SynthEHR-Eviction: Enhancing Eviction SDoH Detection with LLM-Augmented Synthetic EHR Data

Anonymous ACL submission

Abstract

Eviction is a significant yet understudied social determinants of health (SDoH), linked to housing instability, unemployment, and mental health. While eviction appears in unstructured electronic health records, it is rarely coded in structured fields, limiting downstream applications. We introduce *SynthEHR-Eviction*, a scalable pipeline combining LLMs, human-in-the-loop annotation, and automated prompt optimization (APO) to extract eviction statuses from clinical notes. Using this pipeline, we created the largest public eviction-related SDoH dataset to date, comprising 14 fine-grained categories. Fine-tuned LLMs (e.g., Qwen, LLaMA) trained on SynthEHR-Eviction achieved Macro-F1 scores of 88.8% (eviction) and 90.3% (other SDoH) on human validated data, outperforming GPT-4o-APO (87.8%, 87.3%), GPT-4o-mini-APO (69.1%, 78.1%), and BioBERT (60.7%, 68.3%), while enabling cost-effective deployment across model sizes. The pipeline reduces annotation effort by over 80%, accelerates dataset creation, enables scalable eviction detection, and generalizes to other information extraction tasks.

1 Introduction

Modern health care aims at whole-person care and at access to social services (National Center for Complementary and Integrative Health, n.d.; Bokhour et al., 2020). Social and environmental factors matter as much as clinical ones. The World Health Organization defines social determinants of health (SDoH) as the conditions in which people are born, grow, live, work, and age (Organization, n.d.). These factors drive most health outcomes (Hood et al., 2016). Housing and transportation are common examples: unstable housing harms physical and mental health (Tsai et al., 2021; Chen et al., 2022), and poor transportation limits access to care and food (Tsai et al., 2024; Losada-Rojas et al., 2024).

Yet, in current electronic health records (EHRs), SDoH information is often embedded within unstructured free-text clinical notes, lacking systematic, structured documentation (Guevara et al., 2024). This results in "information silos", where crucial social risk factors are underutilized in real-world healthcare delivery, limiting clinicians' ability to assess patients' broader health risks and restricting healthcare institutions and policymakers from designing proactive, targeted intervention strategies (Wikipedia contributors, n.d.).

Among various categories of SDoH, eviction is a highly impactful but long-overlooked factor affecting patient health. Eviction can trigger a chain reaction of adverse outcomes, including housing instability, unemployment, homelessness, mental health issues, and long-term poverty (Desmond and Kimbro, 2015; Desmond, 2016). Its impact extends beyond individuals, affecting entire communities and exacerbating social inequality (Tsai and Huang, 2019). Studies have shown that marginalized populations—such as racial minorities, women, and low-income renters—face a significantly higher risk of eviction (Hepburn et al., 2020). Particularly during the Coronavirus Disease-2019 (COVID-19) pandemic, eviction was directly associated with over 430,000 additional cases and more than 10,000 deaths in the United States, prompting the CDC to enact a federal eviction moratorium between 2020 and 2021 (Tsai, 2024). Despite its profound public health implications, the current healthcare system's ability to identify and address eviction-related risks remained severely limited. Mainstream public EHR datasets, such as MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2020), contain extremely sparse documentation of eviction-related information, with relevant terms appearing in fewer than 1 in 1,000 records. Furthermore, standardized coding systems, such as relying on the International Classification of Diseases, Version 10 (ICD-10), lack specific codes for eviction, con-

tributing to fragmented and underutilized eviction data in clinical practice (for Medicare & Medicaid Services , CMS). Unlocking and systematically incorporating eviction-related SDoH data can transform healthcare delivery by enabling comprehensive health risk assessments, integrating social work, housing support, and psychological interventions into patient care workflows, and providing real-time, high-quality data to support targeted public health interventions, particularly benefiting marginalized populations and advancing accessible healthcare and social service systems.

We present *SynthEHR-Eviction*, a scalable *augment–annotate* pipeline that combines automated prompt optimization (APO) (Ramnath et al., 2025), DSPy (Khattab et al., 2023), large language models (LLMs), and a human-in-the-loop (HITL) check. The pipeline runs in two simple stages and repeats: the augmenter first creates new notes for underrepresented labels and updates prompts with expert feedback to rebalance the distribution; the annotator then verifies labels, refines spans, and writes short chain-of-thought (CoT) reasoning traces with DSPy, which supports structured CoT and fine-grained label editing. This design replaces full manual rewriting and cuts human hours by >80% while keeping quality and interpretability. It is modular and can transfer to other SDoH such as food insecurity, utility shutoff, or intimate partner violence (Magister et al., 2022). Using this pipeline, we build a public dataset with 14 clearly defined SDoH categories. Eviction has fine-grained states like *Eviction Absent*, *Eviction Pending*, and *Mutual Rescission History* (see Table 7). We also include related ICD-10 Z.59 categories such as housing instability and homelessness (see Table 8). In total, the release has 8,000 synthetic train/dev instances and 616 expert-annotated test instances from real and synthetic notes, drawn from MIMIC-IV (Johnson et al., 2020) and PMC-Patients (Zhao et al., 2023) (see Table 3). This mix supports cross-domain testing and downstream fine-tuning of open-source LLMs.

