Uncovering Intervention Opportunities for Suicide Prevention with Language Model Assistants

Anonymous ACL submission

Abstract

The National Violent Death Reporting System (NVDRS) documents suicides in the United States. In a demanding public health data pipeline, annotators manually extract structured information from death investigation records following extensive codebooks (i.e. annotation guidelines) painstakingly developed by experts. In this work, we facilitate data-driven insights from the NVDRS data to support the development of novel suicide interventions by leveraging language models (LM) as assistants to these (a) data annotators and (b) experts. We find that LM predictions match existing data annotations about 85% of the time across 50 NVDRS variables. Where the LM disagrees with existing annotations, our expert review identifies that 38% of these instances reveal inconsistencies between narratives and structured data. Finally, we introduce a human-in-the-loop algorithm that helps experts efficiently build and refine codebooks for new variables by having them only focus on providing feedback for incorrect LM predictions. We apply our algorithm to a real-world case study, and find that about 28K narratives contain evidence of victim interactions with legal professionals, which surfaces a substantial opportunity for upstream intervention that is not captured in the original structured data. Our findings provide evidence that LMs can serve as effective assistants to public health researchers who handle sensitive data in high-stakes scenarios.

1 Introduction

Warning: This paper discusses suicide, which may be distressing to 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)[1], which contains structured and unstructured data for more than 270K suicides (Figure 1; top left). Structured data in NVDRS is manually labeled by annotators (or NVDRS data abstractors[2]) using codebooks—precise definitions and annotation guidelines—developed painstakingly by experts. Given the sensitive nature of the topic and the scale of NVDRS, a data annotator's job is 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 a valuable source for data-driven discovery of *novel* suicide interventions. To uncover actionable insights for new suicide interventions, experts must move beyond existing structured variables and extract evidence from unstructured narratives that support their hypotheses. However, this process requires experts to manually analyze NVDRS narratives, develop a codebook for this new variable, and abstractors to retroactively annotate all 270K cases.

To overcome these challenges, we investigate whether language models (LMs) can help public health researchers make data annotation and analysis for suicide prevention research more efficient and effective. Our investigation is based on the promise LMs have shown in social science (Rytting et al., 2023; Pangakis et al., 2023; Halterman and Keith, 2024; Ziems et al., 2024), and aims to answer 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?* For **RQ1**, we deploy LMs as annotation assistants to data abstractors and find that they achieve an average agreement of 85% with data abstractors across

[1] https://www.cdc.gov/nvdrs/about/index.html

[2] We use the term data abstractors and annotators interchangeably.

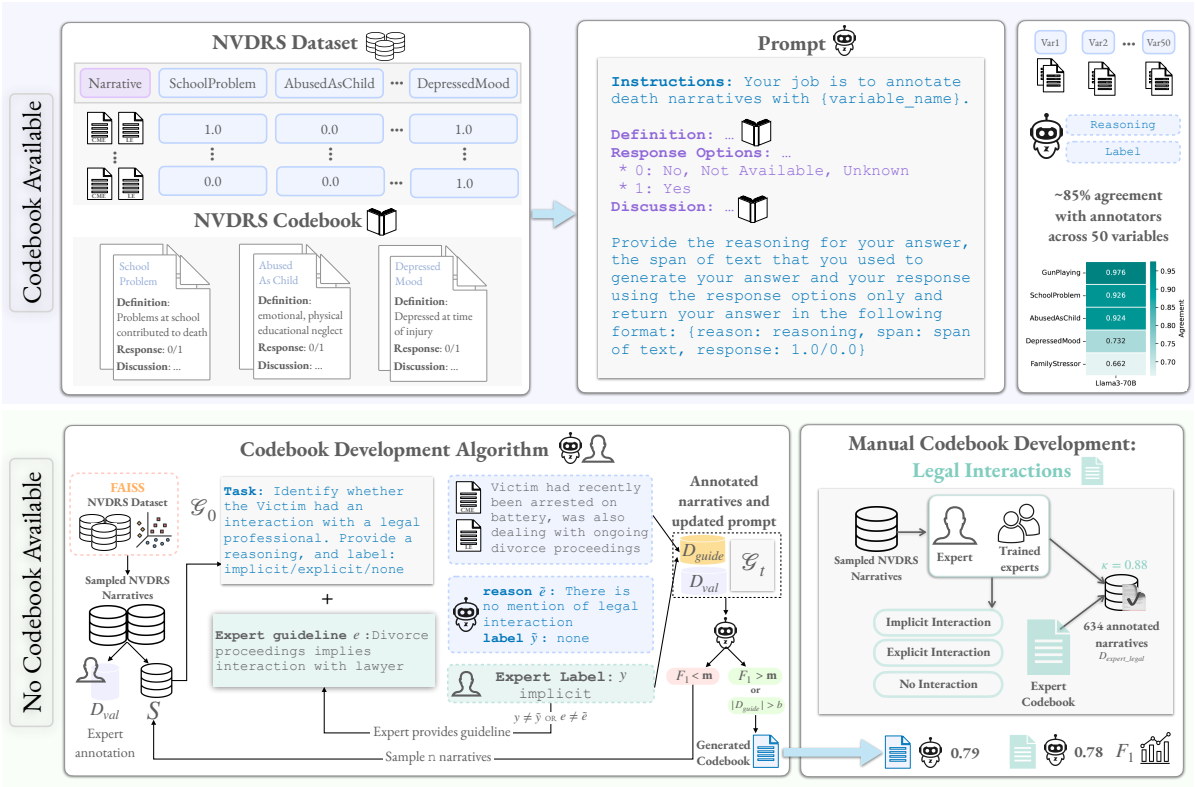

Figure 1: LM assistants can help annotate NVDRS narratives when a codebook is available (top) and help experts efficiently develop new codebooks for novel variables (bottom). When a codebook is available, we incorporate it in the prompt and generate predictions. For a new variable of interest, we propose a human-in-the-loop codebook development algorithm. Here, experts focus on providing feedback for incorrect LM predictions, reducing manual codebook development time from weeks to hours (bottom right). We apply our algorithm to a novel variable characterizing victim interactions with legal professionals. We find that 10.4% of 270K NVDRS narratives contain evidence of such interactions, indicating a new and salient intervention opportunity, while achieving annotation quality comparable to a laborious manual approach.

50 variables (§2.3), making them a reliable peer validator for high-agreement variables. They are also useful for low-agreement variables: from our expert review of six low agreement variables, we find that 38% of disagreement instances reveal inconsistencies between the narratives and the structured data, suggesting that LMs can surface inconsistencies between the two sources of data (§2.4).

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

Suicide prevention research has largely focused on risk factors arising in clinical settings, leaving opportunities for *preventative* care through interactions with nonclinical professionals underexplored (Lybarger et al., 2023; Ralevski et al., 2024; Xu et al., 2024). To examine this gap, we apply our algorithm to study a novel nonclinical factor for suicide prevention: identifying victim interactions with legal professionals (§4). We find that 10.4% of 270K NVDRS narratives contain evidence of such interactions, indicating a substantial and underexplored avenue for new interventions. Furthermore, our algorithm reduces codebook development time from weeks to hours without compromising on annotation quality, demonstrating its potential to efficiently surface and characterize a wider range of nonclinical intervention opportunities (e.g., interactions with other nonclinical professionals).

