5	Client refused services
6	Client dissatisfied with services offered to them or with team
7	Client is dealing with health issues (i.e. mental, physical, substance abuse
8	Client is dealing with social issues (i.e. robbery, lost belongings, family, arrest)
9	Client is dealing with issues with other social services

Table 4: Labels assigned to casenotes by human annotators and LLMs for Challenges theme.

Label	Appointments Definition
0	No appointments mentioned
1	Medical appointment
2	Document retrieval appointment
3	Social benefits appointment
4	Housing appointment
5	Received a check
6	Recertification appointment
7	Domestic partnership appointment
8	Law appointment (i.e. meeting with lawyer, court date)
9	Caseworker appointment

Table 5: Labels assigned to casenotes by human annotators and LLMs for Appointments theme.

Label	Documents Definition
0	No documents mentioned
1	Budget letter
2	ID or SS card
3	Metro card
4	Atm card
5	Housing documents (i.e. housing voucher/HVL)
6	First sighting survey
7	SSI award letter
8	Benefits card
9	Medical documents

Table 6: Labels assigned to casenotes by human annotators and LLMs for Documents theme.