Uncovering Intervention Opportunities for Suicide Prevention with Language Model Assistants

Jaspreet Ranjit[♡] **Hyundong J. Cho**[♠] **Claire J. Smerdon**[♡] **Yoonsoo Nam**[♡]
Myles Phung[♡] **Jonathan May**[♠] **John R. Blosnich**[♣] **Swabha Swayamdipta**[♡]
[♡]Thomas Lord Dept. of Computer Science, University of Southern California
[♠]Information Sciences Institute, University of Southern California
[♣]Suzanne-Dwork School of Social Work, University of Southern California
{jranjit, hjcho, blosnich, swabhas}@usc.edu

Abstract

The National Violent Death Reporting System (NVDRS) documents information about suicides in the United States, including free text narratives (e.g., circumstances surrounding a suicide). In a demanding public health data pipeline, annotators manually extract structured information from death investigation records following extensive guidelines developed painstakingly by experts. In this work, we facilitate data-driven insights from the NVDRS data to support the development of novel suicide interventions by investigating the value of language models (LMs) as efficient assistants to (a) data annotators and (b) experts. We find that LM predictions match existing data annotations about 85% of the time across 50 NVDRS variables. In the cases where the LM disagrees with existing annotations, expert review reveals that LM assistants can surface annotation discrepancies 38% of the time. Finally, we introduce a human-in-the-loop algorithm to assist experts in efficiently building and refining guidelines for annotating new variables by allowing them to focus only on providing feedback for incorrect LM predictions. We apply our algorithm to a real-world case study for a new variable that characterizes victim interactions with lawyers and demonstrate that it achieves comparable annotation quality with a laborious manual approach. Our findings provide evidence that LMs can serve as effective assistants to public health researchers who handle sensitive data in high-stakes scenarios.[1]

1 Introduction

Warning: This paper discusses topics of suicide and suicidal ideation, which may be distressing to some readers.

Each year, approximately 50,000 people in the United States fall victim to suicide [Cammack, 2024]. The Centers for Disease Control and Prevention (CDC) documents this information in the National Violent Death Reporting System (NVDRS)[2], which contains structured (e.g., more than 600 demographic and circumstance variables) and unstructured data (e.g., narrative summaries surrounding the circumstances of death) for more than 270,000 suicides (Figure 1; top left). Structured data in NVDRS is manually labeled by annotators (or NVDRS data abstractors[3]) using codebooks—precise definitions and annotation guidelines—developed painstakingly by experts. Given the sensitive

[1]Project page: https://dill-lab.github.io/interventions_lm_assistants/
[2]https://www.cdc.gov/nvdrs/about/index.html
[3]We use the term data abstractors and annotators interchangeably.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health Potential, Trust, and Policy Compliance.

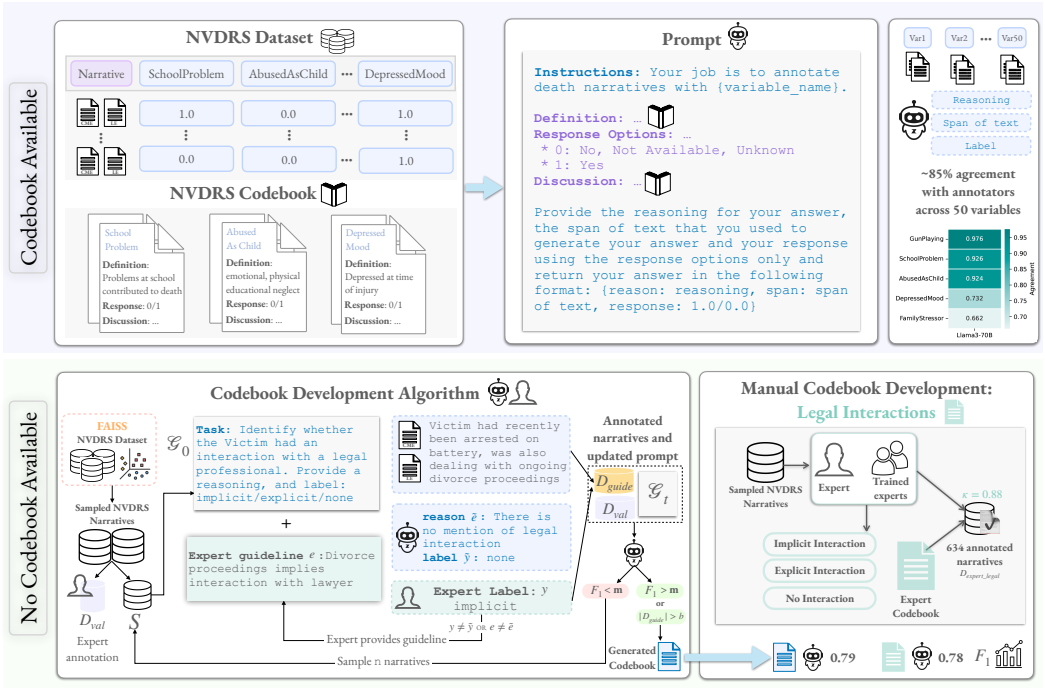

Figure 1: LM assistants can reduce the emotional burden of annotating NVDRS narratives when a codebook is available (top) and help experts efficiently develop new codebooks for novel variables (bottom). When a codebook is available, we incorporate it in the LM instruction for Chain-of-Thought predictions. For a new variable without an existing codebook, we propose an efficient, LM-assisted codebook development algorithm. Here experts focus on providing feedback for incorrect LM predictions, reducing manual codebook development time from weeks to hours (bottom right).

nature of the topic and the scale of NVDRS, a data annotator's job can be extremely demanding and emotionally taxing [Fincham et al., 2008, Murthy, 2024, Nazarov et al., 2019]. Moreover, variation in manual reporting and data abstraction leaves room for annotation inconsistencies [Dang et al., 2023].

Given its scale, NVDRS is an extremely valuable source for data-driven discovery of novel opportunities for suicide intervention. However, identifying such opportunities in NVDRS is not trivial due to complex and overlapping risk factors in the victim's circumstances, such as mental health and life disruptions [cdc, 2024, Wang et al., 2024]. To uncover actionable insights for new suicide interventions, experts must move beyond existing structured variables and extract new, data-driven evidence from qualitative narratives recorded as unstructured variables. However, systematically uncovering new variables require experts to manually analyze NVDRS narratives, develop a codebook for each variable, and abstractors to retroactively annotate each of the 270K cases.

In this work, we ask: *Can language models help public health and social work researchers improve efficiency and reduce the emotional labor of manual data annotation and analysis for suicide prevention research?* Our investigation is based on the promise language models have shown in social science [Rytting et al., 2023, Pangakis et al., 2023, Halterman and Keith, 2024, Ziems et al., 2024]. Specifically, we are interested in two research questions:
RQ1: Can a language model alleviate the burden of annotating variables with existing reference codebooks in NVDRS and surface discrepancies between the structured and unstructured data?
RQ2: Can a language model assist experts in the annotation of new variables that go beyond current NVDRS codebooks, potentially leading to new intervention opportunities?

For **RQ1**, we deploy LMs as annotation assistants to data abstractors to reduce the emotional burden of analyzing sensitive and explicit data, as well as surface potential inconsistencies between the narratives and the annotated structured variables. Our LM assistant achieves an average agreement of 85% with data abstractors across 50 variables (§2.3), making it a reliable peer validator for high-agreement variables. The assistant is also useful for low-agreement variables: from our expert review of six variables with the lowest agreement, we find that our LM assistant surfaces annotation inconsistencies for 38% of the instances where the LM disagrees with the data abstractor (§2.3.1).

We address **RQ2** by introducing a human-in-the-loop codebook development algorithm that helps experts iteratively develop codebook guidelines for new variables by providing feedback for incorrect LM predictions (§3.1). We first demonstrate the effectiveness of our algorithm by developing codebooks for variables with existing guidelines using only the variable name as our starting point. Our experiments show that our algorithm produces guidelines that enable LMs to achieve an average 80% agreement with data annotators, outperforming LMs conditioned on the existing NVDRS guidelines (75% agreement) for 12 variables (§3.2). Figure 1 illustrates our approach.

We further demonstrate that our algorithm is effective at characterizing new variables with a real-world case study: identifying (explicit or implicit) victim interactions with legal professionals (§3.3). We compare our algorithm to a fully manual codebook development process, where we collect annotations for characterizing victim interactions with legal professionals from a suicide prevention research expert on 634 NVDRS narratives [Halterman and Keith, 2024, Rytting et al., 2023]. We find that our algorithm reduces codebook development time from weeks to hours without compromising on annotation quality, demonstrating its potential to efficiently develop guidelines for a wider range of variables (e.g., interactions with other nonclinical professionals).

In summary, our results indicate the ability of language models to serve as efficient assistants that enable experts to mobilize the NVDRS dataset more effectively, and accelerate the identification of novel risk factors for developing data-driven interventions. However, despite the promising results we demonstrate in building LM assistants for experienced professionals in social work and public health, we strongly caution against forgoing expert supervision for LMs in sensitive domains given the high stakes involved.

2 Annotating Predefined Variables in NVDRS Narratives

We describe the NVDRS narratives and structured data (§2.1), and address our first research question (§2.2): Can LMs alleviate the burden of manual annotation for NVDRS death narratives? This setting includes established guidelines and human labels, allowing us to investigate whether LMs can follow preexisting annotation criteria.

2.1 The National Violent Death Reporting System (NVDRS)

NVDRS contains case records for 270,000 suicide cases between 2003-2019 in 42 US states, the District of Columbia, and Puerto Rico [Liu, 2023, Wilson, 2022]. Each case is potentially characterized by more than 600 structured variables which include demographics, and circumstances surrounding the victim's death [Liu, 2023]. Suicide circumstance (e.g., eviction or loss of home) and crisis (e.g., events occurring within two weeks of death) variables are annotated using four data sources: death certificates, coroner/medical examiner (CME) records, law enforcement (LE) records, and crime laboratory records. Narratives are recorded using information from the CME/LE records [Paulozzi et al., 2004]. Data abstractors annotate cases by following an extensive codebook containing definitions and guidelines for each variable. However, the likelihood of annotation discrepancies is high due to the demanding nature of the task [Wang et al., 2023, 2024]; further, only 5% of NVDRS annotations are validated by two annotators [Liu, 2023]. For our study, we consider the subset of variables that have been manually annotated based on information from CME/LE records. Since the narratives are also derived from these records [Paulozzi et al., 2004], they should capture the same information as the structured data. However, variations in recording practices can lead to discrepancies where some details in the structured data may be missing from the narrative, and vice versa, resulting in inconsistencies between the two sources.