In summary, our results indicate the ability of LMs to serve as efficient assistants that enable experts to mobilize the NVDRS dataset more effectively, and accelerate the identification of novel intervention opportunities[3]. The promising results we demonstrate in building LM assistants for professionals in public health emphasize the value of collaborative approaches between LMs and human experts, especially in high-stakes domains.

2 Predicting 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 narratives? This setting includes established codebooks and abstractor labels, allowing us to evaluate whether LM predictions can follow the same.

2.1 The National Violent Death Reporting System (NVDRS)

The NVDRS contains records for 270K suicide cases between 2003-2019 in 42 US states, the District of Columbia, and Puerto Rico (Liu, 2023; Wilson, 2022). Each case is characterized by more than 600 structured variables which include victim demographics, and circumstances surrounding their death (Liu, 2023). Suicide circumstance (e.g., eviction or loss of home, alcohol problem) 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 annotation guidelines for each variable.

2.2 Experimental Setup

We consider a subset of variables that have been manually annotated by NVDRS abstractors, based on information from CME / LE records. Since the narratives are also derived from these records (Paulozzi et al., 2004), they should ideally capture the same information as the structured data.[4]

[3]Our code will be made publicly available: `https://anon`

[4]We selected 50 binary variables (36 circumstance and 14 crisis variables) that appeared in at least 300 cases to ensure reliable evaluation.

Model	Mean Agreement	SD
Llama-3-70B	**0.85, [0.82, 0.87]**	0.09
Qwen2.5-14B	0.79, [0.76, 0.82]	0.10
Qwen2.5-7B	0.73, [0.70, 0.75]	0.08
Mistral-7B	0.73, [0.71, 0.75]	0.07
Llama-3-8B	0.71, [0.69, 0.74]	0.09

Table 1: Mean agreement with data abstractor annotations, 95% confidence intervals (bootstrap, 10K iters), and standard deviation across 50 variables (SD). `Llama-3-70B` achieves the highest agreement. Performance is reported on a balanced evaluation set of 500 narratives per variable (D_{balanced}).

We use locally-hosted, open-weight LMs from various model families and sizes as annotation assistants, given the private and sensitive nature of the NVDRS records. We curate LM prompts for each variable by adapting the NVDRS codebooks in a standardized format with instructions, definitions, response options, and discussion, as shown in Figure 1 (top panel). For the input, we concatenate the CME and LE narratives, and generate Chain-of-Thought reasoning (CoT; (Wei et al., 2022)), and a label in a zero-shot setting. Our approach offers a lightweight alternative to finetuning models (Wang et al., 2023): it is easily adaptable to codebook changes and does not require retraining models, which is computationally expensive.

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. 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 LM Annotation Results and Analysis

Table 1 shows the average agreement between LM predictions and abstractor annotations for D_{balanced} across 50 variables, with 95% confidence intervals. The complete results on D_{random} are included in Table 4 in Appendix A. We find that `Llama-3-70B` achieves the highest average agreement: 85% for D_{balanced} and 82% for D_{random}. We performed paired t-tests comparing `Llama-3-70B` against each model across 50 variables. All comparisons were statistically significant ($p < 0.0125$ under Bon-

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

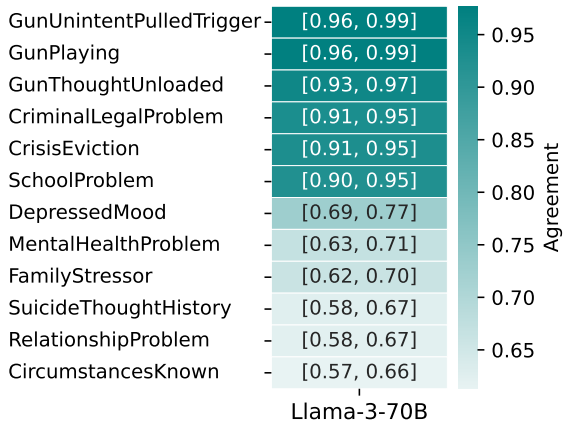

Figure 2: Per variable agreement (95% confidence intervals; bootstrapped with 10K iterations) for 12 highest and lowest agreement variables. `Llama-3-70B` has low agreement with data abstractors for mental health-related variables (e.g., SuicideThoughtHistory, MentalHealthProblem) and relatively higher agreement for firearm-related variables (e.g., GunPlaying, GunUnintentPulledTrigger) (See Figure 4 in Appendix for results across all models and variables).

ferroni correction). Additionally, `Llama-3-70B` outperforms all other models on at least 38 out of 50 variables. `Llama-3-70B` is also highly self-consistent across three runs with varying temperature values (i.e. 0.2, 0.5, 0.7) producing identical predictions 99% of the time. Given the fairly high agreement with abstractors, our findings suggest that LM assistants can serve as a peer validator in the absence of additional human verification for NVDRS annotations, especially given only 5% of NVDRS annotations are validated by two annotators (Liu, 2023). Furthermore, the LM's CoTs highlight relevant context in long narratives that abstractors may overlook during initial annotation. For example, for the FamilyStressor variable, the CoT connects domestic violence history and concerns about the daughter's suicidal ideation as indicators of a stressful family environment.

From analyzing the 6 highest and 6 lowest agreement variables from D_{balanced} (illustrated in Figure 2), we observe that all models have relatively higher agreement for firearm-related variables (e.g., GunPlaying, GunUnintentPulledTrigger) and low agreement for mental health-related variables (e.g., SuicideThoughtHistory, MentalHealthProblem). The true positive rate (TPR) and false negative rate (FNR), as shown in Table 3, also indicate that concrete, observable events with explicit lexical cues (e.g., GunUnintentPulledTrigger) achieve very high TPR (>0.96) and low FNR (<0.03). In contrast, variables related to mental health and emotional state (e.g., MentalHealthProblem, DepressedMood, SuicideThoughtHistory) exhibit low TPRs and high FNRs (>0.7).

To understand why LMs fall short in some cases for low agreement variables, we conducted a targeted qualitative analysis for two of them: MentalHealthProblem and DepressedMood. We manually reviewed 100 `Llama-3-70B` predictions and their chains-of-thought (CoT), and identified two recurring failure modes. In false negative cases, references to mental health problems or depressed moods appeared in only one of the two concatenated narratives (i.e., CME or LE), reducing the model's ability to detect the variable in longer context narratives. In false positive cases, the model's CoTs frequently cited circumstances such as financial problems, substance use, and interpersonal conflict as evidence for mental health problems.

2.4 LM Disagreements Surface Annotation Inconsistencies

Given variations in recording practices, prior work by Wang et al. (2023, 2024b) reveal inconsistently annotated NVDRS variables among abstractors. Specifically, Wang et al. (2024b) identify cases where circumstances described in the narratives are not reflected in the structured data, resulting in inconsistencies between the two sources. Given their finding, we hypothesize that disagreement between the LM assistant and the abstractor can surface inconsistencies between the narratives and structured data. For six variables with low agreement between the LM (`Llama-3-70B`) and abstractor (i.e. CircumstancesKnown, RelationshipProblem, 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 get a second opinion from a suicide prevention researcher on our team with 15 years of experience.

