A Study of Large Language Models for Extraction of Themes from Homeless Shelter Case Notes

Madhumitha Selvaraj, Teale Masrani, Yani Ioannou, & Geoffrey Messier
Department of Electrical and Software Engineering
University of Calgary
Calgary, AB T2N 1N4, Canada
{madhumitha.selvaraj,teale.masrani2,yani.ioannou,gmessier}@ucalgary.ca

Abstract

Homeless shelters generate large amounts of unstructured text data in the form of case notes, which are challenging to analyze using traditional methods due to their variability and domain-specific language. This study explores the use of Large Language Models (LLMs) to extract abstract themes related to client behaviour and experiences from these notes. We focus on prompt engineering techniques and evaluate the performance of smaller LLMs against human-generated labels. Our results demonstrate that for certain themes requiring contextual understanding, smaller LLMs offer advantages over simpler methods such as keyword search or Naive Bayes. However, discrepancies between model predictions and human labels remain, with models occasionally making broad assumptions that may be undesirable. Overall, our findings highlight the role of prompt design in optimizing model performance and demonstrate the potential of LLMs to effectively understand complex homelessness data.

1 Introduction

A key strength of large language models (LLMs) is their ability to interpret contextual nuance and extract meaning from text in a manner comparable to humans. In public sector settings such as homeless shelters, data is inherently unstructured, nuanced, and often text-based in the form of case notes. While these notes contain valuable insights into shelter operations and client needs, their content varies substantially due to the unpredictable nature of client interactions. Current approaches to analyzing shelter case notes are either difficult to scale (e.g., manual reviews) or fail to account for context (e.g., keyword searches). This makes it difficult to use shelter data to answer questions such as "How many clients have expressed the desire to find housing but have not made efforts to do so?" or "How many have demonstrated violent behaviours?".

Overall, this research aims to answer the question: How well do small, open-source LLMs extract abstract themes from case notes compared to humans, and what role does prompt engineering play in their effectiveness?

2 Related Work

The use of large language models for theme extraction or classification has been widely studied for clinical notes, where themes are often more concrete and related to social determinants of health (SDOH) or clinical diagnosis. These studies typically explore variations in prompt formats and different levels of contextual information provided to the model (Ralevski et al., 2024; Ramachandran et al., 2023; Keloth et al., 2025; Gu et al., 2025; Zhang et al., 2024; Hu et al., 2024).

Annotation guidelines are commonly developed either to guide initial human labeling efforts, or to construct the LLM prompts (Relins et al., 2024; Hu et al., 2024; Zhang et al., 2024). These guides often define what to look for (e.g., indicators) and what to ignore (e.g.,

exclusions). Prompt engineering efforts often start with simpler prompts and gradually incorporate more complex instructions. Both zero-shot and one-shot prompting have been tested, though some studies have found that one-shot prompting does not always yield significant performance improvements (Zhang et al., 2024). Variations in prompt techniques include step-by-step reasoning (Zhang et al., 2024), conditional logic (e.g., "if and only if") (Gu et al., 2025), and bullet-pointed lists of relevant concepts (Relins et al., 2024).

Most studies experiment with larger, closed source models, such as GPT-4. Due to the privacy and resource constraints of the homelessness not-for-profit sector, our work examines smaller, open-source LLMs, which have not been as thoroughly studied for this task. While some studies highlight the importance of evaluating smaller models for use in limited-resource settings (Relins et al., 2024), few use models as small as used in our work.

Some related work in adjacent domains, such as police narratives (Relins et al., 2024), share certain similarities with case notes from homeless shelters. However, themes extracted in our study are more abstract and specific to the domain of homelessness support. Previous studies were often able to demonstrate high performance with concrete themes and larger models, however it remains unclear whether those results generalize to smaller models and the complex/abstract themes found in homelessness.

3 Human Labelling of Data

The data used in this work consists of over 60,000 case notes from a major North American shelter. The notes vary widely in both content and length, ranging from brief summaries of daily interactions to detailed accounts of medical emergencies, altercations, and mental health crises. In some instances, notes also reference multiple clients, which can further complicate interpretation due to redacted client names.

The themes for extraction were identified through consultation with shelter staff about their questions regarding client behaviour, and by referencing prior work on extracting SDOH terms from clinical datasets (Ramachandran et al., 2023; Gu et al., 2025). This resulted in 34 themes grouped into eight categories: Behavioural, Emotional, Employment, Housing, Mental Health, Physical Health, Relationships, and Substances. The full list is available in Appendix B. For the purposes of the theme extraction task, we focus on eight themes of focus as indicated in Appendix A.

The theme labelling approach consisted of first randomly selecting a sample of 400 notes to manually label. The length distribution of this sample was consistent with the full set of notes. A team of three coders conducted the labelling over four iterative rounds using an annotation guideline provided in Appendix B. This guideline defines each theme and outlines relevant indicators and exclusions, helping address the context-dependent nature of the notes where themes often appear in varied forms.

To evaluate the consistency of the labelling and better understand where disagreements occurred, we utilized a set of agreement metrics. Thematic analysis of shelter case notes is inherently challenging and some level of disagreement is to be expected. Table 1 shows the distribution of the eight themes of focus across the notes, along with metrics that quantify the level of coder agreement. The column "Total" indicates the number of notes where at least one of the three coders identified the theme. The next three columns show agreement among those notes: "Unanimous agreement" (all three coders), "Majority agreement" (at least two coders), and "Lone agreement" (only one coder). "Basic Agreement", is based on all 400 labelled notes and reflects the percentage of notes where coders either all agreed that the theme was present or all agreed it was absent. The labelling analysis for all 34 themes is provided in Appendix G. Overall, we observe that the most frequently selected themes fall under the categories of Behavioural, Substance Use, Emotional, and Housing, which is consistent with commonly discussed concerns within homeless shelter contexts. Some themes, however, present greater labelling challenges. For example, T4 has a significanlty lower majority agreement score in comparison to the other themes. Additionally, T4 also shows the highest Lone Agreement, indicating a larger proportion of labels assigned by a single coder. This could be mitigated in future iterations by having the coders refine

Theme	Total	Unanimous (%)	Majority (%)	Lone (%)	Basic (%)
T1 Harmful/Non Violent	103	42.7%	60.2%	39.8%	85.5
T2 Substance use	80	53.8%	76.3%	23.8%	90.9
T3 Working to Find Housing	49	44.9%	63.3%	36.7%	93.4
T4 High Instability	43	16.3%	32.6%	67.4%	91.1
T5 Violence	32	40.6%	62.5%	37.5%	95.3
T6 Mental Health	25	24.0%	60.0%	40.0%	95.32
T7 Harmful Relationships	13	38.5%	53.8%	46.2%	98.03
T8 Interactions with EMS	16	81.3%	93.8%	6.3%	99.26

Table 1: Label distribution and agreement metrics for each theme.

the annotation guide to achieve consensus specifically for these challenging themes. The majority metric demonstrates greater coding consistency for concrete themes (eg. T2, T8) than for abstract themes (eg. T4, T6, T7) which serves as a predictor of where the LLMs will be challenged.