2.2 Experimental Setup

We use open-weight LMs that vary in model family and parameter size, as annotation assistants to label 36 circumstance and 14 crisis variables for CME and LE death narratives. We select 50 binary variables based on whether they are annotated using CME/LE narratives and appear in at least 300 cases to ensure reliable evaluation.[4] We curate LM prompts for each variable by adapting the NVDRS codebook guidelines in a standardized format with instructions, definitions, response options, and

[4]The remaining demographic/toxicology variables are derived from death certificates and lab reports which we do not have access to.

discussion, as shown in Figure 1 (top panel). We concatenate the CME and LE narratives to form the input, and generate Chain-of-Thought reasoning (CoT; Wei et al. [2022]), a span of supporting text, and a label in a zero-shot setting.

For evaluation, we consider a balanced dataset ($|D_{\text{balanced}}| = 25{,}000$) where we sample 500 narratives per variable with equal representation across 0/1 classes, to address the high class imbalance for each variable across narratives. We also evaluate on a random sample of narratives ($|D_{\text{random}}| = 1{,}000$), which more closely reflects real-world deployment conditions where class distributions are heavily skewed. In both settings, we report agreement[5] with data abstractor annotations per variable.

2.3 LMs validate data abstractor annotations

Table 1 shows the average agreement of LM predictions with abstractor annotations for D_{balanced} across 50 variables, with 95% confidence intervals (bootstrapped with 10K iterations). We find that `Llama-3-70B` achieves the highest average agreement: 85% for D_{balanced} and 82% for D_{random}, outperforming all other models on at least 38 out of 50 variables.[6] All comparisons are statistically significant ($p < 0.0125$ under Bonferroni correction). Figure 2 illustrates a more detailed overview for a subset of 20 variables from D_{balanced}. We observe that all models have low agreement with abstractors for mental health- and emotional state- related variables (e.g., `SuicideThoughtHistory`) and relatively higher agreement for firearm-related variables (e.g., `GunPlaying`), which have concrete, observable events with explicit lexical cues. We report further detailed results in the Appendix in Table 3.

Model	Mean Agreement$_{S.D.}$, (95% CI)
Llama-3-70B	$0.85_{0.09}$, **[0.82, 0.87]**
Qwen2.5-14B	$0.79_{0.10}$, [0.76, 0.82]
Qwen2.5-7B	$0.73_{0.08}$, [0.70, 0.75]
Mistral-7B	$0.73_{0.07}$, [0.71, 0.75]
Llama-3-8B	$0.71_{0.09}$, [0.69, 0.74]

Table 1: Mean agreement, 95% CI, and standard deviation across 50 variables for different models on D_{balanced}. `Llama-3-70B` achieves the highest agreement of 85% with data annotators.

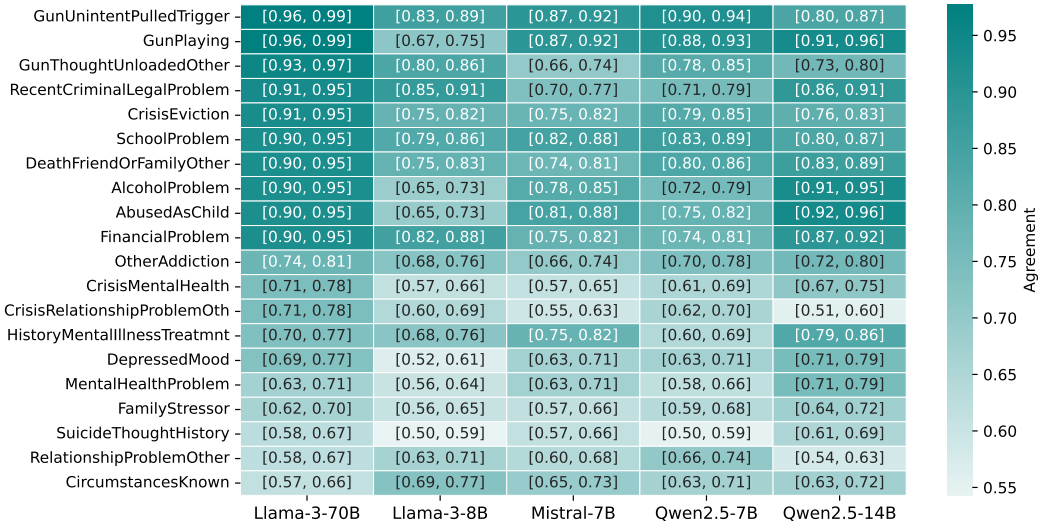

Figure 2: Per variable agreement (95% confidence intervals; bootstrapped with 10K iterations) for 20 highest and lowest agreement variables. All models have low agreement with abstractors for mental health-related variables and relatively higher agreement for firearm-related variables. Complete results for all 50 variables are included in Figure 5 in the Appendix.

To better understand the shortcoming of LMs on low-agreement variables, we conduct a targeted qualitative analysis for two of them: `MentalHealthProblem` and `DepressedMood`. A manual review of `Llama-3-70B` predictions reveal two recurring failure modes. First, in false negative cases, references to mental health problems or depressed moods appear in only one of the two concatenated

[5]Agreement is computed as the ratio of matching labels to total number of instances.

[6]The complete results on D_{random} are included in Table 4 in Appendix B.

narratives (i.e., CME or LE), reducing the model's ability to detect the variable in longer context narratives. Second, in false positive cases, the model often incorrectly infers mental health problems based on circumstantial evidence such as financial problems, bankruptcy, substance use, interpersonal conflict, or life stressors. This raises concerns about the reliability of LMs for mental health related variables.

2.3.1 Surfacing Annotation Inconsistencies Through LM Disagreements with Abstractors

Prior work by Wang et al. [2023, 2024] reveal inconsistently annotated NVDRS variables among data abstractors. For example, Wang et al. [2024] apply a cross-validation approach to identify cases where suicide circumstances described in free-text narratives are not reflected in the structured data. Given their findings, we hypothesize that the disagreement between the LM assistant and the abstractor may surface inconsistencies between the narratives and structured data. For six variables with low LM (`Llama-3-70B`) agreement with the abstractor (i.e. `CircumstancesKnown`, `RelationshipProblemOther`, `FamilyStressor`, `MentalHealthProblem`, `DepressedMood`, `HistoryMentalIllnessTreatment`), we sample 150 narratives where the LM and abstractor disagree and another 150 where they agree. For each of these settings, we ask for a second opinion from a suicide prevention research expert. In cases where the LM and abstractor disagree, the expert finds that the original annotation is inconsistent with the information contained in the narrative 38% of the time, compared to only 13% of the time when the LM and abstractor are in agreement. We conclude from a bootstrap hypothesis test for equality of means [Efron and Tibshirani, 1994] that this difference is statistically significant ($p < 0.05$), indicating the potential for our assistant to surface annotation inconsistencies between the narratives and structured data.

Based on our findings, we recommend practitioners to abstain from relying on LM-generated annotations for low-agreement variables. In practice, models should flag high uncertainty predictions or abstain altogether, prioritizing safety over coverage to minimize the risk of misleading outputs in expert workflows. Furthermore, for high-agreement variables, we recommend that annotators validate LM-generated annotations where reviewing the model's predicted label and chain-of-thought is considerably less labor-intensive than annotating narratives from scratch. Our objective is to reduce the burden on annotators, while preserving human oversight rather than eliminating it.

3 Characterizing New Variables in NVDRS Narratives

The NVDRS codebook does not exhaustively cover all contextual information contained in the death narratives [Dang et al., 2023]; analyses that hinge on information not yet coded as structured data have to go through a rigorous annotation process for all 270K narratives. Although the codebook guidelines have continuously evolved to expand the scope of NVDRS variables, the updates are infrequent and incremental [Steenkamp et al., 2006]. Therefore, it is paramount that new variables are coded efficiently to enable data-driven research, test novel hypotheses in prevention research, and inform intervention strategies [Blair et al., 2016].

To address these challenges, we introduce an algorithm to develop codebook guidelines for new variables (§3.1) using expert (human or LM) feedback. Given the reference NVDRS codebook and human labels for existing NVDRS variables, we then validate the effectiveness of our algorithm by generating codebooks for 12 NVDRS variables using LM feedback (§3.2). This approach achieves comparable or better agreement with data abstractors compared to using the reference codebook, validating its effectiveness. We further validate our approach by applying our algorithm to a real-world case study with human expert feedback to characterize a new variable: victim interactions with legal professionals (§3.3).

3.1 Codebook Development Algorithm for Characterizing Variables

Our codebook development algorithm (Algorithm 1) leverages a language model's ability to surface evidence from narratives that can validate expert hypotheses on new variables affecting suicide. Our algorithm efficiently synthesizes codebook guidelines and iteratively refines them based on feedback from either an LM in a fully automated setup (§3.2) or a human expert (§3.3). We repeat this process until the language model achieves a pre-defined target performance on an evaluation set using the refined codebook.

