Towards Interventions for Suicide Prevention with LLM Assistants in Social Work

Jaspreet Ranjit
USC Computer Science
jranjit@usc.edu

Justin Cho
USC Computer Science
hjcho@usc.edu

Myles Phung
USC Computer Science
maphung@usc.edu

John R. Blosnich
USC Suzanne Dworak-Peck
School of Social Work
blosnich@usc.edu

Swabha Swayamdipta
USC Computer Science
swabhas@usc.edu

Suicide is one of the leading causes of death in the US: one person dies by suicide every 11 minutes and one is hospitalized for a suicide attempt every 54 seconds (cdc, 2024b). Unfortunately, identifying risk factors contributing to suicide is difficult due to its complex, multi-dimensional causes, which includes factors such as mental health conditions and life disruptions (cdc, 2024a). Current prevention and intervention research has primarily focused on the clinical sector (Labouliere et al., 2018), although an equally important yet underexplored opportunity lies in the interactions between victims and nonclinical professionals (e.g. attorneys) (Chen and Roberts, 2021). For example, relationship failures and divorce are positively associated with suicide (Næss et al., 2021); however, there has been little to no engagement of family law attorneys in suicide prevention. Social work experts hypothesize that life disruptions are a significant factor contributing to suicide, and that identifying interactions between victims and nonclinical professionals could serve as an entry point for new preventive measures (Chen and Roberts, 2021). In this extended abstract, we will outline our ongoing efforts to *identify and characterize interactions* between suicide victims and nonclinical professionals *to assist social work researchers in the development of suicide interventions.* To this end, we propose employing large language models (LLMs) to extract relevant information from large-scale corpora to verify hypotheses posed by social work researchers.

The NVDRS Corpus The National Violent Death Reporting System (NVDRS) is created by the Centers for Disease Control and Prevention (CDC), containing information about 240K suicides (Murthy, 2024; Nazarov et al., 2019). It contains structured data (i.e., over 300 coded variables) and unstructured data in the form narratives, which are brief summaries about the decedent's death as documented by death investigator, medical examiners and coroners. Although NVDRS narratives contain helpful information about nonclinical interactions, the need for manual processing and the emotionally burdensome nature of the analysis limits its scale. Our work addresses the following research questions for NVDRS narratives:

RQ1 Can we reliably identify and characterize nonclinical interactions in NVDRS narratives?

RQ2 Can we enable hypothesis-driven analysis of annotated NVDRS narratives to assist social work researchers in data-driven development of intervention strategies?

1 Characterizing Nonclinical Interactions

Identifying interactions with nonclinical professionals is challenging because they are often obscured by references to life disruptions (e.g. custody battles) and thus overlooked by existing methods (Kafka et al., 2023). Furthermore, relying on keyword searches with professions such as "lawyer" or "attorney" may lead to false positives cases (e.g. victim is a lawyer by profession). In **preliminary work**, we collaborated with researchers from USC's School of Social Work to develop annotation guidelines to identify and characterize victims' interactions with attorneys, under different life disruptions (Halterman and Keith, 2024; Rytting et al., 2023). Our guidelines capture indirect references to life disruptions by classifying interactions into three categories: implicit (indirect interactions inferred from life disruptions), explicit (direct meetings), and no interaction. In our initial efforts, we collected human annotations for victim interactions with attorneys where we achieved an average inter-annotator agreement (Fleiss' κ) of 0.86 (Fleiss, 1971). Using our expert-developed guidelines in the prompt, we find that open-source models such as Llama-3-70B-Instruct achieve a Macro F_1 of 0.87, outperforming a guidelines-agnostic baseline (Macro F_1 of 0.55), on a test set of 150

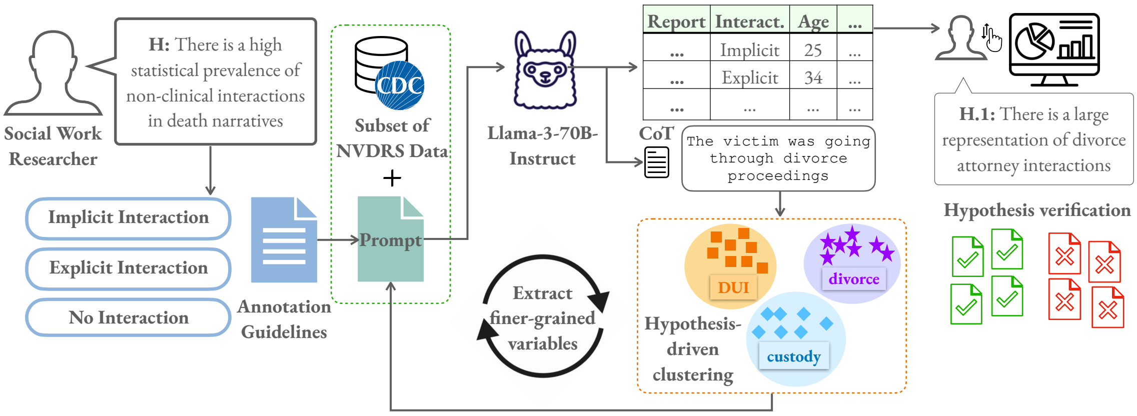
Social Work Researcher
H: There is a high statistical prevalence of non-clinical interactions in death narratives
Implicit Interaction
Explicit Interaction
No Interaction
Annotation Guidelines
Subset of NVDRS Data
CDC
+
Prompt
Llama-3-70B-Instruct
Report
Interact.
Age
...
Implicit
25
Explicit
34
CoT
The victim was going through divorce proceedings
Extract finer-grained variables
DUI
divorce
custody
Hypothesis-driven clustering
H.1: There is a large representation of divorce attorney interactions
Hypothesis verification

Figure 1: **Hypothesis-driven data annotation and analysis.** Given a hypothesis, social work researchers first define explanatory variables and annotation guidelines, we extract these variables using Llama-3-70B-Instruct on a small subset of data. We use the generated chains-of-thought to perform hypothesis-driven clustering and extract finer-grained variables. Finally, we verify hypotheses by counting the incidence of supporting narratives.

narratives. This indicates promising potential for further large-scale evaluation. We plan to expand our human annotations for large-scale evaluations and refine our guidelines to cover a wider range of nonclinical interactions (e.g., financial advisors).