In cases where the LM and the 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 the 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 (see Appendix A.1 for discussion on the robustness of our hypothesis test). Future work is needed to systematically account for and rectify annotation inconsistencies in the NVDRS annotations across a larger number of variables with LM assistants.

3 Characterizing New Variables in NVDRS Narratives

Although NVDRS codebooks have continuously expanded their scope, they are not exhaustive and updates are infrequent and incremental due to the scale and diversity of narratives (Steenkamp et al., 2006). Therefore, it is paramount to develop efficient methods for characterizing new variables to enable data-driven analysis for testing novel hypotheses in prevention research and informing intervention strategies (Blair et al., 2016).

To this end, we introduce an algorithm that efficiently creates codebooks for new variables by refining an initially coarse codebook. In this section, we elaborate on our algorithm (§3.1) and demonstrate its usefulness by validating LM-developed codebooks through simulations with existing variables (§3.2). In the following section, we present a case study to extract real-world, actionable insights for a new variable capturing victims' interactions with legal professionals (§4).

3.1 Codebook Development Algorithm for Characterizing Variables

Our algorithm, formally defined in Algorithm 1, uses an LM to (i) generate predictions using the current codebook in the same manner as §2, and (ii) iteratively refine the codebook by synthesizing per-sample feedback from experts on those LM predictions.[6] The main advantage of our algorithm over a traditional manual codebook development (§4.2) is that the expert only needs to focus on providing feedback on LM predictions for a few samples in each iteration, making it more efficient. The codebook is automatically refined by the same LM with a codebook update prompt based on the expert's feedback. We repeat this process until the codebook results in LM predictions that achieve a predefined target performance on an evaluation set.

Initialization. First, we collect expert annotations on a subset of narratives, $\mathcal{D}_{val}$, which serves as a held-out validation set, containing at least j instances per class. The likelihood of finding a true positive can be low depending on the expert hypothesis. To address this, 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). We set our initial codebook $\mathcal{G}_0$ using a template that only includes the variable name and label classes (see Table 8 in Appendix E). The LM uses this to predict a label $\tilde{y}$ along with its chain-of-thought reasoning $\tilde{e}$ for n sampled narratives at each iteration t.

Algorithm 1: Codebook Development

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

Output: $\mathcal{D}_{\text{LM}}$: LM-annotated data , $\mathcal{G}_t$: final guideline

$t \leftarrow 0$ // Iteration 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 $\mathcal{D}_{\text{LM}}, \mathcal{G}_t$

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

Codebook update with feedback. For each incorrect LM prediction, our expert provides the corrected label y, and a free-text rationale e explain-

[6] We use `Llama-3-70B` with different instructions for both generating predictions and refining the codebook.

ing their label. The current $\mathcal{G}_t$ is then updated by prompting an additional language model to incorporate e into $\mathcal{G}_{t+1}$. 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

Simulation Setup. To validate the usefulness and generalizability of our codebook development algorithm before we apply it to real-world new variables, we first evaluate it on a subset of existing variables by treating them as new variables that are not yet characterized using simulated expert feedback from an LM. To scale our experiments across multiple NVDRS variables, we use existing NVDRS abstractor labels y (see §2) as reference labels and LM-based expert feedback (e) as a proxy for human feedback. While we have reference y, we do not have corresponding reference reasoning, e. To address this, we simulate an expert reasoning via the LM's CoT (as generated in §2) for e (refer to further details in Appendix B). We apply our algorithm in this simulated setting to generate codebooks for 12 variables, using a subsample $\mathcal{D}$ of 150 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 each from low (0.6-0.7), medium (0.7-0.8) and 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 Appendix D.

We measure the effectiveness of the generated codebooks by measuring their corresponding LM prediction performance using the same prompt template and held out test set D_{balanced} from §2 and compare them to the performance achieved using the reference NVDRS codebooks.

Simulation Results. Paired t-tests ($p < 0.025$ under Bonferroni correction) across 12 variables show that our generated codebooks significantly outperform the NVDRS codebook for random and coverage-based sampling as shown in Figure 3. Furthermore, the maximum accuracy on $\mathcal{D}^t_{guide}$ is reached between 10-15 iterations, as shown in Figure 5 (see Figure 6 in Appendix D for results on all variables). We observe that our generated codebooks provide finer-grained instructions to the model by including examples from the narratives, which helps resolve ambiguities that may be underspecified in NVDRS codebooks (see Table 10 in the Appendix F).

Our simulation results provide preliminary evidence that our algorithm enables efficient collaboration between LMs and experts that result in effective codebooks. However, simulation results are not sufficient to validate the usefulness of our algorithm. In the following section, we apply our algorithm for a real-world case study to derive a codebook that is used to collect data-driven evidence for uncovering new intervention opportunities.

4 Case Study: Legal Interactions of Suicide Victims

Suicide prevention experts have identified legal professionals as part of a broader set of nonclinical '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), which primarily engages in *remedial* care, leaving opportunities for *preventive* care through interactions with nonclinical professions underexplored (Lybarger et al., 2023; Ralevski et al., 2024; Xu et al., 2024). A recent survey found that 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). However, NVDRS lacks structured variables that capture these interactions, preventing a direct analysis which could inform new preventive measures (e.g., training lawyers to spot risk factors in clients).

Identifying interactions with legal professionals is challenging because they are often indirectly implied by references to life disruptions (e.g. custody battles), where there is a lot of ambiguity (e.g., DUIs may not always involve legal interac-

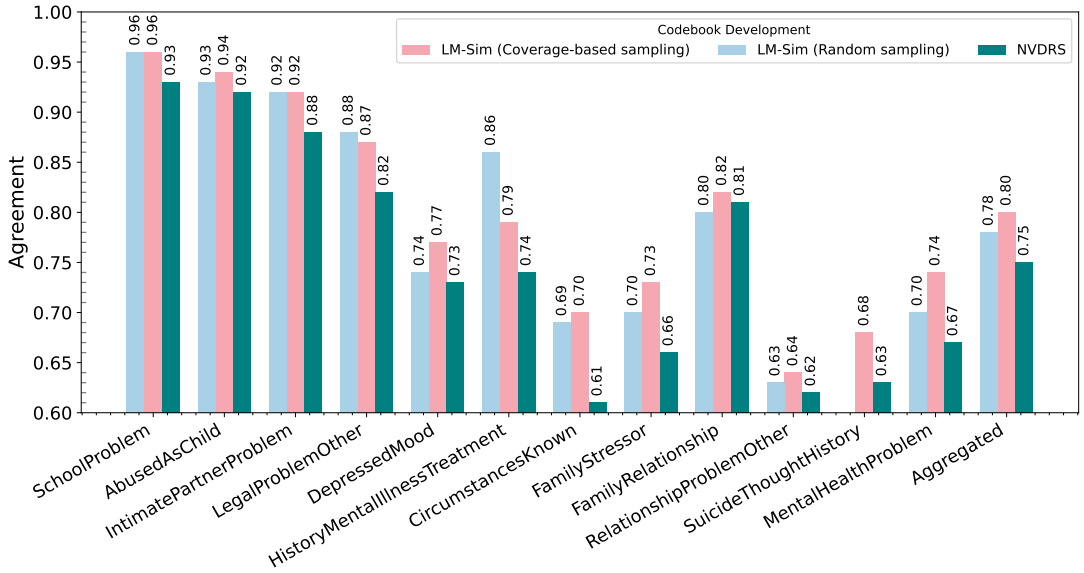

Figure 3: `Llama-3-70B` performance on D_{balanced} using the generated codebooks (LM-Sim) for random and coverage-based sampling on 12 variables, compared to performance achieved using the reference NVDRS codebook. Our generated codebooks from the simulated setting are just as effective as using the NVDRS codebook for all 12 variables as shown by the 'Aggregated' agreement (∼0.80 for LM-Sim vs ∼0.75 for NVDRS).