Algorithm 1: Codebook Development Algorithm

Input: π_θ: LM, $\mathcal{D} = \{x_i \mid 1 \leq i \leq N\}$: full set of unannotated NVDRS narratives, $\mathcal{D}^0_{guide} = \emptyset$: validated NVDRS narratives, $\mathcal{D}_{val}$: annotated NVDRS narratives, $\mathcal{G}_0$: initial guideline prompt, $\mathcal{U}$: guideline update prompt, m: target accuracy, k: minimum evaluation set size, b: budget, t: iteration index, $\mathcal{F}$: feedback

Output: $\mathcal{D}_{\text{LM}}$: LM-annotated data

$t \leftarrow 0$
while *True* **do**
  $\mathcal{S} \sim \mathcal{D} \setminus \mathcal{D}^{t-1}_{guide}, |S| = k$ // Sample from $\mathcal{D}$
  $\mathcal{I} \leftarrow \emptyset$
  for $x \in \mathcal{S}$ **do**
    $\tilde{y}, \tilde{e} \sim \pi_\theta(\mathcal{G}_t(x))$ // Generate LM label and reasoning
    $y, e \sim \mathcal{F}(x, \tilde{y}, \tilde{e})$ // Feedback provides correct label and reasoning
    if $\tilde{y} \neq y$ ***or*** $\tilde{e} \neq e$ **then**
      $\mathcal{I} = \mathcal{I} \cup \{(x, \tilde{y}, y, \tilde{e}, e)\}$ // Collect LM errors
    $\mathcal{D}^t_{guide} = \mathcal{D}^{t-1}_{guide} \cup \{(x, y, e\}$
  $\mathcal{G}_{t+1} \leftarrow \pi_\theta(\mathcal{U}(\mathcal{G}_t, I))$ // Update guideline based on LM errors
  $t \leftarrow t + 1$
  if $\mathcal{D}^t_{guide} \neq \emptyset$ **then**
    $acc \leftarrow \frac{\sum_{(x,y,e) \in \mathcal{D}_{val}} \mathbb{1}[\pi_\theta(\mathcal{G}_t(x)) = y]}{|\mathcal{D}_{val}|}$ // Check for stopping criteria
    if $acc \geq m$ & $|\mathcal{D}^t_{guide}| \geq k$ ***or*** $|\mathcal{D}^t_{guide}| > b$ **then**
      break
$\mathcal{D}_{\text{LM}} \leftarrow \{(x, y, e) \mid y, e \sim \pi_\theta(\mathcal{G}_t(x)), x \in \mathcal{D} \setminus \mathcal{D}^t_{guide}\}$ // Annotate $\mathcal{D}$ with final $\mathcal{G}$
return $\mathcal{D}_{\text{LM}}$

First, we ask the expert to annotate a subset of narratives, $\mathcal{D}_{val}$, which serves as a held-out validation set and is not used for guideline development. Experts annotate until $\mathcal{D}_{val}$ contains at least j instances per class. Given NVDRS contains 270K narratives, the likelihood of finding a true positive is quite low depending on the expert hypothesis (e.g., occurrence of legal interactions in 270K NVDRS narratives is 15%, as shown in Figure 8 in Appendix F). As a result, we initialize $\mathcal{D}$ by upsampling instances based on expert-defined keywords relevant to the variable of interest, using similarity search with FAISS [Douze et al., 2024]. As illustrated in Figure 1 and specified in Algorithm 1, we set our initial codebook prompt $\mathcal{G}_0$ using a template that only includes the variable name and response options, i.e. labels (see Table 8 in Appendix F). With this prompt, the LM[7] generates a chain-of-thought reasoning $\tilde{e}$ and label $\tilde{y}$ for n sampled narratives.

We consider random and coverage-based sampling [Gupta et al., 2023] to select n samples to annotate in each iteration. The latter selects n narratives that are most dissimilar to those seen in prior iterations ($\mathcal{D}^t_{guide}$) so that experts can avoid redundant cases. We detail our coverage-based sampling in Appendix D.

For each incorrect LM prediction, an expert provides the corrected label y, and a free-text rationale e explaining their label. The current $\mathcal{G}_t$ is then updated by concatenating e to the prompt. Each expert-validated annotation is added to the growing evaluation set $\mathcal{D}^t_{guide}$ and the LM's performance is evaluated using $\mathcal{G}_t$ on $\mathcal{D}^t_{guide}$ and $\mathcal{D}_{val}$ upon each update to $\mathcal{G}_t$. This process is repeated until performance on $\mathcal{D}_{val}$ exceeds a specified target performance m, or the number of expert-validated annotations in $\mathcal{D}^t_{guide}$ exceeds a predefined budget b. The budget refers to the maximum number of samples the expert can annotate in the duration of the algorithm. We set a minimum size k for $|\mathcal{D}^t_{guide}|$ to account for the target m being met prematurely. A simplified overview of the process is shown in the bottom of Figure 1.

3.2 Simulated Codebook Development: Existing NVDRS Variables

To validate the usefulness and generalizability of our codebook development algorithm, we first evaluate it via simulations on a subset of existing variables, treating them as new variables that are not yet characterized. To scale our simulations to multiple variables, we use LM feedback (y, e) in lieu

[7] LM refers to `Llama-3-70B` for all codebook development experiments.

of a human expert. We have reference labels y for existing NVDRS variables (from §2). However, e is not provided. To address this, we consider the LM CoT generated in §2 as a synthetic proxy for feedback and use it as e. We apply our codebook development algorithm in this simulated setting to generate codebooks for 12 variables, using a subsample $\mathcal{D}$ of 150 randomly sampled narratives per variable, balanced across 0/1 classes. The 12 variables are chosen to cover varying degrees of average LM-annotator agreement as reported in §2 (four from low agreement (0.6-0.7), four from medium agreement (0.7-0.8) and four from high agreement (0.8-0.9)) so that we can evaluate how well the generated codebooks perform across different levels of annotation difficulty. In our simulated setting, the algorithm terminates after reaching the annotation budget (b=150 narratives). We include our hyperparameter settings in Table 6 in Appendix E.

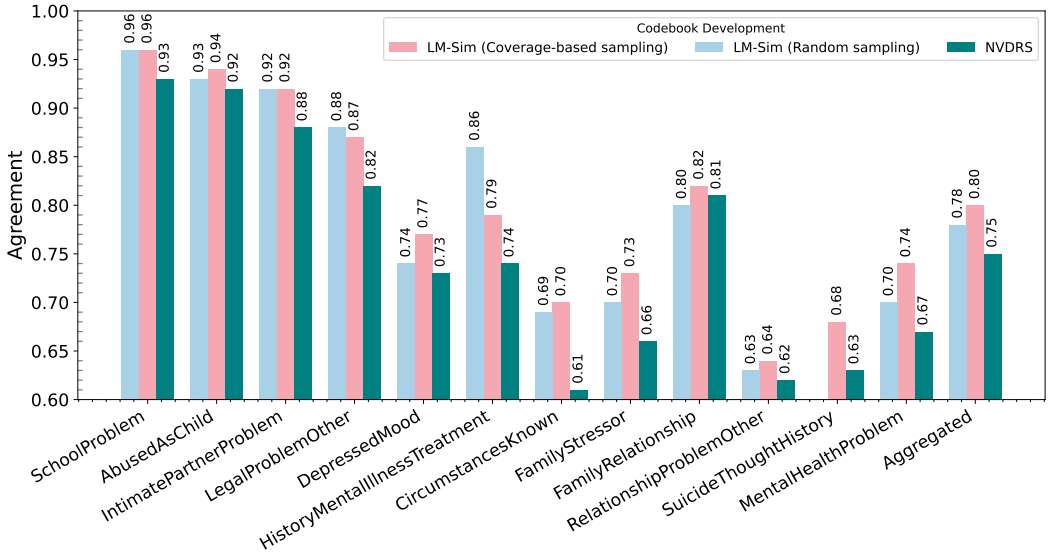

Figure 3: `Llama-3-70B` performance on D_{balanced} using the generated codebooks in the prompt (LM-Sim) for random and coverage-based sampling on 12 variables. We compare this to the performance achieved using the reference NVDRS codebook in the prompt for the selected variables. Our generated codebooks from the simulated setting are more effective for an LM than using the NVDRS codebook in the prompt for all 12 variables as shown by the 'Aggregated' agreement ($\sim$0.80 for LM-Sim vs $\sim$0.75 for NVDRS).

We measure the effectiveness of the generated codebooks by evaluating LM performance on our held out test set (D_{balanced} from §2) using the generated codebooks in the prompt. `Llama-3-70B` is used to generate the codebooks and perform the annotation task for evaluation. We compare this to the performance achieved using the NVDRS codebook in the prompt in §2 for the selected variables.

Paired t-tests ($p < 0.025$ under Bonferroni correction) across 12 variables show that our generated codebooks significantly outperformed the NVDRS codebook for random and coverage-based sampling as show in Figure 3. Moreover, the maximum accuracy on $\mathcal{D}^t_{guide}$ is reached between 10-15 iterations, as shown in Figure 6 (see Figure 7 in Appendix E for results on all variables). We observe that our generated codebooks provide finer-grained instructions to the model, and include examples from the narratives as guidelines which helps address ambiguities that might be missed by the NVDRS codebook as shown in Table 10 in the Appendix G. Although codebook development must never be fully automated, our findings suggest that LMs can help experts surface contextual clues embedded in long narratives to refine and augment NVDRS codebooks for existing variables. Furthermore, our simulated setting enables scalable evaluation of our codebook development algorithm without requiring extensive expert feedback, allowing us to assess its effectiveness across variables with varying levels of annotation difficulty.

3.3 Human-in-the-Loop Codebook Development: Case Study of Victim-Lawyer Interactions

To further showcase the effectiveness of our algorithm, we explore its use for a case study on a new variable where feedback comes from a human expert, i.e. human-in-the-loop (`HitL`). Suicide preven-

tion experts have identified legal professionals as part of a broader set of non-clinical 'industries of disruption' (e.g. financial advisors, homeless shelters) that frequently interact with at-risk individuals. Yet, prevention efforts largely focus on the clinical sector [Labouliere et al., 2018, Wang et al., 2023, Consoli et al., 2024, Guevara et al., 2024, Lybarger et al., 2023, Ralevski et al., 2024, Xu et al., 2024, Kafka et al., 2023], which primarily engages in *remedial* care, leaving non-clinical professions (e.g. attorneys) under-explored. NVDRS lacks structured variables to capture these interactions, despite their potential importance: a recent survey found that over 40% of attorneys had a client die by suicide, 70% had concerns about a client's suicide risk, but 65% had never received suicide prevention training [Blosnich et al., 2024]. Suicide prevention experts have hypothesized that the identification of new variables, such as the presence of victim / nonclinical professional interaction, can lead to novel suicide prevention measures [Chen and Roberts, 2021].

For our case study, we focus specifically on victim interactions with legal professionals (e.g. lawyers, attorneys). 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 positive cases (e.g., when the victim is a lawyer by profession). Thus, interactions with legal professionals can be characterized by three classes: implicit (indirect interactions inferred from life disruptions), explicit (direct interactions), and no interaction.