2 Hypothesis-driven Analysis

Social work researchers developing suicide interventions are interested in exploring nuanced hypotheses (e.g., risk factors related to life disruptions) that are difficult to be addressed by the existing set of coded variables. Prior work on topic modeling and concept induction (Lam et al., 2024; Shankar et al., 2024) has identified high-level themes in unstructured data using a bottom-up approach (e.g., demographics of populations), but these are insufficient to test conceptual hypotheses that social work researchers develop prior to data analysis (Jun et al., 2022). To address these challenges, we propose to expand our hypothesis-driven, LLM-based framework (**RQ1**) to extract finer-grained information that can help develop new interventions as illustrated in Figure 1. For example, based on a hypothesis regarding the prevalence of legal interactions, social work researchers define explanatory variables (e.g., implicit, explicit, no interaction). We extract these variables from a small subset of data, along with LLMs' chain-of-thought (CoT) explanations (Wei et al., 2022), in **RQ1**. We hypothesize that CoT explanations from implicit interactions may yield key insights into the life disruptions leading to suicide, as they offer more context about the victim's circumstances. We propose clustering CoT explanations (in high-dimensional embedding space) and summarizing each cluster (again, using LLMs) to explain representative patterns in the nature of the interaction (e.g. custody battles, restraining order). We will extract finer-grained concepts by iteratively clustering CoTs enabling the investigation of a wider range of hypotheses. Finally, we will verify hypotheses by counting the incidence of supporting narratives.

We will evaluate two different types of LLM responses: (1) identity and character of interactions, and (2) cluster summaries indicating higher-order concepts that can assist in the verification of social work hypotheses, using both intrinsic and extrinsic measures. Our extrinsic measures include the evaluation of LLM generations by human annotators employed for (1) and social work researchers for (2). Our work aims to identify entry points for new interventions in "industries of disruption," a term coined by Dr. Blosnich to describe professionals assisting in challenges such as divorce and foreclosures (Blosnich et al., 2024). Our hypothesis-driven, human-in-the-loop data annotation and analysis pipeline addresses a pressing challenge for social science practitioners: reliably extracting data-driven insights from unstructured data to efficiently test concrete hypotheses. Our pipeline will enable social scientists to process large amounts of unstructured data that would otherwise be unmanageable due to the overwhelming cognitive load incurred from manual analysis in domains containing sensitive and explicit content. We hope our work will provide support in development of enhanced interventions for nonclinical professionals, and ultimately lower rates of suicide.

References

2024a. Risk and protective factors for suicide. Centers for Disease Control and Prevention. Accessed: 2024-10-17.

2024b. Suicide data and statistics. Centers for Disease Control and Prevention. Accessed: 2024-10-17.

John R Blosnich, Alexandra M Haydinger, Harmony Rhoades, and Susan M De Luca. 2024. Differences in beliefs about suicide by occupation in a representative sample of adults in the united states, general social survey 2002–2021. *Archives of suicide research*, 28(2):439–453.

Tony Chen and Karl Roberts. 2021. Negative life events and suicide in the national violent death reporting system. *Archives of suicide research*, 25(2):238–252.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Andrew Halterman and Katherine A Keith. 2024. Codebook llms: Adapting political science codebooks for llm use and adapting llms to follow codebooks. ArXiv.

Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. 2022. Hypothesis formalization: Empirical findings, software limitations, and design implications. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(1):1–28.

Julie M Kafka, Mike D Fliss, Pamela J Trangenstein, Luz McNaughton Reyes, Brian W Pence, and Kathryn E Moracco. 2023. Detecting intimate partner violence circumstance for suicide: development and validation of a tool using natural language processing and supervised machine learning in the national violent death reporting system. *Injury prevention*, 29(2):134–141.

Christa D Labouliere, Prabu Vasan, Anni Kramer, Greg Brown, Kelly Green, Mahfuza Rahman, Jamie Kammer, Molly Finnerty, and Barbara Stanley. 2018. "zero suicide"–a model for reducing suicide in united states behavioral healthcare. *Suicidologi*, 23(1):22.

Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Vivek Murthy. 2024. National strategy for suicide prevention.

Oybek Nazarov, Joseph Guan, Stanford Chihuri, and Guohua Li. 2019. Research utility of the national violent death reporting system: a scoping review. *Injury epidemiology*, 6:1–12.

Erik Oftedahl Næss, Lars Mehlum, and Ping Qin. 2021. Marital status and suicide risk: Temporal effect of marital breakdown and contextual difference by socioeconomic status. *SSM - Population Health*, 15:100853.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, et al. 2023. Towards coding social science datasets with language models. ArXiv:2306.02177.

Shreya Shankar, Aditya G. Parameswaran, and Eugene Wu. 2024. Docetl: Agentic query rewriting and evaluation for complex document processing.

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