tions). Existing methods based on retrieval methods (Kafka et al., 2023, 2024) cannot capture these implicit instances. Therefore, we characterize interactions with legal professionals into three classes: explicit (direct interactions), implicit (indirect interactions inferred from life disruptions), and no interaction. Since feedback comes from a human expert for our algorithm in a real-world case study setting, we refer to our algorithm-based codebook development as human-in-the-loop (`HitL`) codebook development.

4.1 `HitL` Codebook Development

We apply our algorithm to NVDRS narratives to answer how many cases contain evidence of victim interactions with legal professionals, to determine whether such nonclinical contexts are relevant for prevention. Given a budget of 150 instances, and batch size of 5, we first sample narratives using FAISS (Douze et al., 2024) with the following keywords: '`lawyer, attorney`.' We start with a simple codebook ($\mathcal{G}_0$) as shown in Figure 1. In our pilot, the human expert exhausts the entire budget b.

4.2 Manual Codebook Development

We conduct this case study with a manual approach that does not leverage human-LM collaboration to create a held-out test set and assess the efficiency of our algorithm. Based on the expertise of the suicide prevention researcher on our team (who is a co-author of this work), we manually develop codebooks to identify and characterize victims' interactions with legal professionals (Halterman and Keith, 2024; Rytting et al., 2023). The process involves regular discussions between our expert and two trainees (CS graduate students), who independently analyze and annotate a sample of 150 narratives to delineate precise codebooks for legal interactions. This process is repeated twice on different sets of narratives until no further refinements are made to the codebooks. These guidelines are then used by all three to annotate 634 narratives ($D_{\text{expert\_legal}}$), as an evaluation set for our algorithm. On this set, we achieve high inter-annotator agreement: Krippendorff's $\alpha = 0.88$ (Krippendorff, 1970).

Model	$\mathcal{G}_0$	$\mathcal{G}_{\texttt{HitL}}$	$\mathcal{G}_{\textbf{expert}}$
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 for implicit-, explicit-, and no-interaction reported on $D_{\text{expert\_legal}}$ using no codebook ($\mathcal{G}_0$), the `HitL` generated codebook at the 25th iteration ($\mathcal{G}_{\texttt{HitL}}$), and the expert codebook ($\mathcal{G}_{\text{expert}}$). The generated codebook achieves performance on par with the expert codebook across all models.

4.3 `HitL` Results on Legal Interactions

In Table 2, we compare performance on $D_{\text{expert\_legal}}$ across three prompts: (*i*) $\mathcal{G}_{\textbf{0}}$: No Codebook, (*ii*) $\mathcal{G}_{\textbf{expert}}$: Expert Codebook (§4.2), and (*iii*) $\mathcal{G}_{\texttt{HitL}}$: `HitL` Codebook at the 25th iteration. We observe that $\mathcal{G}_{\texttt{HitL}}$ performance (Macro F_1) is on par with $\mathcal{G}_{\text{expert}}$ across all models suggesting that our guide-

lines can be generalized beyond the model that was used to develop them. Our `HitL` codebook development pilot took 3.5 hours to complete, compared to several weeks in the manual setting, and our results show that we do not compromise on annotation quality for victim-legal interactions.

Most importantly, our algorithm finds that **10.4% of all 270K suicide narratives in NVDRS contain evidence of legal interactions with victims**. More concretely, `Llama-3-70B`, prompted using our `HitL` codebook, detects implicit victim-lawyer interactions in 10% of the narratives and explicit interactions in 0.4% of the narratives (see Figure 7 in the Appendix E). These findings support the efficiency and effectiveness 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). Most importantly, our finding that interactions between suicide victims and legal professionals are relatively common, may provide the required empirical evidence for designing novel interventions in suicide prevention, such as training for legal professionals. We emphasize that such efforts should be pursued through collaboration with domain experts, where LMs should augment (e.g. via validation) rather than replace human expertise.

5 Related Work

Characterizing NVDRS Data using NLP. Despite the scale of NVDRS narratives, relatively few studies have examined the effectiveness of LMs for characterizing them (Dang et al., 2023). Most prior work has focused on using AI tools to identify risk factors such as social determinants of health using electronic health record (EHR) and medical notes (Kirtley et al., 2022; Lejeune et al., 2022; Ehtemam et al., 2024; Melia et al., 2025; Johns et al., 2023; Zhou et al., 2023; Consoli et al., 2024; Guevara et al., 2024; Wang et al., 2025), which are limited to individuals who have engaged with healthcare systems. For NVDRS narratives, NLP tools have been used to characterize these narratives by narrative length and quality (Arseniev-Koehler et al., 2023) and variation across decedents' demographics (Chance et al., 2025). Others uncover latent themes (Arseniev-Koehler et al., 2022, 2021; Davidson et al., 2021) or classify circumstance and crisis variables from narratives (Wang et al., 2023; Zhou et al., 2023; Kafka et al., 2023; Parker, 2025b). Prior work typically frame LMs as annotation assistants, without enabling the discovery of *new* intervention opportunities. Motivated by these limitations, we examine whether LMs can serve as effective assistants to data abstractors and experts (see Appendix G for further related work).

Qualitative Coding with LMs. Prior work has explored the use of language models 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). Despite these advances, it remains unclear how expert judgment can be effectively integrated into LM-assisted codebook development for real-world, high-stakes domains (Wang et al., 2024a; Ziems et al., 2024). In particular, prior work has not examined LM-assisted codebook refinement in the context of NVDRS, which presents unique challenges due to the sensitivity and heterogeneity of the data, which to our knowledge our work is the first attempt towards.

6 Conclusion

We introduce LM assistants to help human experts efficiently develop codebooks and human abstractors to accurately annotate structured NVDRS variables. Our LM assistant achieves high agreement with NVDRS abstractors while surfacing annotation inconsistencies, and our algorithm produces codebooks that result in annotation performance on par with using the reference codebooks. We further apply our algorithm to a real-world case study on a new variable: interactions between suicide victims and legal professionals. Our findings about the high prevalence of such interactions could open intervention avenues in suicide prevention research in future work. Recognizing the potential risks of using LMs in high-stakes domains, we analyze failure modes to inform responsible practitioner use and advise against relying on LMs as peer validators for low-agreement variables, particularly those related to mental health. Also, models should flag high uncertainty predictions or abstain altogether, prioritizing safety over coverage. Finally, codebook development should always be supervised by domain experts as LM-only approaches pose serious safety risks in high-stakes contexts, thus the success of such systems depend on effective human-LM collaboration, where LMs augment rather than replace human expertise.