4 Theme Extraction Models

4.1 Classic Baseline Algorithms

We begin by evaluating two baseline methods that are more intuitive and less resource-intensive than LLMs but often miss nuances for certain themes. First, we perform keyword search using theme-specific lists of keywords and key phrases based on insights from the manual labelling process. These keyword lists are available in Appendix E. While identifying accurate keywords was relatively straightforward for concrete themes, it was more challenging for abstract themes such as T6 and T3, which are inherently context-dependent and cannot be reliably reflected through specific words or phrases.

Naive Bayes was also implemented as another simple and often effective method of text classification using the majority ground truth labels and the MultinomialNB classifier from scikit-learn (scikit-learn, 2025). Due to the imbalanced nature of our dataset, where most notes for any given theme do not contain that theme, we applied SMOTE to oversample the minority class of "no theme" and reduce bias during training. We trained eight binary classifiers for each theme of focus, and used k-fold cross-validation with five folds for most themes and three folds for those with fewer labelled examples.

4.2 Small Language Models

Due to the resource constraints of the not-for-profit sector and the privacy advantages of running models locally, we consider two small models with promising benchmark performance Phi-4 Mini Instruct from Microsoft (Microsoft et al., 2025) and a publicly available fine-tuned version of LLaMA-3.1-Minitron-4B-Width-Base from Nvidia (Sreenivas et al., 2024), named LLaMA-3.1-Minitron-4B-Chat (rasyosef, 2024). Minitron was created from Llama-3.1-8B using model compression techniques such as pruning and knowledge distillation. This is where redundant parameters of the base model are removed, and the smaller model is trained to mimic the behaviour of the larger one. In contrast, Phi-4 Mini was developed using a compact transformer architecture and training it from scratch on a new curated dataset.

To utilize these models for theme extraction, we designed structured prompts and refined them through multiple rounds of iterative prompt engineering. Prompt configurations and key observations are summarized in Appendix C. The process began with baseline prompts

Theme	Keyword Search				Naive Bayes				Phi-4			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
T1	0.91	0.75	0.61	**0.67**	0.59	0.27	1.00	0.43	0.86	0.53	0.87	0.66
T2	0.91	0.68	0.79	**0.73**	0.66	0.28	0.79	0.41	0.90	0.66	0.72	0.69
T3	0.93	0.60	0.10	0.17	0.80	0.27	0.94	0.42	0.95	0.78	0.45	**0.57**
T4	0.96	0.50	0.07	0.12	0.60	0.08	1.00	0.15	0.93	0.33	1.00	**0.50**
T5	0.94	0.44	0.80	0.57	0.66	0.13	1.00	0.23	0.95	0.49	0.85	**0.62**
T6	0.98	1.00	0.40	**0.57**	0.63	0.09	0.87	0.15	0.85	0.18	0.80	0.29
T7	0.96	0.11	0.14	0.12	0.73	0.05	0.71	0.09	0.88	0.10	0.71	**0.17**
T8	0.94	0.39	1	0.57	0.92	0.30	0.93	0.46	0.96	0.47	1	**0.64**

Table 2: Comparison of performance metrics across methods

to assess general capabilities, followed by progressively adding details from the annotation guideline. While both models exhibited some common output issues, Minitron performance was far inferior to Phi-4 as illustrated in Appendix F. Therefore, Phi-4 is used to generate final results.

Through the prompt engineering process, we concluded that single-theme prompts were essential to achieve acceptable performance. This approach allowed for strategic modifications based on theme-specific performance, which proved more effective than determining prompt changes that would be generalizable across multiple, often highly varied, themes.

Using the single theme prompt format, we conducted an initial round of inference on Phi-4 for the eight themes and reviewed model responses for areas of improvement. A key adjustment across all themes involved rewording vague language from the annotation guideline into more direct and unambiguous phrasing, since the original guideline was designed for human labelling and intentionally left room for interpretation due to the varied ways themes appear in notes. However, this vagueness often led the model to make overly speculative conclusions, such as inferring a person's mental state from minor occurrences of harmful behaviour. Therefore, the guideline was unsuitable for direct use in model prompts.

The revised prompts, used in a second round of inference, focused on clarifying exclusions and making indicators less ambiguous for each theme. For some themes such as T4 High Instability, we implemented more significant changes including a different output format to better accommodate the unique nature of the theme. The final revised prompts for all eight themes of focus are provided in Appendix D.

4.3 Results

The performance of the two baseline methods and Phi-4 is shown in Table 2, where results were compared to majority ground truth labels. As expected, keyword search achieved higher recall on concrete themes like T2 and T5 where relevant keywords commonly appear in notes. However, precision was low for some concrete themes such as T8, where the presence of acronyms like "EMS" did not always indicate interactions with the client specifically. In contrast, abstract themes like T3, T4, and T7 had much lower recall, as identifying them required a contextual understanding beyond simple keywordmatching. In some cases, most notably T6, keyword search did outperform Phi-4 Mini, suggesting that overly broad assumptions sometimes made by LLMs can hurt performance.

Naive Bayes classifiers showed generally high recall across the themes but very low precision. This suggests that the MultinomialNB model, which estimates class probabilities based on word frequency, is poorly suited to this dataset because word count often does not correspond to the presence of a theme. While some relevant features like "drug paraphernalia" for T2 and "his son" for T7 were weighted highly, many top features were common but irrelevant words.

Phi-4's precision was relatively high for concrete themes like T2 and T8, but declined sharply for abstract themes such as T6 and T7. This suggests the model often inferred without sufficient evidence in the notes for abstract themes, which require accurate contextual

inference to identify. Phi-4 generally tended to over-identify themes, showing higher recall than precision. The exception was T3, where the model achieved higher precision. In most of these missed cases the model concluded that the note `\does not provide evidence of any concrete steps being taken to find housing.`". This likely reflects the prompt's emphasis on identifying explicit actions rather than only interest.

In comparison to the first round of inference, the revised prompts reduced false positives across all themes but also led to a slight increase in missed notes. However, false positives for T7 decreased only slightly, despite clarifications in excluding staff members, suggesting the need for additional strategies. For T4, rather than clarifying indicators or exlusions, precision improved more after modifying the output format to justify why there was overreaction. Exploring variations in output format may also improve performance across other themes. Finally, for some themes, like T4, precision improved considerably with minimum ground truth labels rather than majority. This difference highlights the impact of disagreements during labelling, which is evident in the high lone agreement for T4 in Table 1.

5 Discussion

This work highlights the importance of a structured labeling process with well-defined themes to support accurate annotation and meaningful model evaluation As shown in Table 1, full agreement among annotators remained challenging despite multiple rounds of discussions, underscoring the need for multiple coders to capture ambiguity and minimize individual bias in the labelling process.