3.3.1 `HitL` Codebook Development

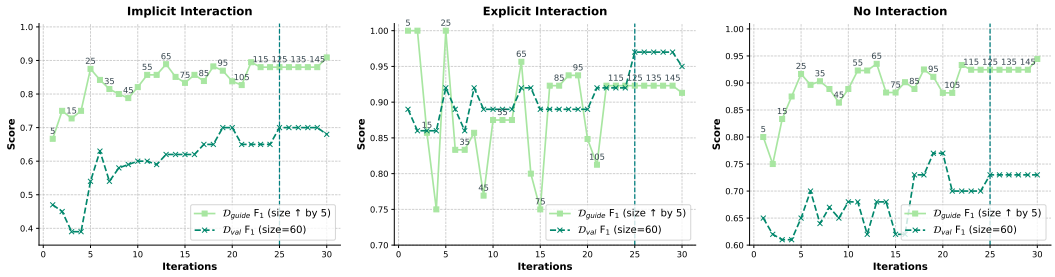

Figure 4: `Llama-3-70B` performance on $\mathcal{D}_{val}$ and $\mathcal{D}^t_{guide}$ across 30 iterations for `HitL` codebook development for detecting victim-lawyer interactions. Data labels for $\mathcal{D}_{guide}$ represent the cumulative size of $\mathcal{D}^t_{guide}$ at each iteration t. Max overall macro F_1 on $\mathcal{D}_{val}$ is reached by the 25th iteration (Macro F_1 of 0.8).

We apply our codebook development algorithm, this time with human expert feedback in the `HitL` pipeline, to efficiently develop guidelines for victim interactions with legal professionals. Given a budget of 150 instances, and batch size of 5, we first subsample narratives using FAISS [Douze et al., 2024] with the following keywords: '`lawyer, attorney`'. We start with a baseline prompt ($\mathcal{G}_0$) as shown in Figure 1, and use `Llama-3-70B` for guideline development. In our pilot, the expert exhausts the entire budget.

We plot the F_1 on $\mathcal{D}^t_{guide}$ and $\mathcal{D}_{val}$ across iterations for all three interaction types in Figure 4.[8] As expected, the performance is unstable in the first few iterations given that $\mathcal{D}^t_{guide}$ only contains a few narratives. However, we observe that peak performance is reached by the 25th iteration (Macro F_1 on $\mathcal{D}_{val}$ was 0.8).[9] We hypothesize that performance on explicit interactions is less stable because most of the guidelines pertain to implicit interactions, as shown in Table 9, resulting in limited instruction to identify explicit interactions.

3.3.2 Baseline: Manual Codebook Development

As a baseline to our LM-assisted codebook development algorithm, we collaborate with a suicide prevention research expert to manually develop a codebook to identify and characterize victims' interactions with legal professionals, under different life disruptions [Halterman and Keith, 2024, Rytting et al., 2023]. The codebook development process involves a subset of the authors: one

[8] Please see Table 7 in Appendix F for $\mathcal{D}^{30}_{guide}$, $\mathcal{D}_{val}$ (j=20) and $D_{\text{expert\_legal}}$ class distributions.

[9] Please see Table 6 in Appendix E for further discussion on hyperparameter selection for our case study.

social work expert (suicide prevention professor) and two experts-in-training (CS PhD students). who independently analyze and annotate a sample of 150 narratives to delineate precise definitions and guidelines for legal interactions. This process is repeated twice on different sets of narratives until no further refinements were made to the guidelines. Using our codebook, the team of experts annotates 634 narratives ($D_{\text{expert\_legal}}$) with high inter-annotator agreement (Krippendorff's $\alpha = 0.88$; [Krippendorff, 1970]).

3.3.3 Evaluating HitL Codebook Development Against our Manual Baseline

In Table 2, we compare performance across three prompts: (1) No Codebook ($\mathcal{G}_0$), (2) Expert Codebook ($\mathcal{G}_{\text{expert}}$) (§3.3.2), and (3) HitL Codebook ($\mathcal{G}_{25}$), and we evaluate on $D_{\text{expert\_legal}}$. We observe that $\mathcal{G}_{25}$ performance (Macro F_1) is on par with $\mathcal{G}_{\text{expert}}$ across all models suggesting that our guidelines can be generalized beyond the model that was used to develop the guidelines in the HitL pipeline.

Model	$\mathcal{G}_0$	$\mathcal{G}_{25}$	$\mathcal{G}_{\text{expert}}$
Meta-Llama-3-70B	0.57	0.80	0.78
Qwen2.5-32B	0.68	0.76	0.77
Qwen2.5-14B	0.63	0.75	0.77

Table 2: Macro F_1 reported on $D_{\text{expert\_legal}}$ using no codebook ($\mathcal{G}_0$), the HitL generated codebook at the 25th iteration ($\mathcal{G}_{25}$), and the expert codebook ($\mathcal{G}_{\text{expert}}$). The generated codebook achieves performance on par with the expert codebook.

Our HitL codebook development pilot takes 3 hours to complete (compared to several weeks in the manual setting), and our results show that we do not compromise on annotation quality on the new variable (victim-legal interactions). These findings support the application of our algorithm for experts to verify novel hypotheses, such as victim interactions with additional nonclinical stakeholders (e.g. financial institutions, housing shelters) [Sinyor et al., 2024].

A promising direction for future work is to explore how thematic saturation might be approximated by investigating alternative stopping criteria. In this context, thematic saturation refers to the point at which the guidelines sufficiently capture all relevant cases observed for a given variable [Glaser and Strauss, 1967], where annotating additional narratives does not result in meaningful changes to the guideline content. While thematic saturation cannot be guaranteed with our algorithm, we recommend practitioners take into account the subjectivity of the variable and the manual effort involved when determining the budget, target performance and the minimum size of $\mathcal{D}^t_{guide}$ [Glaser and Strauss, 1967]. For instance, subjective variables may demand more iterations to refine guidelines and thus, exhibit slower convergence in model performance. Alternative stopping criteria might include enforcing a minimum number of labeled examples and target performance per class to ensure balanced outcomes across all classes rather than relying solely on overall performance metrics. Another alternative is to track how many consecutive iterations occur without new guideline additions, and to stop once this number surpasses a predefined threshold. Future work is needed to further explore the trade-off between efficiency and performance for additional variables.

4 Conclusion

We introduce an LM assistant to validate expert annotations for NVDRS variables, and to aid experts in developing guidelines for novel variables going beyond the current NVDRS codebook. We show that our LM assistant achieves high agreement with data abstractors in annotating narratives with NVDRS variables and surfaces annotation inconsistencies for a subset of variables. We also introduce a human-in-the-loop codebook development algorithm to assist experts in characterizing new variables. We validate our algorithm with existing NVDRS variables and show that LM annotation performance using our generated codebook is on par with LM performance when using the reference codebook. We test our algorithm for a real world case study to characterize a new variable: victim interactions with legal professionals. Our findings motivate the use of our LM assistant in practice as additional validation for annotators who are annotating sensitive and explicit data. We hope practitioners use our framework to accelerate the discovery of data-driven insights for socially sensitive tasks, ultimately leading to new opportunities for interventions.

Acknowledgments and Disclosure of Funding

Support for this work came from a research award from the National Institute of Mental Health (DP2MH129967; PI Blosnich). The content is solely the responsibility of the authors and does not necessarily represent the views of the funders or institutions.

References

Risk and protective factors for suicide. Centers for Disease Control and Prevention, 2024. URL https://www.cdc.gov/suicide/risk-factors/index.html. Accessed: 2024-10-17.

Alina Arseniev-Koehler, Jacob Gates Foster, Vickie M Mays, Kai-Wei Chang, and Susan D Cochran. Aggression, escalation, and other latent themes in legal intervention deaths of non-hispanic black and white men: Results from the 2003–2017 national violent death reporting system. *American journal of public health*, 111(S2):S107–S115, 2021.

Alina Arseniev-Koehler, Susan D Cochran, Vickie M Mays, Kai-Wei Chang, and Jacob G Foster. Integrating topic modeling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences*, 119(10):e2108801119, 2022.

Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Felipe Zambrano, Alexandra L Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, et al. Chatgpt for education research: exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*, pages 134–149. Springer, 2024.

Janet M Blair, Katherine A Fowler, Shane PD Jack, and Alexander E Crosby. The national violent death reporting system: overview and future directions. *Injury prevention*, 22(Suppl 1):i6–i11, 2016.

John Blosnich, Jeanne Ward, Alexandra Haydinger, Melissa Perkins, and Susa De Luca. Industries of disruption: New avenues for upstream suicide prevention. In *APHA 2024 Annual Meeting and Expo*. APHA, 2024.

Alison L Cammack. Vital signs: Suicide rates and selected county-level factors—united states, 2022. *MMWR. Morbidity and Mortality Weekly Report*, 73, 2024.

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

Bernardo Consoli, Xizhi Wu, Song Wang, Xinyu Zhao, Yanshan Wang, Justin Rousseau, Tom Hartvigsen, Li Shen, Huanmei Wu, Yifan Peng, et al. Sdoh-gpt: Using large language models to extract social determinants of health (sdoh). *arXiv preprint arXiv:2407.17126*, 2024.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*, 2023.

Linh N Dang, Eskira T Kahsay, LaTeesa N James, Lily J Johns, Isabella E Rios, and Briana Mezuk. Research utility and limitations of textual data in the national violent death reporting system: a scoping review and recommendations. *Injury epidemiology*, 10(1):23, 2023.

Nabarun Dasgupta. Ghost in the machine: The emotional gravity of conducting mortality research, 2021.

Judy E Davidson, Gordon Ye, Felicia Deskins, Heather Rizzo, Christine Moutier, and Sidney Zisook. Exploring nurse suicide by firearms: A mixed-method longitudinal (2003–2017) analysis of death investigations. In *Nursing forum*, volume 56, pages 264–272. Wiley Online Library, 2021.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Ben Fincham, Jonathan Scourfield, and Susanne Langer. The impact of working with disturbing secondary data: Reading suicide files in a coroner's office. *Qualitative Health Research*, 18(6): 853–862, 2008.

Barney G. Glaser and Anselm L. Strauss. *Discovery of Grounded Theory Strategies for Qualitative Research*. AldineTransaction, London, 1967.

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

Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.930. URL https://aclanthology.org/2023.findings-emnlp.930/.

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

Lily Johns, Chuwen Zhong, and Briana Mezuk. Understanding suicide over the life course using data science tools within a triangulation framework. *Journal of psychiatry and brain science*, 8(1): e230003, 2023.

Julie M Kafka, Mike D Fliss, Pamela J Trangenstein, Luz McNaughton Reyes, Brian W Pence, and Kathryn E Moracco. 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, 2023.