Limitations

We show that LM assistants can support the large-scale validation of sensitive death narrative annotations, and enable the discovery of data-driven evidence for new intervention opportunities. However, given the scope, real-world implications, and unique challenges of the suicide prevention domain, we acknowledge some important limitations.

First, the evaluation of our real-world case study relies on 634 expert-annotated narratives, which required 3 months to annotate. Obtaining ground-truth validation for all model predictions across 270K narratives was infeasible at this rate, given the emotionally demanding nature of the task.

Another limitation of our work is the scope of our empirical evaluation—our real-world case study intentionally focuses on a single hypothesis within suicide prevention: victim interactions with legal professionals. A larger range of non-clinical factors may also be relevant for prevention. More broadly, while our codebook development algorithm is not domain specific, evaluating its effectiveness beyond the suicide prevention context was outside the scope of this work.

Additionally, while our codebook development algorithm supports iterative guideline development, it does not incorporate multi-turn interaction between experts and LMs in an effort to minimize expert burden. However, we recognize that this design choice may have limited opportunities to clarify codebooks for ambiguous cases.

We do not explicitly assess whether codebook development achieves thematic saturation: the point at which the codebooks sufficiently capture all relevant cases observed for a given variable (Glaser and Strauss, 1967). In practice, assessing thematic saturation is challenging as it would require additional human-in-the-loop experiments across many hyperparameter settings, which was not feasible for our case study. In the future, we recommend practitioners take into account the subjectivity of the variable and the manual effort involved when determining the stopping conditions and hyperparameters (i.e., budget, target performance and the minimum size of $\mathcal{D}^t_{guide}$ and $\mathcal{D}_{val}$), which affect how quickly thematic saturation is achieved.

Our evaluation in §2.3 is limited to a subset of 50 binary NVDRS variables that occur in at least 300 cases, and that can be inferred from the narrative data which we have access to. As a result, we do not evaluate the full set of 600+ NVDRS variables which are derived from external data sources such as death certificates or toxicology reports which we do not have access to. This constraint may limit the applicability of our findings to only those variables that were evaluated in our study.

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). This project was reviewed by our Institutional Review Board, who 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.

Annotator Well-Being. Given the sensitive nature of suicide death narratives, we implemented several precautions in line with current best practices (Dasgupta, 2021) 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.).

It is important to emphasize that NVDRS data *rarely* contain suicide notes (Pestian et al., 2012; Rockett et al., 2018); in some scant instances, only paraphrases may appear. 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).

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 collabora-

tion with the annotators. All annotators independently labeled the same set of narratives, achieving strong inter-annotator agreement (Krippendorff's $\alpha = 0.88$). Disagreements were resolved through consensus discussions. Given the sensitive nature of suicide death narratives, we implemented multiple safeguards to support annotator well-being (see previous paragraph). 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.

Privacy. The NVDRS data is already de-identified for protecting the privacy of the victims and their survivors. 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.

References

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

Alina Arseniev-Koehler, Jacob Gates Foster, Vickie M Mays, Kai-Wei Chang, and Susan D Cochran. 2021. 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.

Alina Arseniev-Koehler, Vickie M Mays, Jacob G Foster, Kai-Wei Chang, and Susan D Cochran. 2023. Gendered patterns in manifest and latent mental health indicators among suicide decedents: 2003–2020 national violent death reporting system (nvdrs). *American journal of public health*, 114(S3):S268–S277.

Amanda Barany, Nidhi Nasiar, Chelsea Porter, Andres Felipe Zambrano, Alexandra L Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, and 1 others. 2024. 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.

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

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

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

Christina Chance, Alina Arseniev-Koehler, Vickie M Mays, Kai-Wei Chang, and Susan D Cochran. 2025. Measuring narrative complexity among suicide deaths in the national violent death reporting system (2003–2021 nvdrs). *Information*, 16(11):989.

Bernardo Consoli, Xizhi Wu, Song Wang, Xinyu Zhao, Yanshan Wang, Justin Rousseau, Tom Hartvigsen, Li Shen, Huanmei Wu, Yifan Peng, and 1 others. 2024. Sdoh-gpt: Using large language models to extract social determinants of health (sdoh). *arXiv preprint arXiv:2407.17126*.

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

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

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

Judy E Davidson, Gordon Ye, Felicia Deskins, Heather Rizzo, Christine Moutier, and Sidney Zisook. 2021. 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.

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

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

Houriyeh Ehtemam, Shabnam Sadeghi Esfahlani, Alireza Sanaei, Mohammad Mehdi Ghaemi, Sadrieh Hajesmaeel-Gohari, Rohaneh Rahimisadegh, Kambiz Bahaadinbeigy, Fahimeh Ghasemian, and Hassan Shirvani. 2024. Role of machine learning algorithms in suicide risk prediction: a systematic review-meta analysis of clinical studies. *BMC medical informatics and decision making*, 24(1):138.

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

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

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

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

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

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

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.

Julie M Kafka, Kathryn Elizabeth Moracco, Brian W Pence, Pamela J Trangenstein, Mike Dolan Fliss, and Luz McNaughton Reyes. 2024. Intimate partner violence and suicide mortality: a cross-sectional study using machine learning and natural language processing of suicide data from 43 states. *Injury Prevention*, 30(2):125–131.

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

Olivia J Kirtley, Kasper van Mens, Mark Hoogendoorn, Navneet Kapur, and Derek De Beurs. 2022. Translating promise into practice: a review of machine learning in suicide research and prevention. *The Lancet Psychiatry*, 9(3):243–252.

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

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.

Alban Lejeune, Aziliz Le Glaz, Pierre-Antoine Perron, Johan Sebti, Enrique Baca-Garcia, Michel Walter, Christophe Lemey, and Sofian Berrouiguet. 2022. Artificial intelligence and suicide prevention: a systematic review. *European psychiatry*, 65(1):e19.

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

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

Ruth Melia, Katherine Musacchio Schafer, Megan L Rogers, Emma Wilson-Lemoine, and Thomas Ellis Joiner. 2025. The application of ai to ecological momentary assessment data in suicide research: Systematic review. *Journal of Medical Internet Research*, 27:e63192.

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.

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

Susan T Parker. 2025a. Assessing supervised natural language processing (nlp) classification of violent death narratives: Development and assessment of a compact large language model (llm) approach. *medRxiv*, pages 2025–01.

Susan T Parker. 2025b. Supervised natural language processing classification of violent death narratives: Development and assessment of a compact large language model. *JMIR AI*, 4:e68212.

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

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

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

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

Jaspreet Ranjit, Brihi Joshi, Rebecca Dorn, Laura Petry, Olga Koumoundouros, Jayne Bottarini, Peichen Liu, Eric Rice, and Swabha Swayamdipta. 2024. OATH-frames: Characterizing online attitudes towards homelessness with LLM assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13033–13059, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. 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.

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, and 1 others. 2018. Discerning suicide in drug intoxication deaths: Paucity and primacy of suicide notes and psychiatric history. *PLoS one*, 13(1):e0190200.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, and 1 others. 2023. Towards coding social science datasets with language models. ArXiv:2306.02177.

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

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

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

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2024a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *arXiv preprint arXiv:2402.01908*.