Phi-4 Mini significantly outperformed Minitron, likely due to differences in training data quality or its design as a condensed model rather than one distilled from a larger architecture. However, while Phi-4 shows promise, it did not outperform keyword search on some concrete themes and struggled with some abstract ones. Crafting effective prompts, manually interpreting results, and iterating through experiments also required significant effort. However, LLMs can supplement manual review by flagging potentially relevant notes these, particularly when keyword search is not viable. Consistent with (Relins et al., 2024), these models are particularly effective at identifying cases where a theme is *not* present. This makes them well-suited for narrowing down large datasets for future human review, which remains essential due to low model precision. Rather than replacing manual review, LLMs could help accelerate the process in shelter settings where scalable tools are limited. Future work could explore prompt tuning, few-shot learning, or fine-tuning, and assess whether challenges with abstract themes persist across larger models or other datasets. Comparing small model development strategies may also provide valuable insights into model selection for complex tasks such as abstract theme extraction.

6 Conclusion

Thematic analysis of unstructured homeless shelter data can allow service providers to better understand the challenges faced by the population of people experiencing homelessness. LLMs are the only viable tool since thematic labeling is both a large-scale problem (beyond manual labeling) and requires an understanding of language context (beyond simple keyword or word frequency analysis).This paper demonstrates both the potential of small LLMs for secure, local deployment while also highlighting the challenges that must be met before these models realize that potential.

Social Impacts Statement

Unstructured text data is inherent to many homeless shelter settings. LLMs have greater potential to analyze this data at a population level compared to other methods, offering valuable insights into broad themes such as mental health challenges, substance use, and harmful past relationships. This technique can significantly benefit homeless shelters and front-line workers by enhancing their understanding of client needs and allowing for more

targeted support. Additionally, rather than directly questioning individuals experiencing homelessness, automated analysis can leverage data that currently exists in the form of case notes and therefore allow for less invasive solutions that serve vulnerable populations.

Acknowledgments

The authors would like acknowledge support from the Calgary Drop In Centre, the Government of Alberta and Making the Shift. This study is based in part on data provided by Alberta Community and Social Services. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the Government of Alberta. Neither the Government of Alberta nor Alberta Community and Social Services express any opinion related to this study.

References

Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J. Desai. Scalable information extraction from free text electronic health records using large language models. *BMC Medical Research Methodology*, 25(1):23, Jan 2025. ISSN 1471-2288. doi: 10.1186/s12874-025-02470-z. URL https://doi.org/10.1186/s12874-025-02470-z.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, 01 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocad259. URL https://doi.org/10.1093/jamia/ocad259.

Vipina K. Keloth, Salih Selek, Qingyu Chen, Christopher Gilman, Sunyang Fu, Yifang Dang, Xinghan Chen, Xinyue Hu, Yujia Zhou, Huan He, Jungwei W. Fan, Karen Wang, Cynthia Brandt, Cui Tao, Hongfang Liu, and Hua Xu. Social determinants of health extraction from clinical notes across institutions using large language models. *npj Digital Medicine*, 8(1):287, May 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01645-8. URL https://doi.org/10.1038/s41746-025-01645-8.

Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. Automating thematic analysis: How llms analyse controversial topics, 2024. URL https://arxiv.org/abs/2405.06919.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. Technical report, 2025. URL https://arxiv.org/abs/2503.01743.

Alexandra Ralevski, Nadaa Taiyab, Michael Nossal, Lindsay Mico, Samantha Piekos, and Jennifer Hadlock. Using Large Language Models to Abstract Complex Social Determinants of Health From Original and Deidentified Medical Notes: Development and Validation Study. *Journal of Medical Internet Research*, 26:e63445, November 2024. ISSN 1438-8871. doi: 10.2196/63445.

Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nic Dobbins, Ozlem Uzuner, and Meliha Yetisgen. Prompt-based extraction of social determinants of health using few-shot learning. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky (eds.), *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 385–393, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.41. URL https://aclanthology.org/2023.clinicalnlp-1.41/.

rasyosef. Llama-3.1-minitron-4b-chat, 2024. URL https://huggingface.co/rasyosef/Llama-3.1-Minitron-4B-Chat.

Sam Relins, Daniel Birks, and Charlie Lloyd. Using Instruction-Tuned Large Language Models to Identify Indicators of Vulnerability in Police Incident Narratives, December 2024. URL http://arxiv.org/abs/2412.11878. arXiv:2412.11878. Manuscript under review.

scikit-learn. Multinomialnb, 2025. URL https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Ameya Sunil Mahabaleshwarkar, Gerald Shen, Jiaqi Zeng, Zijia Chen, Yoshi Suhara, Shizhe Diao, Chenhan Yu, Wei-Chun Chen, Hayley Ross, Oluwatobi Olabiyi, Ashwath Aithal, Oleksii Kuchaiev, Daniel Korzekwa, Pavlo Molchanov, Mostofa Patwary, Mohammad Shoeybi, Jan Kautz, and Bryan Catanzaro. Llm pruning and distillation in practice: The minitron approach, 2024. URL https://arxiv.org/abs/2408.11796.

Xiaodan Zhang, Nabasmita Talukdar, Sandeep Vemulapalli, Sumyeong Ahn, Jiankun Wang, Han Meng, Sardar Mehtab Bin Murtaza, Dmitry Leshchiner, Aakash Ajay Dave, Dimitri F Joseph, Martin Witteveen-Lane, Dave Chesla, Jiayu Zhou, and Bin Chen. Comparison of prompt engineering and Fine-Tuning strategies in large language models in the classification of clinical notes. *AMIA Jt Summits Transl Sci Proc*, 2024:478–487, May 2024. URL https://pubmed.ncbi.nlm.nih.gov/38827053/.

A Themes of Focus for Theme Experimentation

We identified 8 out of the 34 case note themes as themes of focus for the theme extraction experimentation. These are outlined in Table 3.

Theme	Full Theme Name
T1 Harmful/Non Violent	Displays harmful, but non-violent behaviour towards others.
T2 Substance Use	Substance use.
T3 Working to Find Housing	Actively working to find housing.
T4 High Instability	Indicates a high level of mental or emotional instability.
T5 Violence	Displays physically violent behaviour towards others.
T6 Mental Health	Poor mental health or experiencing mental health issues.
T7 Harmful Relationships	Has harmful personal relationships.
T8 Interactions with EMS	Has interactions with EMS (emergency medical services).

Table 3: Overview of themes of focus for theme extraction.

B Annotation Guideline

B.1 Harmful Relationships

This theme applies when there is evidence of the client being involved in ongoing relationships that may negatively impact their well-being or safety. Relationships can include family members, friendships, romantic partners, etc. These relationships can be within the shelter, or external.

Indicators:

- Descriptions of the client's relationship containing harmful interpersonal dynamics, such as recurring conflicts, violence, or manipulation.
- Patterns of negative interactions that suggest a sustained relationship rather than isolated incidents.