To summarize, our contributions:

- **Largest public eviction-focused SDoH resource with a clear schema.** We release a fine-grained dataset with 14 categories, including eviction states (*Eviction Absent*, *Eviction Pending*, *Mutual Rescission History*) and related Z.59 labels (e.g., housing instability, homelessness). The package includes 8,000 synthetic train/dev notes and 616 expert-checked test notes across MIMIC-IV and PMC-Patients, with labeling guidelines and tabled definitions (Tables 7, 8, 3). It enables fair, low-cost cross-domain evaluation and housing-risk research at scale.
- **A reusable APO+DSPy+HITL pipeline with CoT supervision.** Two-stage *augment→annotate* loops use expert-in-the-loop prompt updates to balance long-tail labels, DSPy to generate and store CoT traces and to refine labels, and simple span checks for quality. This reduces human effort by >80% while keeping interpretable rationales and stable label rules.
- **Strong results and transferable synthesis insights for BioNLP IE.** Models trained on *SynthEHR-Eviction* outperform strong baselines on both synthetic and real EHR notes, and fine-tuned small open-source models approach larger API systems while keeping CoT-based interpretability. We also report two simple knobs that generalize to similar IE tasks: mix a small portion of real notes to boost out-of-domain transfer, and target a 3k–5k sample range before returns taper; moreover, CoT provides larger gains for smaller models than for 7–8B models.

2 Related Work

Public SDoH datasets focus on common categories such as smoking, substance use, alcohol, and homelessness (Patra et al., 2021; Bompelli et al., 2021; Yu et al., 2022; Mitra et al., 2023; Han et al., 2022). In public EHR corpora like MIMIC-III/IV, eviction mentions are very sparse, and ICD-10 has no specific code (Johnson et al., 2016, 2020; for Medicare & Medicaid Services , CMS). Prior eviction work built models on private data (e.g., VHA) but did not release resources (Yao et al., 2023b). Synthetic SDoH corpora exist, but many omit eviction and lack temporal or legal subtypes (Mitra et al., 2024). We fill this gap with a *public*, fine-grained schema and with expert-checked labels that mark present vs. history and related states.

LLMs can generate and label data at low cost through few-shot generation, filtering and relabeling, adversarial diversification, and human–AI co-annotation (Yoon et al., 2023; Sahu et al., 2022; Hartvigsen et al., 2022; Chung et al., 2023). In clinical NLP, knowledge-infused prompting and data ↔ label generation improve control and specificity (Xu et al., 2023a; Li et al., 2023b). SDoH extraction with instruction-tuned or fine-tuned models reduces cost and improves agreement and fair-

ness (Guevara et al., 2024). But rare, temporally nuanced labels like eviction still need class balancing and legally grounded phrasing. Our augment–annotate loop targets these needs and adds CoT to expose decision logic that helps smaller models.

Prompt optimization spans soft prompts, LoRA, and gradient search such as AutoPrompt (Li and Liang, 2021; Lester et al., 2021; Hu et al., 2021; Shin et al., 2020). RL-based and gradient-free strategies work for closed models via edit, back-translation, evolution, and LLM rewriting (Deng et al., 2022; Zhang et al., 2022b; Prasad et al., 2022; Xu et al., 2022; Guo et al., 2023; Fernando et al., 2023; Zhou et al., 2022; Xu et al., 2023b; Pryzant et al., 2023; Yang et al., 2023; Wang et al., 2023; Dong et al., 2023; Li et al., 2023a; Sun et al., 2023). DSPy decouples task logic and prompt policy and provides reusable optimizers (Khattab et al., 2023). We adopt this stack and combine it with HITL checks and CoT supervision to keep interpretability while reducing labor.

3 Methods

Our method follows the label schema in Figure 2 and the pipeline in Figure 1. The system has two stages: *Augment* then *Annotate*. We use DSPy with Chain-of-Thought (CoT), exact-match evaluation, and a few-shot optimizer (Khattab et al., 2023; Ramnath et al., 2025). Full details are in Appendix A.