Song Wang, Yifang Dang, Zhaoyi Sun, Ying Ding, Jyotishman Pathak, Cui Tao, Yunyu Xiao, and Yifan Peng. 2023. 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.

Song Wang, Yishu Wei, Haotian Ma, Max Lovitt, Kelly Deng, Yuan Meng, Zihan Xu, Jingze Zhang, Yunyu Xiao, Ying Ding, and 1 others. 2025. A multi-stage large language model framework for extracting suicide-related social determinants of health. *Communications Medicine*, 5(1):404.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. 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.

Richard Li Xu, Song Wang, Zewei Wang, Yuhan Zhang, Yunyu Xiao, Jyotishman Pathak, David Hodge, Yan Leng, S Craig Watkins, Ying Ding, and 1 others. 2024. Analyzing social factors to enhance suicide prevention across population groups. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 189–199. IEEE.

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

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

Appendix

A Agreement Across Models

We use LMs as annotation assistants to label 36 circumstance and 14 crisis variables in NVDRS death narratives in §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 4. 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).

A.1 Surfacing Annotation Inconsistencies with LMs

The suicide prevention expert analyzed 300 individual examples sampled from D_balanced across six variables where `Llama-3-70B` either disagreed or agreed with the original human annotation. We calculated the overall proportion in the disagreement / agreement cases where the expert agreed with the model. Our bootstrap hypothesis test for equality of means accounts for sampling variability and shows a statistically significant difference ($p < 0.05$) between the 38% of cases where the model surfaced annotation inconsistencies on disagreement cases and the 13% rate of joint human-model inconsistencies on agreement cases. We make the assumption that small shifts in model agreement (within the 9% CI range) are unlikely to shift the underlying error patterns substantially and would only affect the variability in the number of samples that constitute the distinct 13% and 38% conditional probabilities. While simulating robustness under lower model accuracy is possible (e.g., randomly flipping predictions to reduce agreement and directly recalculating 'model corrects human' and 'joint error' rates), this would introduce a new assumption that accuracy changes occur via random prediction flips rather than a more systematic variance that occurs for ambiguous or difficult samples. Therefore, we rely on the results of our bootstrap hypothesis test as a signal of robustness.

B LM-Simulated Codebook Development

The expert feedback e we use in the simulated setting is the LM CoT, and in order to avoid faulty e, we only keep the samples for which the LM predictions matched the abstractor labels. This can be considered a disadvantage to the codebook that gets generated from our codebook development algorithm compared to the reference codebook because the algorithm is limited to receiving feedback on potentially easy cases for which the LM was already able to predict correctly using the reference codebook. However, despite this disadvantage, the generated codebook outperforms the reference codebook as shown in the results in Figure 3. We suspect this stronger performance to be attributed to our algorithm taking a systematic approach to refining the codebook based on individual examples and therefore providing better granular instructions until a performance threshold is reached, as opposed to the reference codebook which is developed more holistically and therefore overlook details.

Figure 5 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 $\mathcal{D}_{guide}$.

C 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[7](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.

[7] https://sbert.net/index.html

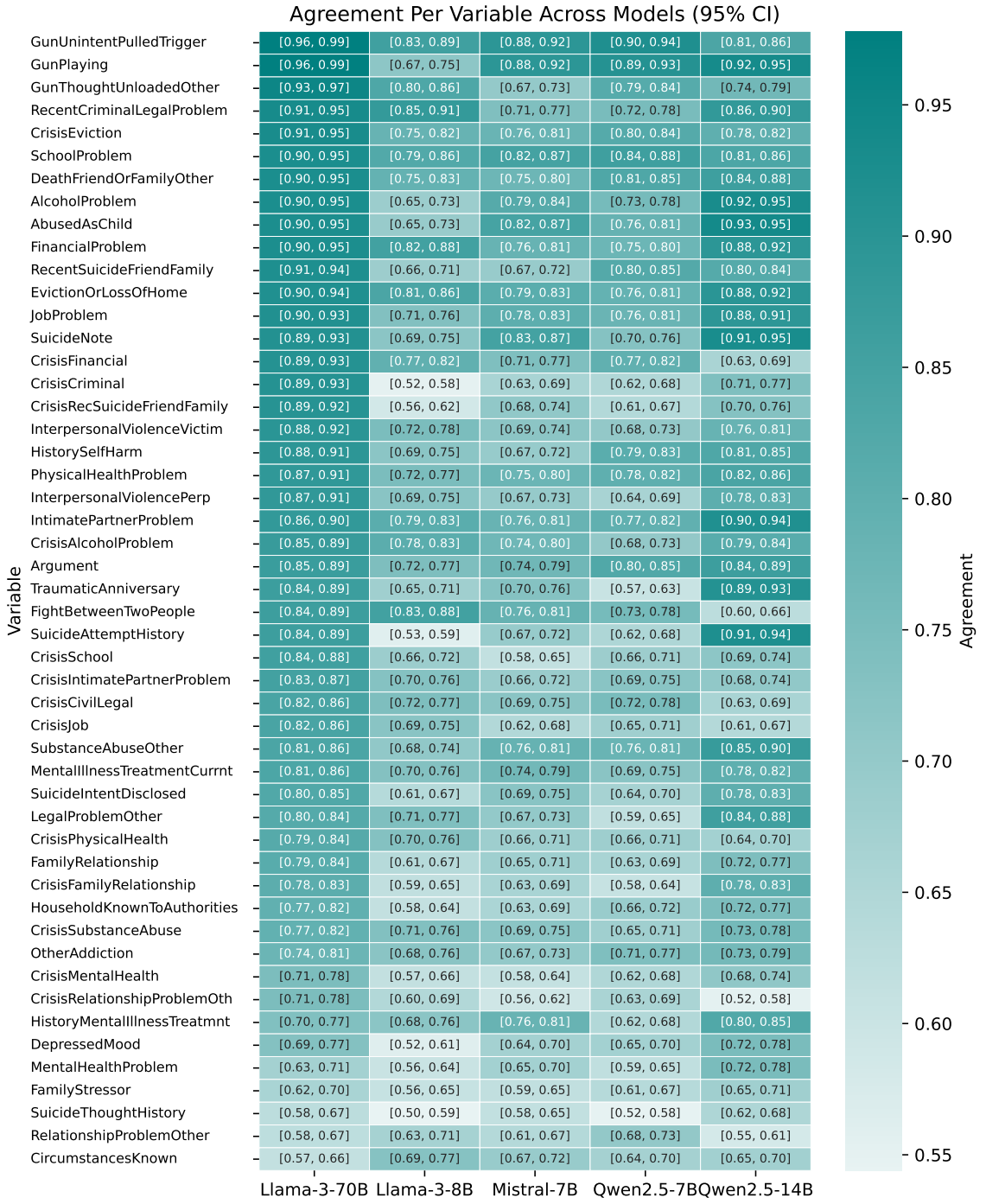

Figure 4: Per variable agreement (95% confidence intervals; bootstrapped with 10K iterations) for for 50 NVDRS variables across different models. Agreement is reported on D_{balanced}. We find that on average, `Llama-3-70B` has the highest agreement with data annotators out of all evaluated models.