Exclusions

- Situations where there isn't clear evidence of an ongoing relationship (e.g., an altercation with another individual at the shelter, or with shelter staff).

B.2 Positive Relationships

This theme applies when the client is involved in supportive and positive relationships with others that have a positive influence on their overall well-being. Relationships can include family members, friendships, romantic partners, etc. This can include relationships within the shelter, or external.

Indicators:

- Mentions of the client having ongoing positive interactions or connections with others.
- Evidence of ongoing supportive behaviour such as offering or receiving help, visits, assistance with tasks, etc.
- Positive relationships with others being described as ongoing over an extended period (e.g., "friends for X years", "supported by their parents for X years", etc.).

Exclusions

- Isolated friendly interactions or one-off positive exchanges that do not indicate an ongoing relationship. A more accurate theme for these cases would be Theme 8.
- Mentions of family members without clear evidence of positive or supportive involvement, particularly if the relationship is distant.
- In general, friendly or respectful behaviour towards shelter staff.

B.3 Isolated from others

This theme applies if the client lacks a positive support system, either through a lack of family or friendships, or through intentional isolation.

Indicators:

- Explicit mention that the client does not have family or friends that could provide them with support.
- Mentions of family members that are uninvolved in the client's life or distant.

B.4 Displays physically violent behaviour towards others

This theme applies when the client exhibits or has a history of physically violent actions, either towards others within the shelter or in external environments.

Indicators:

- Mentions of the client engaging in physical attacks against others, including fights, pushing, hitting, or any other form of violent contact.
- Incidents where the client had to be physically restrained to prevent harm to others.
- Descriptions of aggressive physical behaviour that may have escalated into violence.

Exclusions

- Non-violent aggressive behaviour, such as verbal threats or emotional outbursts, without clear evidence of physical harm.

B.5 Displays harmful, but non-violent behaviour towards others

This theme applies when the client exhibits harmful behaviours that negatively affect others, but these actions do not involve violence.

Indicators:

- Mentions of non-violent harmful behaviour, such as arguments, verbal threats, insults, or hate speech.
- Engaging in activities such drug dealing that can negatively impact the well-being of others at the shelter.
- Actions that create a toxic atmosphere in the shelter, such as bullying.
- Instances of disrespect towards others or shelter staff, for example repeatedly refusing to follow instruction.
- Non-violent violation of rules or laws that impact other clients (i.e. stealing).

Exclusions

- Instances where the client expresses frustration or disagreement without resorting to harmful language or actions.
- Anything that falls under Theme 7.

B.6 Displays inappropriate behaviour towards others

This theme applies when the client exhibits inappropriate or socially unacceptable behaviour towards others, either in the form of verbal comments, actions, or interactions.

Indicators:

- Mentions of the client making inappropriate comments, such as offensive jokes or using crude language.
- Inappropriate behaviour such as invading personal space or making inappropriate gestures.
- Behaviour that makes others uncomfortable or disturbs others, even if the client did not intend to cause harm.

Exclusions

- Hate speech or derogatory remarks – This would fall under Theme 5.

B.7 Perpetrator of Sexual harassment

This theme applies when there is evidence and description of the client sexual harassing others.

Indicators:

- Direct mentions of the client being a perpetrator of sexual harassment towards others
- Descriptions of behaviour that can be deemed as sexual harassment, such as unwanted touching.
- If the note only mentions "Unwanted touching", then label as Theme 7. However, if the "unwanted touching" occurred as part of larger situation, such as a fight, then will need to consider the context of the situation

Exclusions

- Making other uncomfortable with behaviour that would not be considered as "sexual harassment", this would be labelled with Theme 6.

B.8 Displays positive behaviours toward others

This theme applies when the client demonstrates positive, supportive, or friendly behaviours toward others.

Indicators:

- Friendly, respectful interactions with others, including staff and other clients.

B.9 Has interactions with EMS (emergency medical services)

This theme applies when the client has direct or indirect involvement with emergency medical services such as EMS personnel, ambulance services, or emergency hospital visits.

Indicators:

- Emergency medical interventions, such as an ambulance or EMS being called for the client.
- Mentions of the client interacting with EMS workers or paramedics

Exclusions

- Mentions of possible calls to EMS or 911. For example, if shelter staff describe a situation where they were thinking about calling 911/EMS but ultimately did not, then the note should not be labelled as Theme 9.

B.10 Has interactions with the police or justice system

This theme applies when the client has interactions with either law enforcement, or the justice system.

Indicators:

- Instances of police officers being involved in situations related to the client, such as responding to physical altercations or other incidents requiring intervention.
- Mentions of client being approached by law enforcement outside of the shelter environment.
- Mentions of the client appearing in court, meeting with lawyers, or being required to participate in legal proceedings.

B.11 Is unable to provide self-care (e.g., poor hygiene, incontinence, hoarding)

This theme applies when there are clear indications that the client struggles to meet basic self-care needs and maintaining hygiene.

Indicators:

- Descriptions of the client displaying poor hygiene, especially if this is described as a persistent issue.
- Examples of poor hygiene could include strong body odour or wearing visibly soiled clothing
- References to issues with incontinence or an inability to manage personal bodily functions
- Instances that describe the client needing repeated reminders to maintain hygiene

Exclusions

- Self-directed violence or other harmful behaviour

B.12 Is able to provide strong self-care (e.g., cleanliness)

This theme applies when the client demonstrates behaviours that maintain self-cleanliness, hygiene, and overall wellness.

Indicators:

- Descriptions of the client being consistently clean or well-groomed.
- Mentions of the client effectively managing their personal hygiene on their own, i.e. demonstrating independence in this area.

Exclusions

- Instances where clients are only able to maintain self-cleanliness with assistance, i.e. is unable to consistently uphold that on their own.

B.13 Currently employed

This theme applies when there is evidence that the client is actively employed.

Indicators:

- Direct mentions of the client having a job or occupation (e.g., references to specific roles, job titles, or workplaces).
- Descriptions of the client going to work, returning from work, or discussing their work schedule.
- Mentions of paycheques or their income for rendering some form of service.

B.14 Currently unemployed

This theme applies when there is evidence that the client is currently unemployed and not earning income but are capable of working.

Indicators:

- Descriptions of the client actively looking for work or expressing a desire to work but unable to secure a job.
- Mentions of the client choosing not to work, or not having a desire to search for employment.

Exclusions

- Clients who are unable to work due to a disability (this would fall under Theme 15).

B.15 Unable to work due to disability (or other reasons)

This theme applies when clients are unable to work due to a disability or other significant barriers impacting their ability to gain or maintain employment.

Indicators:

- Mentions of the client receiving income support such as AISH.
- Mentions of the client bring up a disability and specifically mentioning that they are struggling to find work because of it. This can include chronic illnesses, physical limitations, mental health struggles, etc.
- Clients discussing how their disability or other significant barriers (e.g., caregiving responsibilities, legal restrictions) have affected their employment history or ability to stay employed.