Andrew Katz, Gabriella Coloyan Fleming, and Joyce Main. Thematic analysis with open-source generative ai and machine learning: A new method for inductive qualitative codebook development. *arXiv preprint arXiv:2410.03721*, 2024.

Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70, 1970.

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

Grace S Liu. Surveillance for violent deaths—national violent death reporting system, 48 states, the district of columbia, and puerto rico, 2020. *MMWR. Surveillance Summaries*, 72, 2023.

Kevin Lybarger, Oliver J Bear Don't Walk IV, Meliha Yetisgen, and Özlem Uzuner. Advancements in extracting social determinants of health information from narrative text, 2023.

Vivek Murthy. National strategy for suicide prevention. 2024.

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

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*, 2023.

Leonard J Paulozzi, J Mercy, Lorraine Frazier, and J Lee Annest. Cdc's national violent death reporting system: background and methodology. *Injury prevention*, 10(1):47–52, 2004.

Mike Perkins and Jasper Roe. The use of generative ai in qualitative analysis: Inductive thematic analysis with chatgpt. *Journal of Applied Learning and Teaching*, 7(1), 2024.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5:BII–S9042, 2012.

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

Jaspreet Ranjit, Brihi Joshi, Rebecca Dorn, Laura Petry, Olga Koumoundouros, Jayne Bottarini, Peichen Liu, Eric Rice, and Swabha Swayamdipta. OATH-frames: Characterizing online attitudes towards homelessness with LLM assistants. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13033–13059, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.724. URL https://aclanthology.org/2024.emnlp-main.724/.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Ian RH Rockett, Eric D Caine, Hilary S Connery, Gail D'Onofrio, David J Gunnell, Ted R Miller, Kurt B Nolte, Mark S Kaplan, Nestor D Kapusta, Christa L Lilly, et al. Discerning suicide in drug intoxication deaths: Paucity and primacy of suicide notes and psychiatric history. *PLoS one*, 13(1): e0190200, 2018.

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

Mark Sinyor, Morton Silverman, Jane Pirkis, and Keith Hawton. The effect of economic downturn, financial hardship, unemployment, and relevant government responses on suicide. *The Lancet Public Health*, 9(10):e802–e806, 2024.

Malinda Steenkamp, Lorraine Frazier, N Lipskiy, M DeBerry, S Thomas, L Barker, and Debra Karch. The national violent death reporting system: an exciting new tool for public health surveillance. *Injury prevention*, 12(suppl 2):ii3–ii5, 2006.

Robert H Tai, Lillian R Bentley, Xin Xia, Jason M Sitt, Sarah C Fankhauser, Ana M Chicas-Mosier, and Barnas G Monteith. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23:16094069241231168, 2024.

Song Wang, Yifang Dang, Zhaoyi Sun, Ying Ding, Jyotishman Pathak, Cui Tao, Yunyu Xiao, and Yifan Peng. An nlp approach to identify sdoh-related circumstance and suicide crisis from death investigation narratives. *Journal of the American Medical Informatics Association*, 30(8): 1408–1417, 2023.

Song Wang, Yiliang Zhou, Ziqiang Han, Cui Tao, Yunyu Xiao, Ying Ding, Joydeep Ghosh, and Yifan Peng. A natural language processing approach to detect inconsistencies in death investigation notes attributing suicide circumstances. *Communications Medicine*, 4(1):199, 2024.

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

Rebecca F Wilson. Surveillance for violent deaths—national violent death reporting system, 42 states, the district of columbia, and puerto rico, 2019. *MMWR. Surveillance Summaries*, 71, 2022.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 75–78, 2023.

Richard Li Xu, Song Wang, Zewei Wang, Yuhan Zhang, Yunyu Xiao, Jyotishman Pathak, David Hodge, Yan Leng, S Craig Watkins, Ying Ding, et al. Analyzing social factors to enhance suicide prevention across population groups. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 189–199. IEEE, 2024.

Weipeng Zhou, Laura C Prater, Evan V Goldstein, Stephen J Mooney, et al. Identifying rare circumstances preceding female firearm suicides: validating a large language model approach. *JMIR mental health*, 10(1):e49359, 2023.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1): 237–291, 2024.

A Ethics Statement

Our work explores the use of language models to facilitate data-driven insights from suicide death narratives in the National Violent Death Reporting System (NVDRS). The University of Southern California Institutional Review Board reviewed this project and deemed it as not human subjects research; therefore approval was not necessary. In the following sections, we outline ethical considerations, including annotation procedures, annotator well-being, model limitations, and data privacy protections.

A.1 Annotator Well-Being

Given the sensitive nature of suicide death narratives, we implemented several precautions to support annotator well-being. The supervising expert engaged with the study team on a bi-weekly basis to have debriefs and check-ins regarding any feelings of emotional or mental heaviness of the task. Aside from formal team meetings, the supervising expert was also available to all study team members for individual appointments to discuss any concerns, emotional reactions, or mental strains from annotating. Specifically, the supervising expert instructed the annotators to be mindful of their reactions and feelings while annotating and, if they felt even the slightest inclination to stop annotating, then they should stop and engage in some activity that they enjoy (e.g., exercise, watch television, be with friends, etc.).

Research on concerns about working with mortality data have focused on one very specific type of data: suicide notes Pestian et al. [2012]. Suicide notes can range from cryptic to graphic, and notes are found in less than one-third of suicide deaths Rockett et al. [2018]. It is important to emphasize that NVDRS data rarely contain suicide notes; in some scant instances, brief quotes or paraphrasing of a suicide note may be included in the narrative. It is not standard practice that, if a suicide note was found on the scene, the text of said note is transcribed into the official reports or narrative. In fact, due to the potential for personally identifiable information included in a note, verbatim transcription of the note in totality would be ill-advised.

While the NVDRS narratives, themselves, may be graphic, researchers who have worked with both suicide notes and coroner reports indicate that suicide notes are emotionally heavier pieces of data to process than coroner reports Fincham et al. [2008]. In fact, while scholars who conduct secondary analysis with death documentation files acknowledge the emotional weight of the work, importantly, they also note: 'Our relationship to the files changed over time, allowing us to introduce the intellectual distance necessary for critical analysis, while the routines of mutual support we established helped us not to lose the empathy necessary for qualitative research' Fincham et al. [2008]. Based on these factors and prior literature Dasgupta [2021], our team engaged in current best practices for working with potentially emotionally taxing mortality data by having team debriefs, fostering open communication, encouraging breaks without imposing arbitrary limits (e.g., coding only 50 narratives at a time), and emphasizing the importance of listening to their own system signals and having the full agency to heed to those signals.

A.2 LMs as Annotation Assistants and Intended Usage

We conducted human annotation with one expert in suicide prevention and two trained annotators. The expert, a practitioner in the field, provided training, ongoing supervision, and led the codebook development process in collaboration with the annotators. All annotators independently labeled the same set of narratives, achieving strong inter-annotator agreement (Krippendorff's α = 0.88). Disagreements were resolved through consensus discussions. Given the sensitive nature of suicide death narratives, we implemented multiple safeguards to support annotator well-being.

While we explore the potential of language models to assist with expert annotation, we do not advocate for their deployment as a replacement for human annotators, but as tools to support more efficient and informed expert analysis. Recognizing the potential risks of LM-assisted annotation, we analyzed failure modes in both structured variable annotation (§2.3) and codebook development (§3.3.3), to better inform safe and effective practitioner use. Upon consulting with our suicide prevention expert collaborator, we emphasize the importance of expert oversight and validation to catch errors and maintain contextual nuance that LMs may overlook. We advise practitioners to abstain from relying on LM-generated annotations for low-agreement variables. In practice, models should be required to flag high uncertainty predictions or abstain altogether, prioritizing safety over coverage to minimize the risk of misleading outputs in expert workflows. Furthermore, we emphasize that codebook development should never be fully automated. Our objective is to reduce the burden on annotators and accelerate the discovery of data-driven insights, while preserving expert oversight.

A.3 Privacy

We followed NVDRS protocols for responsible data handling, and all experiments were conducted locally with open-weight models, ensuring that no data was shared with any LM API providers. To further protect data privacy, we do not include any qualitative narrative excerpts in the paper and all narratives were de-identified.

B Agreement Across Models

We use LMs as annotation assistants to label 36 circumstance and 14 crisis variables in NVRDS death narratives in §section 2. We evaluated LM agreement with data annotators on a test set: D_{balanced} composed of 500 narratives per variable (with equal representation across 0/1 classes). The per variable agreement across models is shown in Figure 5. On average, `Llama-3-70B` has the highest agreement with data annotators. However, there are a few outstanding cases where smaller models such as `Qwen2.5-14B` have higher agreement with the annotator (e.g. HistoryMentalIllnessTreatment).

Variable	Agreement	TPR	FPR	FNR
GunUnintentPulledTrigger_c	0.978	0.984	0.028	0.016
GunPlaying_c	0.976	0.968	0.016	0.032
GunThoughtUnloadedOther_c	0.954	0.992	0.084	0.008
RecentCriminalLegalProblem_c	0.934	0.908	0.040	0.092
CrisisEviction_c	0.932	0.924	0.060	0.076
SchoolProblem_c	0.926	0.996	0.144	0.004
HistoryMentalIllnessTreatmnt_c	0.736	0.484	0.012	0.516
DepressedMood_c	0.732	0.508	0.044	0.492
MentalHealthProblem_c	0.672	0.352	0.008	0.648
FamilyStressor_c	0.662	0.592	0.268	0.408
SuicideThoughtHistory_c	0.628	0.268	0.012	0.732
RelationshipProblemOther_c	0.624	0.964	0.716	0.036
CircumstancesKnown_c	0.614	0.232	0.004	0.768

Table 3: We report the agreement, true positive rate (TPR), false positive rate (FPR), and false negative rate (FNR) for a subset of variables with highest and lowest agreements. Performance is reported on the balanced evaluation set (D_{balanced}) using `Llama-3-70B`.

Model	Mean Agreement	S.D. across vars.
Llama-3-70B	0.82	0.12
Qwen2.5-14B	0.91	0.09
Qwen2.5-7B	0.56	0.17
Mistral-7B	0.67	0.14
Llama-3-8B	0.54	0.17

Table 4: Mean agreement and standard deviation across 50 variables for different models. `Llama-3-70B` achieves the highest agreement of 82% with data annotators. Performance is reported on a random evaluation set of 1000 narratives with unequal representation across 0/1 classes (D_{random}).