D 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 codebook 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

Variable	Agreement	TPR	FPR	FNR
GunUnintentPulledTrigger	0.978	0.984	0.028	0.016
GunPlaying	0.976	0.968	0.016	0.032
GunThoughtUnloadedOther	0.954	0.992	0.084	0.008
RecentCriminalLegalProblem	0.934	0.908	0.040	0.092
CrisisEviction	0.932	0.924	0.060	0.076
SchoolProblem	0.926	0.996	0.144	0.004
HistoryMentalIllnessTreatmnt	0.736	0.484	0.012	0.516
DepressedMood	0.732	0.508	0.044	0.492
MentalHealthProblem	0.672	0.352	0.008	0.648
FamilyStressor	0.662	0.592	0.268	0.408
SuicideThoughtHistory	0.628	0.268	0.012	0.732
RelationshipProblemOther	0.624	0.964	0.716	0.036
CircumstancesKnown	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, [0.79, 0.85]**	**0.12**
Qwen2.5-14B	0.81, [0.80, 0.82]	0.09
Qwen2.5-7B	0.56, [0.54, 0.58]	0.17
Mistral-7B	0.67, [0.65, 0.69]	0.14
Llama-3-8B	0.54, [0.52, 0.56]	0.17

Table 4: Mean agreement, 95% confidence intervals, 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}).

Config	Assignment
models	**Meta-Llama-3-70B-Instruct** Number of parameters: 70B
	Meta-Llama-3-8B-Instruct Number of parameters: 8B
	Mistral-7B-Instruct-v0.2 Number of parameters: 7B
	Qwen2.5-14B-Instruct-1M Number of parameters: 14B
	Qwen2.5-7B-Instruct-1M Number of parameters: 7B
GPU	NVIDIA RTX A6000
# of GPUs	4 (inference)

Table 5: We provide details about the models used for our experiments in §2.3 and §3.2

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 updating the codebook 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.

E Case Study: Legal Interactions

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 codebooks (left) and our `HitL` codebooks (right).

Figure 7 (left) shows the distribution of narratives with implicit-, explicit-, and no interactions across three data splits. In Figure 7 (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.

We plot the F_1 on $\mathcal{D}^t_{guide}$ and $\mathcal{D}_{val}$ across iterations for all three interaction types in Figure 8.[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

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

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

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 LM-Sim and `HitL` across various parameters.

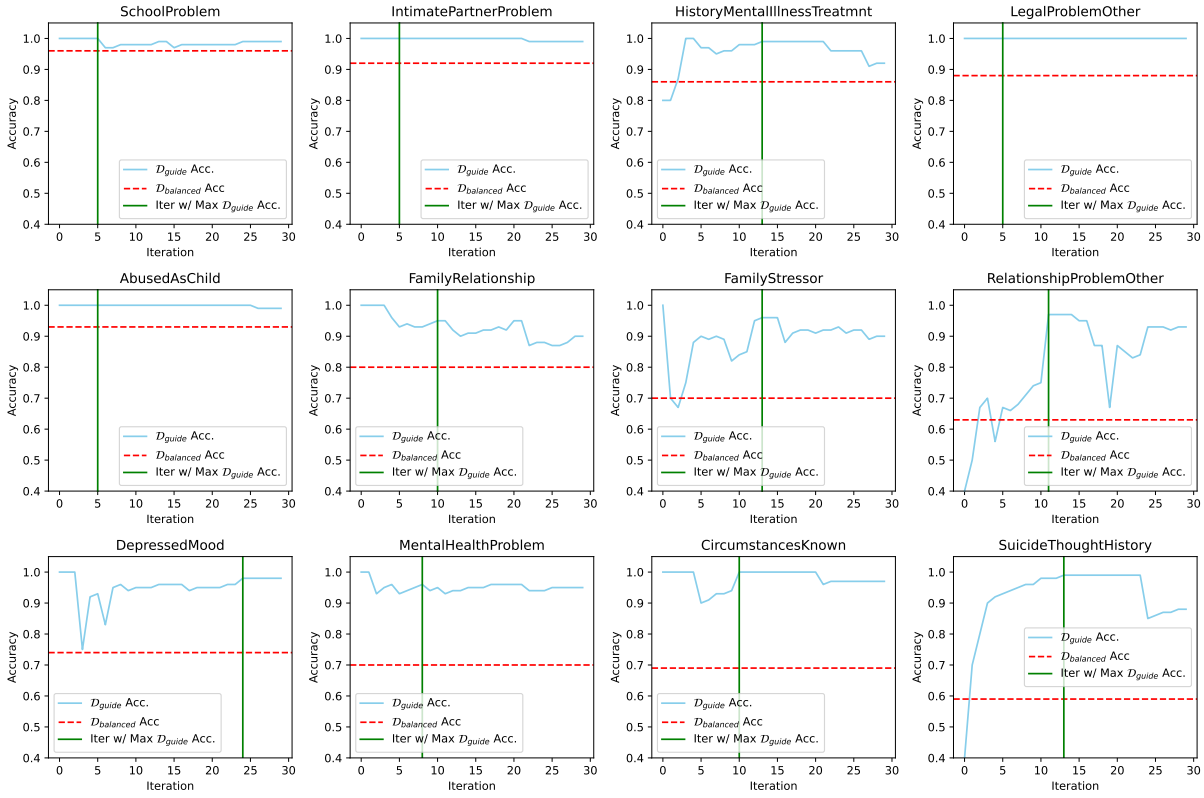

Figure 5: Accuracy on $\mathcal{D}_{guide}$ across 30 iterations for LM-Simulated codebook development for 12 NVDRS variables. The maximum accuracy on $\mathcal{D}_{guide}$ is reached between 10-15 iterations for all variables using random sampling per iteration.

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.

of the guidelines in the codebook pertain to implicit interactions, as shown in Table 9, resulting in limited instruction to identify explicit interactions.

F 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.

G Further Related Work

Prior work has explored the use of language models (Parker, 2025a) to extract a range of suicide-related circumstances from unstructured narratives, including infrequent or highly specific variables such as circumstances preceding female firearm suicide (Zhou et al., 2023) and intimate partner violence (Kafka et al., 2023). Alternatively, Wang et al. (2023) finetune a BERT model to classify circumstance and crisis variables from narratives, while Zhou et al. (2023) use language models to identify infrequent circumstances preceding female firearm suicide using a manually developed codebook. Kafka et al. (2023) applies supervised learning to detect intimate partner violence (IPV), but found that their approach does not capture implicit