Schema (Figure 2). We place *Eviction* under Z59.89 and use 14 classes. These include five eviction states (Absent, Hypothetical, Pending, Mutual Rescission, Present) with time qualifiers (Current, History, Future), plus nine related Z59 items. We give label rules and examples in Tables 7 and 8 (Appendix A.1.2).

Stage 1: Augment (Figure 1, left/bottom-left). We extract *social history* from MIMIC notes and pass each note to a label-specific augmenter. Each augmenter is an LLM with an optimized prompt (APO) that rewrites the note to reflect a target label while keeping clinical style. Human experts check outputs and give feedback to improve prompts. We repeat until accuracy is stable. See Algorithm 11 and Appendix A.1, A.1.3.

Stage 2: Annotate (Figure 1, middle/bottom-right). A small, expert-validated seed set trains a DSPy program on GPT-4o or GPT-4o-mini. We first run a binary gate (Eviction vs. Non-Eviction). Then we run one of two multi-class annotators: an Eviction annotator (7 labels) or a Non-Eviction annotator (7 labels). Each prediction includes a short rationale (CoT). For quality, we use a three-pass consensus rule: if the three outputs agree, we keep the consensus label; if not, we discard the case and return the note to the pool. See Algorithm 11 and Table 2 (Appendix A.2).

Data package. We release synthetic train/dev and expert-checked test notes across MIMIC-IV and PMC-Patients; counts and splits are in Table 3. We use the same DSPy settings for Eviction and Non-Eviction, and only switch datasets and label sets (Appendix A).

Fine-tuning open-source models. We turn the validated set into supervised data and fine-tune BERT and small LLMs. We use parameter-efficient fine-tuning (LoRA via Unsloth) for LLaMA and Qwen, and standard HuggingFace fine-tuning for BERT. Hyperparameters and data mixes are in Appendix A.3.

4 Experimental Setup

We align the evaluation with the pipeline in Figure 1. We run three steps in order: **Step 1** binary gate (Eviction vs. Non-Eviction), **Step 2** Eviction multi-class, and **Step 3** Non-Eviction multi-class. A prediction is counted correct only if Step 1 is correct and the matching Step 2 or Step 3 label is also correct. Metrics are Micro-F1, Macro-F1, and MCC. Full settings appear in Appendix B.

Models and runs. For DSPy runs (GPT-4o/mini backends) we use temperature $= 0$ to annotate, then four extra runs at temperature $= 0.5$ to form 95% CIs. For other LMs/LLMs we vary the random seed across five runs. We keep fine-tuning schedules fixed across sizes and models. We follow PhysioNet guidance for MIMIC use and call Azure OpenAI with human review disabled (see Appendix B.1).

Dev sets and averaging. We report results on three dev sets: Synth, MIMIC, and PMC. We take the simple average of the three to show cross-domain behavior. Keywords for MIMIC/PMC filtering and case construction are in Table 12. See Appendix B.1.

Data size study. We test five training sizes: 200, 1k, 3k, 5k, and 10k samples. BERT sees labels only.