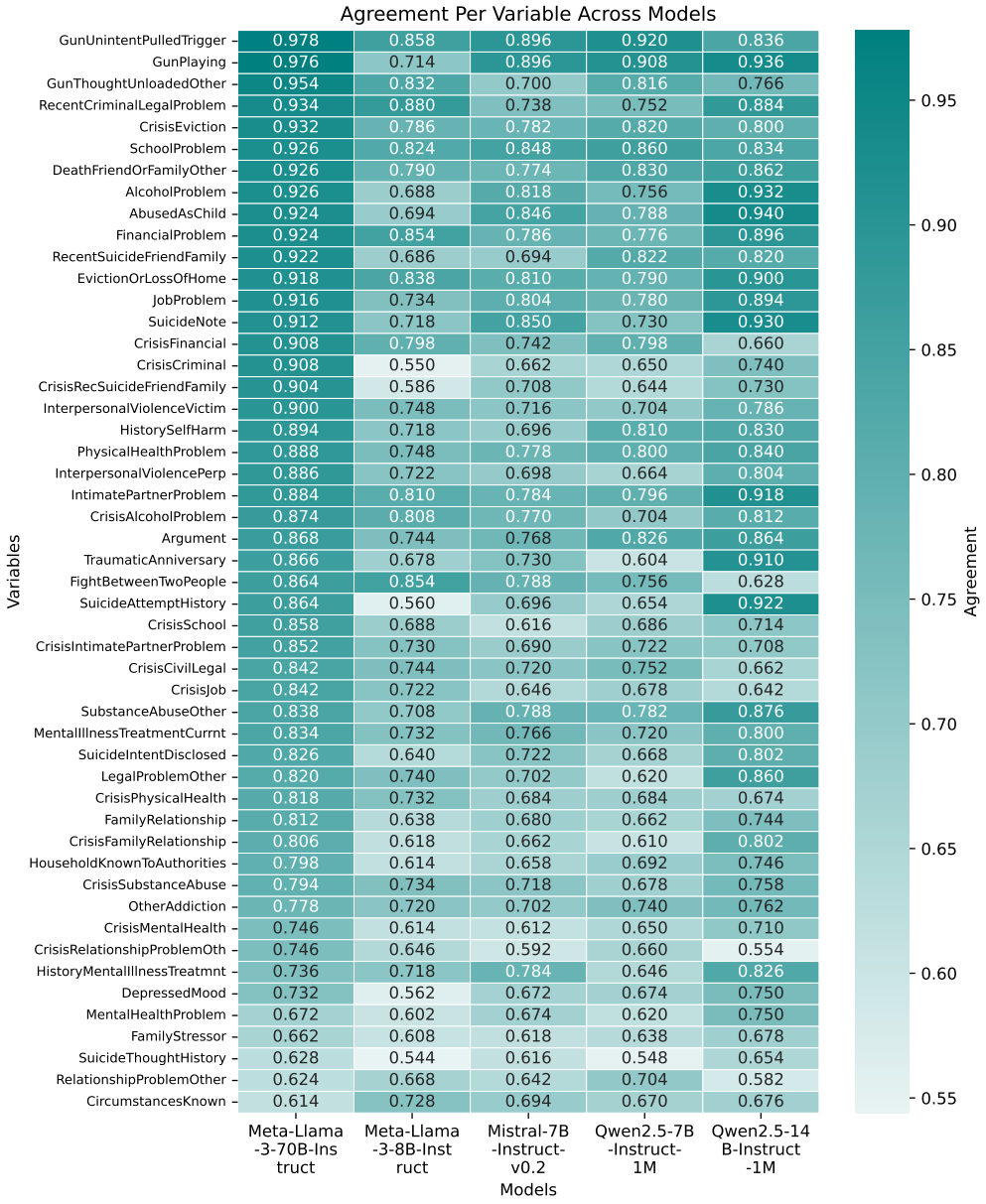

Figure 5: Per variable agreement for 50 NVDRS variables across different models. Agreement is reported on D_{balanced}. We find that on average, `Llama-3-70B` has the highest agreement with data annotators out of all evaluated models.

C LM-Simulated Codebook Development

To validate our codebook development algorithm, we first develop codebooks in a simulated setting for a subset of existing NVDRS variables. Figure 6 shows the accuracy on $\mathcal{D}_{guide}$ across 30 iterations. Most of the variables reach the max accuracy between iterations 10-15. Furthermore, we see greater instability in performance in earlier iterations due to the small size of $\mathcal{D}_{guide}$.

D Coverage-based Sampling

Coverage is defined as how much of a sample's content overlaps in content with another set of samples. Coverage-based sampling is inspired by Gupta et al. [2023], which showed that selecting a

set of samples by collective coverage leads to better performance than naively collecting samples by individual similarity. The main difference with our sampling strategy with Gupta et al. [2023] is that instead of measuring coverage at the token level, we compute coverage at the sentence level of the retrieved sample. Given a set of narratives, each narrative is split into sentences, which are then embedded with the `all-MiniLM-L6-v2` model from SentenceTransformers[10][Reimers and Gurevych, 2019]. The coverage of each sentence in a new sample is then computed by the maximum cosine similarity between the sentence and all other sentences in the set of chosen narratives. The coverage of the entire narrative is then the average value of these similarities. With all the coverage values computed for the set of samples to retrieve from, we select N samples with least coverage to promote diversity.

E Codebook Development Algorithm Hyperparameters

Table 6 provides an overview of hyperparameters for the codebook development algorithm. t^* is the number of iterations that the codebook development algorithm ran for. For all experiments, t was fixed to 30. In practice, t would vary depending on the performance m on $\mathcal{D}_{guide}$. b is the budget—the maximum number of narratives that the algorithm is allowed to iterate over. n is the batch size per iteration. n can be sampled using random or coverage-based sampling. Model_id is the model used for guideline development in our algorithm. k is the minimum size of $\mathcal{D}_{guide}$, and m is the target performance for $\mathcal{D}_{guide}$. We leave it to expert judgment to determine b, k, and m, as these depend on the nature of the variable and thus, the number of iterations required to reach thematic saturation [Glaser and Strauss, 1967]. Experts may consider the number of consecutive iterations without adding a guideline as an additional hyperparameter for determining the stopping condition. For our legal interaction case study, we set b as 150 given experts analyzed 150 narratives manually in one round of qualitative coding to develop the codebook manually. We set k by observing performance trends on $\mathcal{D}_{guide}$ in the LM-simulated codebook development experiments (§3.2 where performance was unstable in the first 5 iterations.

Model Name
Meta-Llama-3-70B-Instruct
Meta-Llama-3-8B-Instruct
Mistral-7B-Instruct-v0.2
Qwen2.5-14B-Instruct-1M
Qwen2.5-7B-Instruct-1M

Table 5: Full names of models used in §2 and §3.1.

Model	t*	b	n	**sampling**	**model_id**	k	m
LM-Sim (NVDRS)	30	150	5	Random/Coverage	`Llama-3-70B`	-	-
`HitL` (Legal)	30	150	5	Coverage	`Llama-3-70B`	30	0.9

Table 6: Configuration details for LLM-Sim and `HitL` across various parameters.

F Case Study: Legal Interactions

We apply our codebook development algorithm to a real world case study and develop guidelines for a new variable: victim interactions with legal professionals. In Table 6, we show the hyperparameter configurations for both our LM-simulated codebook development and `HitL` setting for legal interactions. In Table 8, we show the initial prompt templates $\mathcal{G}_0$, for both the simulated and `HitL` setting. Table 9 shows our manually developed expert guidelines (left) and our `HitL` guidelines (right).

Figure 8 (left) shows the distribution of narratives with implicit-, explicit-, and no interactions across three data splits. In Figure 8 (right), we show the distribution of the proportion of positive occurrences for 50 NVDRS variables in all 270K cases. This distribution has a heavy right skew showing the heavy class imbalance in NVDRS.

[10] `https://sbert.net/index.html`

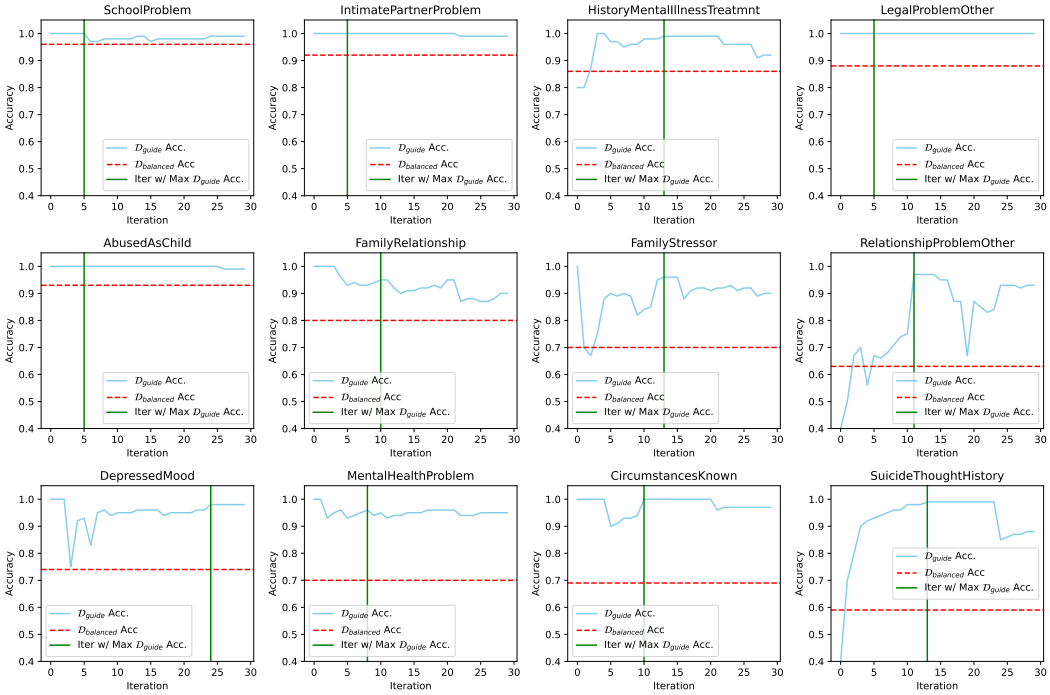

Figure 6: Accuracy on $\mathcal{D}_{guide}$ across 30 iterations for LM-Simulated codebook development for 12 NVDRS variables. The maximum accuracy on $\mathcal{D}_{guide}$ is reached between 10-15 iterations for all variables using random sampling per iteration.