Exclusions

- Mentions of a disability or health condition without any reference to its impact on employment (e.g., descriptions of a physical disability or chronic illness that do not explicitly mention difficulty with work).
- Situations where the client is unemployed, but has no disability or significant barrier preventing them from working (these should fall under Theme 14).

B.16 Poor physical health or experiencing physical health issues

This theme applies if the note describes situations of the client experiencing health issues or having to receive medical care.

Indicators:

- Mentions of the client experiencing health issues where a shelter would have a negative impact on their ability to recover or improve.
  - For example, mentions of the client having a cold with no context on the severity would not be labelled as theme 16.
  - However, more chronic and serious health issues such as diabetes, cancer, bone fractures, etc., would be labelled as theme 16.
- References to the client seeking or receiving medical care, such as visiting a hospital.
- Descriptions of the client being limited in their day-to-day activities due to physical conditions.
- Mentions of the client experiencing long term effects from health issues.

Exclusions

- Notes that only describe poor mental health.
- A prior health problem that the client has recovered from.
- Injuries from acts of violence (acute health issues from events).
- Mentions of medication without context.
- Short term health issues without context on their severity (such as a cold).

B.17 Medical Emergency or is Experiencing Medical Distress

This theme applies if the note described a scenario where the client is experiencing a medical emergency or is medial distress requiring urgent attention.

Indicators:

- Descriptions such as the client exhibiting symptoms of medical distress such as being unresponsive, having difficulty breathing, feeling disoriented, etc.
- Instances where the client needed immediate attention from shelter staff

Exclusions

- Situations where EMS or paramedics were called and are involved in the situation – This would be labelled as Theme 9.
- General mentions of chronic illness without an urgent medical event being described.

B.18 Expressed Interest in Finding Housing

This theme applies when the client expresses interest or a desire to find housing but has yet to take concrete steps or engage with housing services or assistance.

Indicators:

- General statements about wanting to find housing, move out of the shelter, or improve their living situation.
- Asking staff questions regarding housing options or for help getting started with finding housing.

B.19 Actively Working to Find Housing

This theme applies when the client shows they are actively taking steps to secure housing. Compared to Theme 20, this theme focuses more on clear actions and behaviours the clients exhibit that indicates they are trying to find housing, rather than their emotions or beliefs in this effort.

Indicators:

- Making efforts to apply for housing such as filling out the correct forms, submitting applications, etc. Mentions of NSQ application.
- Indicating they want to view housing options and then appearing for those opportunities
- Sorting out their finances to be able to afford housing options.
- Asking for assistance or resources related to housing, such as working with caseworkers or using housing services, or asking for family assistance.

Exclusions

- General statements about wanting to improve their living situation – This would be labelled as Theme 18.

B.20 Has a high belief in housing success or leaving shelter

This theme applies when the client indicates a strong desire or confidence in their ability to secure housing or leave the shelter.

Indicators:

- Expressing positive emotions when describing their process or results in finding housing.
- Expressions of optimism or determination to find housing.
- Showing persistent behaviour during challenges in finding housing.

B.21 Has a low belief in housing success or leaving shelter

This theme applies when the client indicates doubt or lack of confidence in their ability to secure housing or leave the shelter.

Indicators:

- Expressing negative emotions or desire to stop searching for housing.
- Mentions of the client feeling hopeless or frustrated at the process.
- Indications of a desire to stay in the shelter indefinitely or a lack of motivation to transition out.
- Expressing fear or anxiety about life outside the shelter or securing stable housing.

B.22 Hopeful about the future

This theme applies when the client indicates that they are hopeful about their future, either immediate or long term.

Indicators:

- Making plans for life outside the shelter and feeling a sense of optimism around achieving those plans.
- Setting goals for what they want to accomplish either in the shelter or outside.
- Expressing positive outlooks about the potential to change their situation.
- Looking forward to something in the future (e.g. plans they have made, an event, or receiving money).

B.23 Not hopeful about the future

This theme applies when the client demonstrates pessimism, doubt, or a lack of confidence about their immediate or long-term future.

Indicators:

- Statements expressing feelings of hopelessness or resignation regarding their future.
- Lack of motivation towards setting goals or making plans.

B.24 Indicates emotional distress

This theme applies when the client shows signs of emotional distress or has difficulty processing their emotions in response to distressing situations or events. In general, signs of emotional distress are appropriate reactions to negative events.

Indicators:

- Expressions of sadness or frustration, or emotional pain in response to specific events or circumstances.
- Strong emotional reactions such as anger, crying, or withdrawing after experiencing challenges.
- Mentions of feeling overwhelmed or upset by their situation or interactions.

Exclusions

- General complaints made without an emotional element

B.25 Indicates emotional stability

This theme applies when the client shows signs of emotional stability, especially in reaction to difficult situations or events.

Indicators:

- Accepting bad news without becoming overly distressed or agitated.
- Cooperation with staff during challenging situations.
- Showing self-restraint or the ability to regulate their emotions and behaviour.
- Demonstrations of emotional resilience, e.g. maintaining a positive outlook despite challenges.

Exclusions

- De-escalations after a conflict. For example, the client being described as calming down after a fight or following instructions after arguments with a staff member.
- In general, this theme applies only when the client demonstrates emotional stability from the outset (such as avoiding conflicts and showing restraint).

B.26 Indicates a high level of mental or emotional instability

This theme applies when the client demonstrates large, disproportionate reactions and behaviours in the context of situation described in the note. In comparison to emotional distress, the client may not be particularly upset or feel negative emotions.

Indicators:

- Outburst of emotion that are disproportionate to the situation.

- Difficulty calming down or self-regulating.
- Descriptions of the client being in an uncontrollable state.

Exclusions

- Emotional reactions that are appropriate for the situation.
- Relatively quiet reactions or behaviours, even if they could be signs of instability. For example, talking to oneself or singing would fall under Theme 32

B.27 Substance use

This theme applies when there is evidence that the client is using substances, either through direct acknowledgment or indirect signs observed by shelter staff.

Indicators:

- Direct mentions of drug or alcohol use, either inside or outside the shelter.
- Indirect mentions of substance use, such as references to paraphernalia (e.g., tin foil, pipes, or syringes) or behaviours strongly associated with substance use.
- Client explicitly stating they have addiction issues. This theme can be assigned alongside other relevant themes if applicable.
- References to prior struggles with addiction.

B.28 Making efforts to seek substance use treatment

This theme applies when the client expresses a desire to seek substance use treatment or has made efforts to start treatment or explore options.

Indicators:

- Statements from the client indicating a desire to quit or reduce substance use.
- Clients having general interest in substance use treatment. For example, wanting to explore treatment options or start discussions.
- Engaging with staff about enrolling in treatment programs or support groups.

Exclusions

- Mentions of substance use or frustrations with their addiction issues, without statements that indicate a desire to seek treatment or to quit.