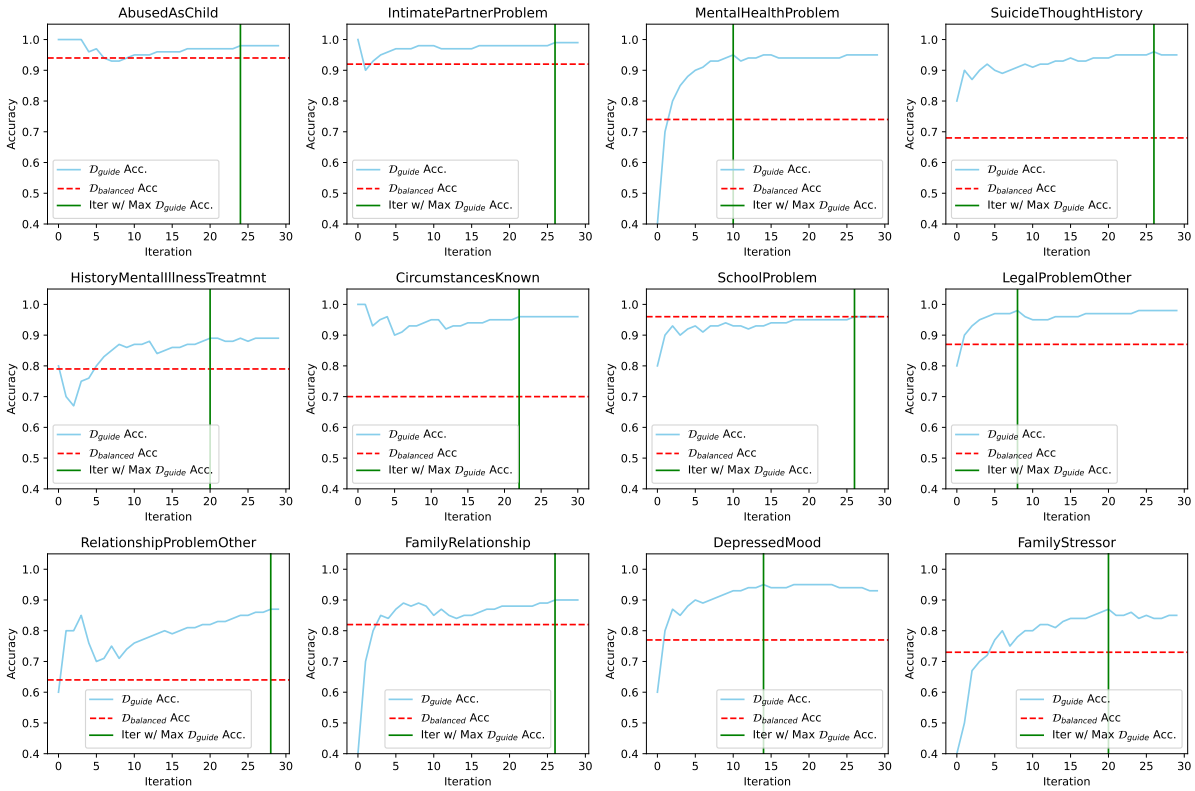

Figure 6: Accuracy on $\mathcal{D}_{guide}$ across 30 iterations for LM-Simulated codebook development for 12 NVDRS variables using coverage based sampling per iteration.

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.

references due to long narrative lengths. For example, Consoli et al. (2024) and Guevara et al. (2024) use LMs with in-context learning to identify SDoH in electronic health record (EHR) and medical notes.

These efforts largely treat existing annotations

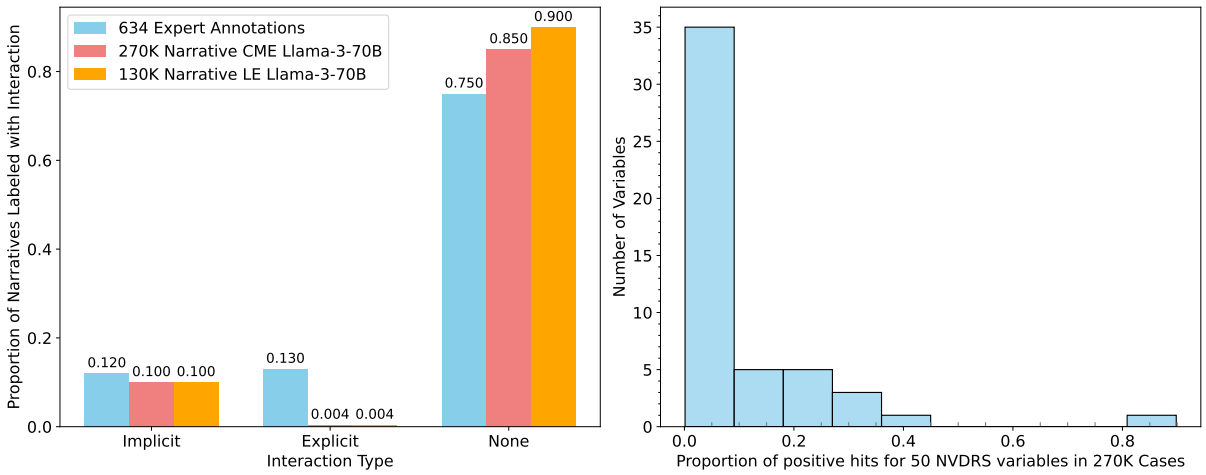

Figure 7: Distribution of narratives containing implicit-, explicit- and no-interaction for 3 data splits - 634 experts ($D_{\text{expert\_legal}}$), 270K CME narratives, and 130K LE narratives (left). Distribution of proportion of positive occurrences for 50 variables in 270K NVDRS cases (right).

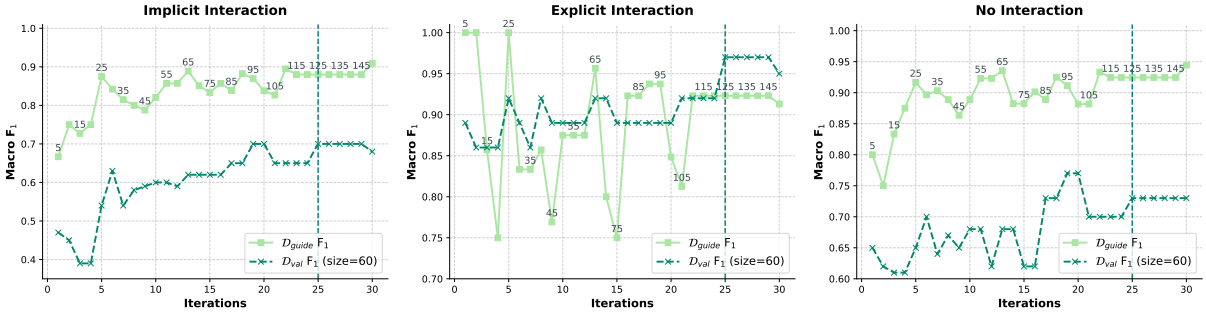

Figure 8: `Llama-3-70B` performance on $\mathcal{D}_{val}$ and $\mathcal{D}^t_{guide}$ across 30 iterations for human-in-the-loop codebook development for detecting victim-lawyer interactions. Data labels for $\mathcal{D}_{guide}$ represent the cumulative size of $\mathcal{D}^t_{guide}$ at each iteration t. Max overall macro F_1 on $\mathcal{D}_{val}$ is reached by the 25th iteration (Macro F_1 of 0.8).

and codebooks as fixed, and do not quantify how model performance varies across variables or characterize corresponding failure modes. More importantly, prior work typically frames LMs as annotation assistants, without enabling the discovery of *new* intervention opportunities. Motivated by these limitations, we examine whether LMs can serve as effective assistants to (i) data annotators by peer-validating existing annotations and surfacing potential discrepancies, and (ii) experts by supporting the development of codebooks for annotating new variables, surfacing new intervention opportunities not captured in the structured data.

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. [truncated ...]

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

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