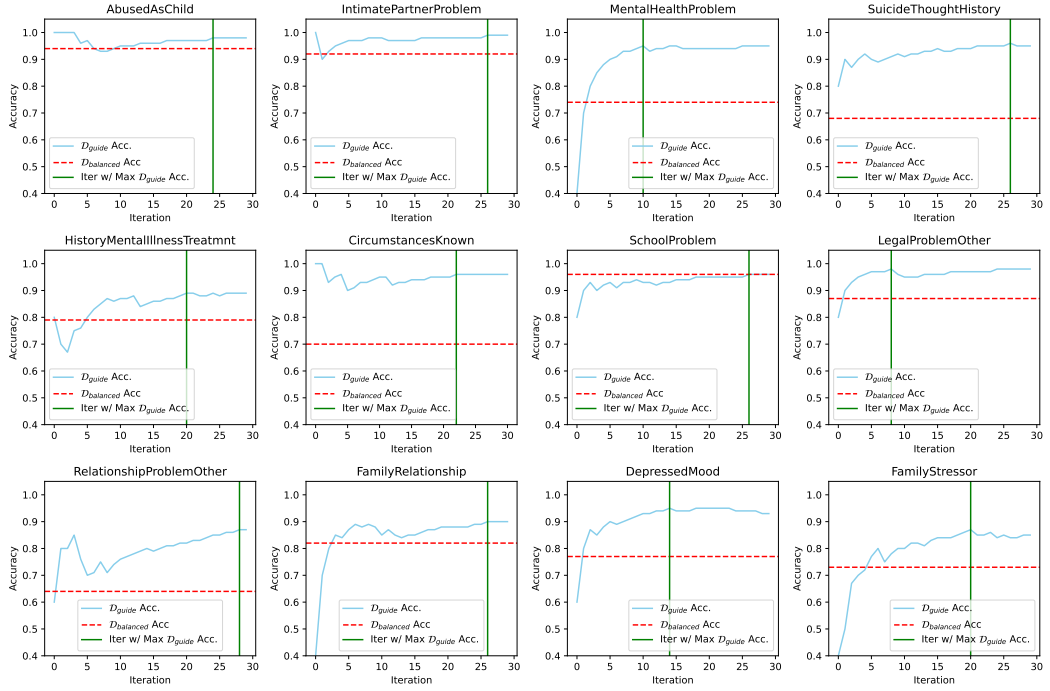

Figure 7: Accuracy on $\mathcal{D}_{guide}$ across 30 iterations for LM-Simulated codebook development for 12 NVDRS variables using coverage based sampling per iteration.

Annotated Set	Implicit	Explicit	None
$\mathcal{D}_{guide}$	55	23	72
$\mathcal{D}_{val}$	20	20	20
$D_{\text{expert\_legal}}$	74	83	477

Table 7: Distribution across implicit, explicit and no-interactions for all annotated sets in legal interaction case study.

LM Simulation (NVDRS)	**HitL (Legal Interactions)**
Prompt Instructions: You are an expert suicide caseworker and your job is to annotate reports with the {code_name} variable. Do not read into the text and stick to the definition of variable strictly. If two reports are provided, use both reports to determine your response but only return one response for both reports with no additional text! Provide the reasoning for your answer, the span of text that you used to generate your answer and your response using the response options only and return your answer in the following format: {'reason': 'reasoning', 'span': 'span of text', 'response': '1.0 or 0.0'} **Guideline Synthesis Prompt**: You are an expert suicide caseworker and your job is to curate a set of guidelines that will be used by another model to label suicide reports with the variable:{code_name}. You will be shown the original set of guidelines, the report that was used to label the variable {code_name}, the model's label, the correct human label, the human's reasoning, and the span of text that the human used from reports to decide their label. The label can be 0.0 or 1.0. You have to return a set of new guidelines using this information which will be used to annotate {code_name} for future reports. Keep the guidelines concise, and use the human reasoning, span, or other information from the report to update the guidelines, make sure to not lose out on information in the original set of guidelines but try not to have too much repetition. You have to return your answer in the following format with absolutely not additional text!: 'Guidelines: *..., *...'.	**Prompt**: Instructions: You are an expert suicide caseworker trained to correctly categorize suicide reports by the victim's interaction with a lawyer or attorney. You have to label each report with only one of the 3 interaction types and return your answer in the following format: {reason: 'reasoning', span: 'span of text', label: 'implicit_interaction, explicit_interaction, no_interaction'}" , with the reason behind your answer, the span of text you used to determine your answer, and a label and no additional text. If two reports are given, only return one answer using both reports using the format and make sure to provide which report you got the span from! Classes: Label: no_interaction • Definition: It is not implied or explicitly stated that V had interactions with a lawyer. Label: implicit_interaction • Definition: V had an implicit interaction with a lawyer where it is implied that V had an interaction with a lawyer. Label: explicit_interaction •Definition: There are explicit mentions of V interacting with a lawyer or attorney.

Table 8: Prompt templates for LM simulated setting (NVDRS variables) and for HitL codebook development for legal interactions.

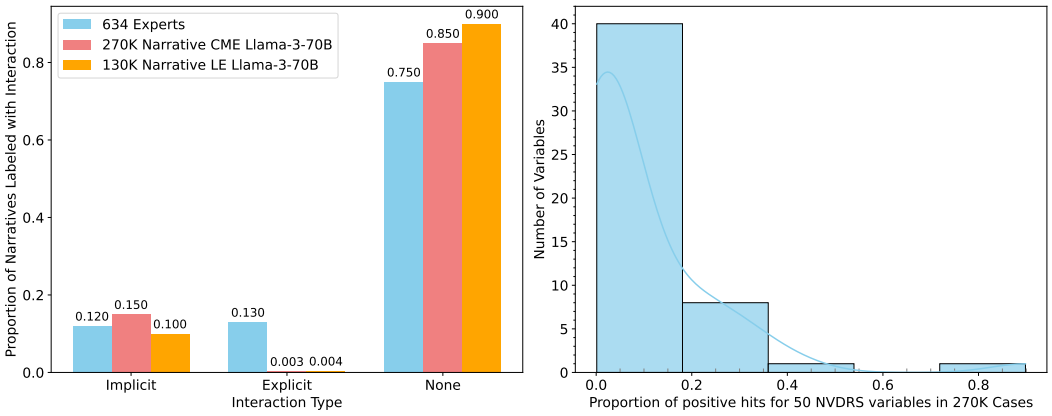

Figure 8: Distribution of narratives containing implicit-, explicit- and no-interaction for 3 data splits - 634 experts ($D_{\text{expert\_legal}}$), 270K CME narratives, and 130K LE narratives (left). Distribution of proportion of positive occurrences for 50 variables in 270K NVDRS cases (right).

Expert (Manual)	HitL
Guidelines: • Label: no_interaction Definition: It is not implied or explicitly stated that V had interactions with a lawyer. Being released from jail, arrest warrants, or being under investigation for a crime should be labeled with no_interaction Positive Example and Justification: 'Censored': Being released from jail does not imply any interaction with a legal professional • Label: implicit_interaction Definition: The report mentions circumstances such as divorce, separation, or issues surrounding child custody/visitation, and should be labeled with implicit_interaction . Additionally, mentions of court proceedings/appearance, court orderings, restraining orders, financial crimes, lawsuits or if V was charged/accused with severe crimes such as DUI/DWI, assaulting an officer, battery, domestic violence/protection orders, ongoing legal problems and arrests for severe crimes within the last 6 months etc. all imply interactions all should be labeled with implicit_interaction . Positive Example and Justification: 'Censored': V was going through a divorce so it is implied they had an interaction with an attorney • Label: explicit_interaction Definition: There are explicit mentions of V interacting with a lawyer or attorney. Only choose this label if the legal professional (i.e. lawyer) is directly mentioned in the report. Positive Example and Justification: 'Censored': It is explicitly stated that V missed appointments with his lawyer, so this is explicit_interaction	**Guidelines**: •Litigation was noted as pending meaning it was scheduled for some future date, therefore it is unclear if victim had actually spoken with a lawyer at the time of death, and this should be labeled as no interaction. •Typically if the victim, themselves, was an attorney, this should be labeled as no interaction. However, when the victim was an attorney and the victim had evidence of legal problems requiring some hearing, court interaction, or need of lawyer service, then this should be labeled as implicit if it is not stated that they did not directly interact with another lawyer or attorney. If they did interact with another lawyer or attorney, then this should be labeled as explicit. •Although the victim has bankruptcy paperwork, it is unclear if this paperwork was filed thus it is unclear if a lawyer or attorney was currently involved, and this should be labeled as no interaction. •Because the victim was facing criminal charges, this likely means a lawyer or attorney was involved in this legal proceeding at the time of death, and this should be labeled with implicit interaction. •Although the victim's sale of his business was not going well, that phrase cannot be interpreted as indicating an implicit interaction with a lawyer or attorney, and this should be labeled as no interaction. •although the narrative mentions the victim had a nasty divorce, which would typically be an implicit interaction, it was noted the divorce was 2 years in the past, which means any current interactions with a lawyer or attorney is unlikely and this should be labeled as no interaction. •In this instance, the mention of lawyer or attorney is in reference to the sister of the victim and not the victim, themselves, and the sister talking to a lawyer seems to have been after the victim's death. The victim, themselves, should be the one who had the interaction, or the family member who talked with a lawyer or attorney should have done so before the victim's death, then this should be labeled as implicit or explicit depending on whether the lawyer or attorney is noted. •In this narrative the IRS issues were framed as they were going to visit, which means this had not happened yet, therefore it is unclear if a lawyer or attorney was yet involved, and this should be labeled as no interaction. •If the victim is in the process of a divorce or if a divorce hearing is pending, then that should be labeled as implicit interaction. •The victim was considering bankruptcy, which means we do not know if a lawyer or attorney was involved, and this should be labeled as no interaction. •There was an ongoing custody issue that included a guardian ad litem, which is a court-appointed representative, so this should be labeled as implicit interaction. •The victim was facing jail time or imprisonment, and this should be labeled as implicit interaction. •Just because the victim was a law student, does not mean there was an interaction and should be labeled as no interaction. •The threat of being sued was not sufficient to imply a lawyer or attorney interaction and should be labeled as no interaction. •Because the narrative explains that victim was awaiting trial, that should be labeled as implicit. •Just because a complaint had been filed, that is not sufficient to assume a lawyer or attorney interaction, so this should be labeled as no interaction. •The victim had looked into getting a lawyer, so this should be an explicit interaction. • The narrative says an attorney was trying to get charges reduced, so this should be coded as an explicit interaction. • The narrative indicates that the victim has recently been arrested, so this should be labeled as implicit interaction. • The lawsuit being within 3 months of the death is strong enough rationale to label this as in implicit interaction. • If the narrative only contains the phrase some kind of legal problems; with no other information about an attorney or lawyer, then this should be labeled as no interaction. • If the narrative only states that legal problems were in the past; with no other information about a lawyer or attorney, then this should be labeled as no interaction. • The victim was working the DA, which stands for district attorney, so this should be labeled as an explicit interaction. • Because divorce proceedings were initiated or in process without any explicit mention of a lawyer or attorney, this should be labeled as an implicit interaction. • If a lawsuit is pending and there is no explicit mention of a lawyer or attorney, then that should be labeled as implicit interaction.