B.29 Receiving substance use treatment

This theme applies when the client is actively receiving treatment for substance use.

Indicators:

- Mentions of the client participating in detox bans, detox treatments, withdrawal management, etc.
- References to methadone.
- Mentions of "witnessed medications".

Exclusions

- General statements made by the client regarding substance use treatment, such as wishing to overcome their substance use struggles, or being interested in treatment options available. Such notes would be labelled as Theme 28.
- This theme only applies if *active* substance use treatment is mentioned.

B.30 Overdose

This theme applies if a drug overdose is described in the note. This includes both confirmed and suspected overdoses where medical intervention was required.

Indicators:

- Direct mentions of a possible or confirmed overdose.
- Descriptions of medical intervention due to a suspected overdose, such as emergency services being called.
- Mentions of naloxone being administered.

Exclusions

- Mentions of clients being at risk of overdose, or past instances where they experienced an overdose. This theme applies for notes describing an *overdose event.*

B.31 Receiving mental health treatment

This theme applies if the client is engaging in treatment for mental health concerns.

Indicators:

- Mentions of the client currently seeking or receiving treatment for mental health concerns (e.g., therapy, counseling, psychiatric care).
- References to attending therapy sessions, support groups, or appointments with mental health professionals.
- Mentions of taking prescribed medication as part of a treatment plan for mental health concerns.

Exclusions

- General statements about mental health struggles without any mention of active treatment efforts.

B.32 Poor mental health or experiencing mental health issues

This theme applies when there are indirect references that the client is experiencing mental health challenges, or if there are descriptions of behaviour that could be considered signs of mental health issues.

Indicators:

- Erratic behaviour such as talking to oneself or singing when it's not appropriate to do so.
- Evidence of a cognitive impairment (i.e. mentions of the client having a learning disability, memory, attention deficit disorder, etc.).
- The client talking about past trauma they have experienced.
- Observations or concerns made by shelter workers regarding the client's mental health, such as noting signs of distress, withdrawal, drastic changes in behaviour or mood, etc.

Exclusions

- Direct mentions of a mental health diagnosis – this would fall under Theme 33.

B.33 Mentions of Mental Health Diagnosis

This theme applies when there are direct references that the client having a diagnosed or self-reported mental health condition, either currently or in the past. experiencing mental health issues either currently or in their past.

Indicators:

- Direct mentions from the client about their mental health struggles, such as feelings of depression, anxiety, PTSD, or other mental health concerns.
- Mentions of self-harm or self-inflicted violence.
- References to the client taking medications used to treat mental health disorders.

Exclusions

- Descriptions or behaviour that appear or could be mental health related – this would fall under Theme 32.

B.34 Indicates Suicide Ideation

This theme applies when the client explicitly expresses or implicitly demonstrates thoughts or behaviours that suggest they are contemplating suicide or engaging in self-harm.

Indicators:

- Direct mentions of suicidal thoughts expressed by the client, or suicide attempts.
- Observations made by shelter staff that raise concern about the client's safety and potential risk of suicide.

Exclusions

- General mentions of mental health issues that do not specifically reference suicide ideation – This would be labelled as the Theme 32 or 33.
- Expressions of hopelessness without direct mention that the client is having suicidal thoughts – This would be labelled as the Theme 23.

C Summary of Tested Prompt Configurations

Variations of prompts for theme extraction were developed and evaluated on both the Minitron and Phi-4 Models. Table 4 presents the three main prompt types with configuration details, performance observations, and key takeaways for designing subsequent prompts.

D Final Single Theme Prompts

The single theme prompts presented here for all eight themes of focus were used to obtain the final performance evaluations for Phi-4.

D.1 Substance use

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the note contains evidence of substance use by the client. This includes both current and past substance use. Only use descriptions of the client's behaviour and actions, not those of others such as friends, to determine if substance use is involved. It

Prompt Type	Configuration Details	Model Observations	Key Takeaways
Full Theme Prompt	A single prompt asked the model to identify relevant themes from a list of all 34 themes. Techniques tested included: (1) using theme names only; and (2) using theme names with definitions. The full annotation guideline was excluded due to memory constraints.	**Minitron:** Demonstrated poor performance with justifications unrelated to the content or themes. **Phi-4:** Similarly poor performance, though it showed slight improvement when definitions were included. Frequently misinterpreted abstract themes.	A list of 34 themes overwhelmed both models. Theme definitions alone were not sufficient for abstract themes. The prompt strategy was revised to reduce the number of themes presented in each prompt.
Theme Group Prompt	Eight prompts were created, each corresponding to a theme group with only 3–5 themes. Techniques tested included: (1) theme definitions only; and (2) full annotation guideline including definitions, indicators, and exclusions.	**Minitron:** Showed no meaningful improvement. **Phi-4:** Performance improved slightly with theme definitions. Including the full guideline made the model overly restrictive, often returning "No Themes" even when relevant themes were present.	Reducing the number of themes per prompt led to slight improvements, but responses remained inconsistent. Full guidelines appeared to overwhelm the models. The prompt strategy was revised to focus on one theme per prompt.
Single Theme Prompt	Each prompt focused on a single theme (eight total). Techniques tested included: (1) binary Yes/No output format with justification and confidence score; (2) step-by-step reasoning (Zhang et al., 2024); (3) confidence scoring (Khan et al., 2024); and (4) simplification and reordering of indicators and exclusions.	**Minitron:** Slight improvement, however justifications remained frequently misaligned with the provided guidelines. **Phi-4:** Significant gains in output quality. Followed the requested output format consistently.	This was the most effective configuration in terms of achieving high output accuracy. Concluded that Minitron was insufficient and poorly suited for our task. This format was further refined into the final prompt configuration, which was focused solely on Phi-4.

Table 4: Summary of Prompt Configurations and Observed Model Performance

is important to not assume a client is using substances based on erratic or unstable behaviour. Use only direct evidence of substance use from the text.

Use the following indicators to identify current or past substance use in the note:

1. Mentions of drug or alcohol use, either inside or outside the shelter. This is often indicated by the term "using".
2. Mentions of drug paraphernalia (e.g., tinfoil, pipes, syringes, contraband). The term "paraphernalia" on its own often refers to drug paraphernalia.
3. The client stating they have addiction issues, or have struggled with addiction in the past.
4. The client appearing intoxicated, drunk, or not sober.

5. Descriptions of the client experiencing withdrawal symptoms.
6. Mentions of an overdose event, or the use of naloxone.

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if the note contains evidence of current or past substance use. Respond with <<NO>> if the note does not contain such evidence. Please enclose your answer in double angle brackets << >>
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.2 Actively Working to Find Housing

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the client is actively working to find housing. This means the client is showing clear actions and behaviours that indicate they are trying to find housing, rather than just desiring housing. It is important to not focus on how the client feels about the process or the housing options available. For example, a client actively looking for housing may feel frustrated with the process, experience setbacks such as scheduling conflicts, or express disappointment with the available options. However, these feelings or challenges do not exclude them from being considered "actively looking for housing".