Table 9: Expert codebook (left) for defining legal interactions and HitL codebook developed with suicide prevention expert (right).

G NVDRS Codebooks vs LM Simulated Codebooks

We provide our generated codebooks (right) for 3 NVDRS variables. Codebooks generated with our LM-simulated pipeline contain finer-grained instruction and examples from narratives which could be helpful in the future for augmenting existing NVDRS codebooks.

Variable	NVDRS Codebook	LL-Simulated Codebook)
AbusedasChild	**Prompt**: The victim had a history of abuse (physical, sexual, or psychological) or neglect (physical, including medical/dental, emotional, or educational neglect; or exposure to violent environments; or inadequate supervision) as a child. This variable more broadly captures victim's experiences of abuse and neglect irrespective of its relationship to the violent death. Code "Yes" if the victim experienced abuse or neglect, but there is no direct link to the violent death, or the link is unknown. •Do NOT code if the abuse or neglect directly causes or precipitated the death, instead code abuse/neglect led to death. • Code as "Yes" if the victim had been the victim of child abuse at any point in the past, even if the victim is currently an adult. •Code "Yes" if the evidence of ongoing abuse is suspected but not confirmed. Code "Yes" if autopsy evidence reported an indication of previous abuse. •Abuse can be physical, psychological, sexual, or other as long as the source document refers to "abuse." Neglect may be medical, physical, or emotional. •Indicate the nature of the abuse or neglect in the incident narrative. • Abuse could have been perpetrated by a caregiver or other individual.	**Prompt**: • label abusedaschild as 1.0 if the report explicitly mentions abuse or neglect in the victim's childhood, •label abusedaschild as 1.0 if the report suggests a potentially abusive relationship in adulthood, such as a history of domestic calls or restraining orders, as this may indicate a pattern of abuse, • label abusedaschild as 0.0 if there is no mention or indication of abuse or neglect in the victim's childhood or adulthood, •consider the entire report, including the victim's history and circumstances surrounding the death, but do not make inferences about childhood experiences based on adult behaviors or circumstances unless there is a clear pattern of abuse, •do not assume childhood abuse based on domestic disputes or arguments in adulthood, •absence of any mention of childhood abuse or neglect in the report indicates a label of 0.0.
FamilyRelationship	**Prompt** • Code as 1.0 if at the time of the incident the victim was experiencing a relationship problem with a family member other than an intimate partner (e.g., a child, mother, in-law), and this appears to have contributed to the death. Suicide example: The victim is despondent over his argument with his parents and dies by suicide. •If the report indicates that the victim was "having relationship problems," these should be assumed to be intimate partner problems, and not problems with other friends or family. •This code should not be endorsed for custody disputes when the victim is a child because the relationship problem in these instances is typically not with a child or other non-intimate partner family member. If the custody dispute affects the relationship of the parent and child, then this should be coded "1.0." • Narrative should contain an explanation of the relationship problem and identify the family member with whom the victim had a problem. •This variable does NOT capture problems related to the family environment that do not explicitly pertain to the victim's relationship with a family member (e.g., the stress of caring for a sick family member).	**Prompt**: • Label as 1.0 if the report mentions a conflict, argument, or strained relationship with a family member (e.g. parent, sibling, child) that contributed to the death, including instances where the family member's death anniversary is mentioned •Label as 1.0 if the report mentions a problematic relationship with a family member that occurred shortly before the incident •Label as 1.0 if the report mentions a loss of custody or a sense of failure towards family members especially children, that contributed to the death •Label as 1.0 if the report mentions the victim being locked out of their residence by family members or being isolated from them, indicating a strained relationship • Label as 0.0 if the report does not mention a strained or problematic relationship with a family member, or if the family member is only mentioned as a discoverer of the body or as someone who had a normal conversation with the deceased prior to the death, or if the family member is mentioned in the context of a past traumatic event that is not directly related to a strained relationship • Consider the context of the report, if the victim's despondency and irrational behavior are related to other factors such as terminal illness, alcoholism, etc., and not directly related to a family member, label as 0.0.
LegalProblemOther	**Prompt** Civil legal (non-criminal) problem(s) appear to have contributed to the death Code as "Yes," if at the time of the incident the victim was facing civil legal problems, such as a divorce, custody dispute, civil lawsuit, or legal problems that were unspecified as either criminal or civil, and these problems appear to have contributed to the death.	**Prompt**: • legalproblemother refers to any legal problem that is not related to criminal justice or immigration, such as civil legal problems like divorce, child custody, eviction notices, or upcoming court appearances, •a legal problem is considered "other" if it is not related to criminal justice or immigration, even if it is not explicitly mentioned as a contributing factor to the suicide, •the presence of a civil legal problem can be inferred from the context of the report, but should be distinguished from financial concerns or medical issues, •specifically look for mentions of court appearances, legal proceedings, or legal issues that are not related to criminal justice or immigration, • eviction notices or other civil legal problems that contribute to feelings of depression or hopelessness should be labeled as 1.0, • if there is no indication of a civil legal problem in the report, and the report only mentions financial or medical issues, label as 0.0.

Table 10: NVRDS codebook guidelines for 3 variables (left) compared to codebooks generated in the LM simulated setting (right) in §3.2

H Further Related Work

H.1 NLP for NVDRS

Prior work has applied natural language processing approaches to identify risk factors for suicide (e.g., social determinants of health - SDoH such as substance abuse, stress, mental health conditions) [Johns et al., 2023, Wang et al., 2023, Consoli et al., 2024, Guevara et al., 2024, Zhou et al., 2023]. Specifically, Wang et al. [2023] finetune BERT in classification setting to characterize circumstance and crisis variables related to social determinants of health using NVDRS narratives. In a more targeted case study, Zhou et al. [2023] used language models to identify infrequent circumstances preceding female firearm suicide and coded 9 infrequent circumstances using a manually developed codebook. Alternatively, Consoli et al. [2024] used GPT-3.5 with in-context learning to characterize medical notes with SDoH, whereas Guevara et al. [2024] use LLMs to identify SDoH in electronic health records. Furthermore, Arseniev-Koehler et al. [2022, 2021], Davidson et al. [2021] use topic modeling approaches to uncover latent themes in NVRDS narratives.

Kafka et al. [2023] applies supervised machine learning to detect cases of intimate partner violence (IPV). However, they find that their approach does not capture implicit references to IPV due to the long narrative lengths. Motivated by this finding, our case study explores how to employ an LM assistant to effectively develop codebooks for new variables to identify variables that contain more implicit references. Given the scale of NVDRS narratives, there have been few studies exploring the utility of LMs in this domain [Dang et al., 2023]. To this end, we hope to build on prior work that has shown the effectiveness of LMs for narrative analysis by employing LMs as efficient assistants to data annotators and experts.

H.2 Qualitative Coding with LMs

Data abstractors face the emotionally demanding and labor-intensive task of annotating graphic and explicit suicide cases in NVDRS. Collecting high-quality annotations using a codebook (e.g. guidelines for variables) for sensitive tasks is essential in downstream analysis. However, codebook development is a time-consuming and laborious task that involves repeated iterations of manual data analysis [Glaser and Strauss, 1967]. Prior work has explored the use of LMs for thematic analysis [Katz et al., 2024, Dai et al., 2023], qualitative coding [Barany et al., 2024, Tai et al., 2024, Xiao et al., 2023, Perkins and Roe, 2024] and annotation of socially sensitive data [Ranjit et al., 2024, Halterman and Keith, 2024, Rytting et al., 2023, Pangakis et al., 2023]. Barany et al. [2024] examine the use of language models in qualitative coding for education, comparing fully manual, fully automated, and LM–human collaborative approaches to codebook development. They find that hybrid approaches perform comparably to manual codebooks and, in some cases, outperform them. Furthermore, Dai et al. [2023] propose a collaborative LM–human framework in which humans provide exemplars for in-context learning and validate LM-generated codes, where they find high inter-annotator agreement in deductive coding using the resulting codebook. Comparatively, Katz et al. [2024] simulate the process of inductive coding and thematic saturation using open source models in a fully automated pipeline applied to hypothetical organizational settings.

While codebook development for highly sensitive tasks should not be fully automated, it remains unclear at which stages of inductive coding LMs can support experts without compromising annotation quality. For NVDRS, we explore both LM-assisted deductive coding (e.g. codebook is available) in §2.2 and §3.2 and inductive coding (e.g. developing a codebook from the data) in §3.3. Given the highly sensitive nature of our task, we cannot rely on fully automated codebook development approaches. Instead, we introduce a collaborative setting between experts and LMs where we rely on the expert to develop the initial set of labels. We use embedding-based similarity [Douze et al., 2024] to sample data and leverage model success and failure modes to assist experts in refining codebook guidelines. Experts only need to provide guidelines for instances where the model makes incorrect predictions, thus reducing their cognitive load in fully manual analysis. We find that our framework helps experts augment existing codebooks by using LMs to surface finer-grained examples from the data and develop guidelines for newly defined codes. Finally, codebook development and refinement have not been explored in the context of NVDRS, which presents challenges due to the explicit and graphic nature of the data, as well as its implications for downstream intervention development. Our work addresses the unique challenges of applying codebook development algorithms to highly sensitive domains using NVDRS.