Use the following indicators to identify the presence of taking clear actions to find housing:

1. Scheduling or attending housing viewings, or actively searching for rental apartments or homes
2. Applying for housing, including filling out housing intake forms or submitting housing applications
3. Mentions of the client filling out an NSQ application
4. Mentions of transitioning to housing from another program
5. Waiting for housing referrals or requesting referrals from staff
6. Working with caseworkers to find housing
7. Seeking assistance from housing services or housing support, or having meetings with housing staff
8. Seeking help from family members to find housing or obtain financial support for housing

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if the note contains clear indicators that the client is actively looking for housing. Respond with <<NO>> if the client is not actively looking for housing. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.3 Displays physically violent behaviour towards others

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to analyze the case note and determine if the client is exhibiting physically violent behaviour. Only consider descriptions of the client's behaviour and actions. Do not consider the outcome of their behaviour, or whether the note states that someone was harmed. For example, the note does not need to mention that someone was injured for the client to be exhibiting physically violent behaviour.

Use the following indicators to identify physically violent behaviour in the note:

1. Descriptions of physical attacks. This can include hitting, punching, kicking, pushing, spitting, slapping, or biting.
2. Descriptions of aggressive physical behaviour. This can include throwing objects or food at someone, forcibly grabbing another person, or physically intimidating another person.
3. Mentions of the client being physically restrained by staff to prevent harm to others.

Please keep in mind that yelling, swearing, and verbal arguments are not physically aggressive behaviour.

Your response must include the following three components:

- YES or NO answer: Respond with `<<YES>>` if the note describes the client displaying physically violent behaviour. Respond with `<<NO>>` if the note does not describe such behaviour. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.4 Displays harmful, but non-violent behaviour towards others

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the client exhibits harmful but non-violent behavior towards others. This applies if the client engages in harmful behavior that negatively affect others, but without physical violence. If the client physically harms someone, flag the note if there are additional signs of non-violent harmful behavior. Do not flag the note if the client only expresses frustration or disagreement without resorting to harmful language or actions.

Some harmful behavior should not be considered, such as clear indications of sexual harassment.

Use the following indicators to determine if the client displays harmful, non-violent behavior towards others:

1. Mentions of non-violent harmful behaviors, such as arguments, verbal threats, insults, or hate speech
2. Engaging in activities, such as drug dealing, that negatively impact the well-being of others at the shelter
3. Actions that create a toxic atmosphere in the shelter, such as bullying
4. Instances of disrespect towards others or shelter staff, such as repeatedly refusing to follow instructions

5. Non-violent violations of rules or laws that impact other clients (i.e. stealing)

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if the note contains clear indicators that the client exhibits harmful but non-violent behavior towards others. Respond with <<NO>> if not. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.5 Indicates a high level of mental or emotional instability

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the client shows signs of a high level of mental or emotional instability. This applies only applies the client displays reactions or behaviours that are extreme, or significantly out of proportion to the situation described in the note. Do NOT flag the note if the client is having a normal emotional response that is appropriate for the situation. For example, expressing sadness after receiving bad news or frustration during a conflict is considered a normal emotional response. Do NOT flag the note based on harmful or negative behaviour unless there is clear evidence of instability.

Use the following indicators to determine if the client shows signs of HIGH mental or emotional instability.

1. Strong outbursts of emotion that are extremely disproportionate to the situation
2. Difficulty calming down or self-regulating after a strong outburst
3. Descriptions of the client being in an uncontrollable state

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if there are clear indicators from the list above that the client shows a high level of mental or emotional instability. Respond with <<NO>> if there are not. Please enclose your answer in double angle brackets << >>.
- Justification: Clearly explain your reasoning by including all three of the following:
  - The situation: Briefly describe the situation the client is responding to in one sentence.
  - The client's reaction: Briefly describe the client's response or feelings towards the situation.
  - Why it indicates instability: Explain why this response appears extreme or significantly out of proportion to the situation.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

D.6 Poor mental health or experiencing mental health issues

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the client is showing signs of poor mental health, or is experiencing mental health issues. Do not flag notes that mention a diagnosed or self-reported mental

health condition. Only flag the note if there are clear and unambiguous indicators of poor mental health. Do not speculate, only use the text provided in the note. Do not flag the note based on behavoiur that could be explained by other factors, such as frustration or stress.

Use the following indicators to determine if the client is showing signs of poor mental health or experiencing mental health issues.

1. Erratic behaviour, such as talking to oneself or singing when it's not appropriate to do so. Minor unusual behaviour, such as being quiet, does not count
2. Evidence of a cognitive impairment (i.e. mentions of the client having a learning disability, memory issues, or attention deficit disorder)
3. The client talking about past trauma they have experienced
4. Explicit comments from shelter workers regarding the client's mental health, or statements expressing concern about the client showing signs of distress or withdrawal

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if the note contains clear indicators that the client is showing signs of poor mental health. Respond with <<NO>> if not. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.7 Has harmful personal relationships

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the client is involved in harmful personal relationships. This includes both current and past relationships. This applies when there is evidence of the client being involved in a relationship that may negatively impact their well-being or safety. Relationships can include family members, friendships, romantic partners, or other individuals outside of shelter staff. Relationships with other shelter residents can be considered, only if there is strong evidence that the relationship is ongoing and significantly harmful. Do not consider isolated incidents or temporary interactions. Do NOT flag notes that involve conflicts with shelter staff. Shelter staff should never be considered part of a harmful personal relationship.

Use the following indicators to determine if the client is involved in harmful personal relationships.

1. Descriptions of the client's relationship containing harmful interpersonal dynamics, such as recurring conflicts, violence, or manipulation
2. Patterns of negative interactions that suggest a sustained relationship rather than isolated incidents

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if there are clear indicators from the list above that the client is involved in harmful personal relationships. Respond with <<NO>> if not. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

D.8 Has interactions with EMS (emergency medical services)

You will be given a case note written by a shelter worker at a homeless shelter. This note reflects the shelter worker's observations and interactions with an individual seeking or receiving support at the shelter. The individual is referred to as the "client".

Your task is to determine if the note mentions the client having interactions with emergency medical services (EMS). It is important to only consider cases of confirmed direct involvement with EMS personnel, ambulance services, or emergency hospital visits. Do not consider cases where EMS or 911 was only considered, but not actually contacted or involved. For example, if a staff member thought about calling EMS but did not follow through, this should not be considered as an interaction.

Also, do not confuse EMS with other support staff or services. The following do NOT count as EMS.

- ACW (Adult Care Worker)
- CPS (Child Protection Services)
- DOAP (Downtown Outreach Addictions Partnership)
- LPN (Licensed Practical Nurse)

Use the following indicators to determine if the client has interactions with EMS:

1. Mentions of an ambulance or "EMS" being called for the client
2. Mentions of the client having an emergency visit to the hospital
3. Mentions of the client interacting with EMS workers or paramedics

Your response must include the following three components:

- YES or NO answer: Respond with <<YES>> if the note contains any of the indicators in the above list that show the client has iteractions with EMS. Respond with <<NO>> if not. Please enclose your answer in double angle brackets << >>.
- Justification: Provide a justification explaining the reasoning for your answer.
- Confidence Score: Provide a confidence score between 0 and 100, where a higher score reflects higher confidence in your answer.

All three components must be present in your response.

E Keyword Lists

The following keywords for each theme of focus was used as part of the simple keyword search algorithm for theme extraction.

T1 Harmful/Non Violent: Yell, Scream, Harmful, Racial, Slur, Uncomfortable, Stalk, Threat, Argue, Argument, Insult, Stole, Steal, Bully, Harass, Discriminate, Offend, Disrespect.

T2 Substance Use: Tinfoil, Drug, Witnessed Medication, Smoke, Smoking, Contraband, Etoh, Sober, Sobriety, Alcohol, Paraphernalia, Naloxone, Overdose, Under the Influence, Drinking, Drunk, Intoxicated, High, Addict, Weed, Injected, Injection, Withdrawal.

T3 Working to Find Housing: Housing Services, NSQ Application, Submitted Housing Application, Apartment Viewing, Visiting Apartment, Applied for Housing, NSQ, Found an Apartment, Found Housing, Viewing Appointment, Viewing Home, Viewing Apartment, Viewing House.

T4 High Instability: Outburst, Overreaction, Self-Regulating, Uncontrollable, Yelling at Staff, Unable to Self-Regulate, Highly Reactive, Volatile, Needed De-Escalation.

T5 Violence: Hit, Punch, Bite, Bitting, Throws Punch, Held Back, Holds Back, Holding Back, Altercation, Kick, Kicked, Push, Slap, Attack, Assault, Fistfight, Fight, Fought.

T6 Mental Health: Erratic, Singing, Talking to Themselves, Trauma, Cognitive Impairment, Mental Health Concerns, Paranoid, Mental Health Crisis, Hearing Voices, Depressed, Anxious, Cognitive Decline.

T7 Harmful Relationships: Abusive Relationship, Manipulative, Toxic, Ex-Partner Assaulted, Controlling, Unsafe Home Environment, Domestic Violence, Repeatedly Harmed By.

T8 Interactions with EMS: EMS, Ambulance, Emergency, Paramedic, First Responder, 911, Transported to Hospital.

F Phi-4 vs. Minitron Peformance Comparison on Theme Extraction

Table 5 compares the performance of Phi-4 and Minitron on theme extraction using single-theme prompts. Due to Minitron's consistently poor results, it was evaluated on only three themes before being excluded from further testing. All results are based on single-theme prompt configurations.

Theme	**Phi-4**				**Minitron**			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
T2 Substance Use	0.90	0.66	0.72	0.69	0.82	0.13	0.03	0.05
T3 Working to Find Housing	0.95	0.78	0.45	0.57	0.86	0.12	0.13	0.12
T5 Violence	0.95	0.49	0.85	0.62	0.74	0.02	0.10	0.04
T1 Harmful/Non-Violent	0.86	0.53	0.87	0.66	-	-	-	-
T4 High Instability	0.93	0.33	1.00	0.50	-	-	-	-
T6 Mental Health	0.85	0.18	0.80	0.29	-	-	-	-
T7 Harmful Relationships	0.88	0.10	0.71	0.17	-	-	-	-
T8 Interactions with EMS	0.96	0.47	1.00	0.64	-	-	-	-

Table 5: Performance comparison between Phi-4 and Minitron on theme extraction

G Full Labelling Analysis Results

Table 6 presents the labeling agreement analysis across all 34 themes annotated in the case notes.

Theme	Total	Unanimous (%)	Majority (%)	Lone (%)	Basic (%)
Displays harmful, but non-violent behaviour towards others	103	42.7%	60.2%	39.8%	85.5%
Substance use	80	53.8%	76.3%	23.8%	90.9%
Displays inappropriate behaviour towards others	61	21.3%	41.0%	59.0%	88.2%
Actively Working to Find Housing	49	44.9%	63.3%	36.7%	93.4%
Indicates a high level of mental or emotional instability	43	16.3%	32.6%	67.4%	91.1%
Indicates emotional distress	39	10.3%	35.9%	64.1%	91.4%
Displays physically violent behaviour towards others	32	40.6%	62.5%	37.5%	95.3%
Expressed Interest in Finding Housing	25	36.0%	72.0%	28.0%	96.1%
Poor mental health or experiencing mental health issues	25	24.0%	60.0%	40.0%	95.3%
Poor physical health, or experiencing physical health issues.	23	39.1%	65.2%	34.8%	96.6%
Indicates emotional stability	21	19.0%	42.9%	57.1%	95.8%
Has positive personal relationships	17	35.3%	88.2%	11.8%	97.3%
Has interactions with EMS (emergency medical services)	16	81.3%	93.8%	6.3%	99.3%
Hopeful about the future	16	25.0%	50.0%	50.0%	97.0%
Has interactions with the police or justice system	16	50.0%	68.8%	31.3%	98.0%
Medical Emergency or is Experiencing Medical Distress	15	6.7%	60.0%	40.0%	96.6%
Has a high belief in housing success or leaving shelter	14	7.1%	35.7%	64.3%	96.8%
Overdose	13	100.0%	100.0%	0.0%	100.0%
Has harmful personal relationships	13	38.5%	53.8%	46.2%	98.0%
Displays positive behaviours toward others	11	36.4%	63.6%	36.4%	98.3%
Is unable to provide self-care (e.g., poor hygiene, incontinence, hoarding)	10	50.0%	50.0%	50.0%	98.8%
Has a low belief in housing success or leaving shelter	7	0.0%	28.6%	71.4%	98.3%
Unable to work due to disability (or other reasons)	6	50.0%	50.0%	50.0%	99.3%
Currently employed	6	50.0%	83.3%	16.7%	99.3%
Making efforts to seek substance use treatment	6	33.3%	83.3%	16.7%	99.0%
Receiving substance use treatment	6	50.0%	50.0%	50.0%	99.3%
Currently unemployed	4	25.0%	25.0%	75.0%	99.3%
Not hopeful about the future	3	0.0%	0.0%	100.0%	99.3%
Perpetrator of sexual harassment	3	33.3%	100.0%	0.0%	99.5%
Mentions of mental health diagnosis	3	0.0%	33.3%	66.7%	99.3%
Isolated from others	2	0.0%	0.0%	100.0%	99.5%
Indicates suicide ideation	2	100.0%	100.0%	0.0%	100.0%
Receiving mental health treatment	1	0.0%	0.0%	100.0%	99.8%
Is able to provide strong self-care (e.g., cleanliness)	0				100.0%

Table 6: Full Label Agreement Summary by theme