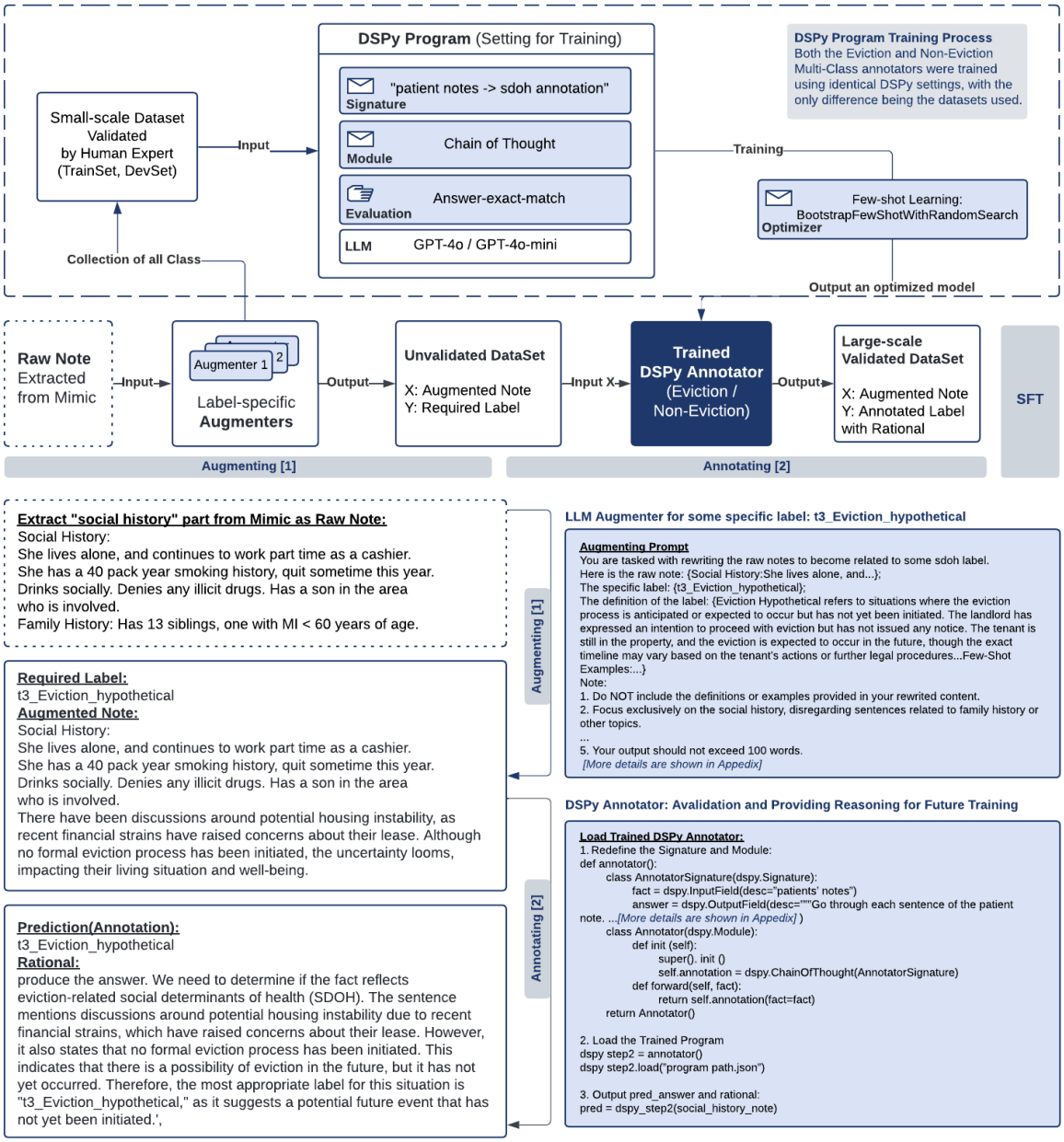

Figure 1: A complete DSPy-based system for training annotators to identify potential eviction cases from clinical notes. **Upper Section (Training Process)**: A small-scale human-validated dataset feeds into the DSPy Program training process, which employs GPT-4 or GPT-4o-mini models. The framework uses Chain of Thought reasoning and answer-exact-match evaluation, with few-shot learning optimization to produce a trained annotator model. **Lower Section (Pipeline Flow)**: **Augmentation Stage [1]**: Raw clinical notes extracted from MIMIC database undergo augmentation through label-specific augmenters to be the augmented notes dataset. **Annotation Stage [2]**: The trained DSPy Annotators (supporting Eviction/Non-Eviction classification separately) processes these augmented notes, outputting a large-scale validated dataset with annotated labels and accompanying rationales. **Bottom Section (Examples)** demonstrates a sample augmentation process for the t3_Eviction_hypothetical label and a detailed DSPy Annotator implementation and output format. **Quality Control**: The example shown represents a case where the annotated label matches the required label from augmentation, allowing direct inclusion in the final validated dataset. However, the system also handles annotation inconsistencies: If the annotated label differs from the required label, the system performs two additional annotation passes. When all three annotation outputs are consistent, the case is included in the final dataset with the annotated label (even if different from the original augmented label). If annotation outputs remain inconsistent, the case is discarded from the current validation cycle and the augmented note is returned to the raw note pool for potential future use. The entire framework maintains consistent settings across both Eviction and Non-Eviction Multi-Class annotators, with the only difference being the specific datasets used for training and test.

LLMs see labels and rationales. We use the same schedule for each size. Design is in Appendix B.1.

Reasoning ablation. We train LLaMA-3.1-8B, LLaMA-3.2-3B, Qwen2.5-7B, and Qwen2.5-3B with and without rationales, while keeping all hyperparameters the same. We report Micro-F1 on both Eviction and Non-Eviction tasks. See Appendix B.2.

Fine-tuning details. We use LoRA with Unsloth for LLaMA/Qwen and standard HuggingFace training for BERT; token lengths, optimizers, and schedules are in Appendix A.3. We also give train/dev splits, seed, and prompt templates there.

5 Results

We report empirical results for our method, organized around the figures and tables in the main text; full numbers and extended analyses are in Appendix C.

RQ1. Can models reliably detect any *eviction mention*? Table 1 (top) show that the binary task is easy for most models. GPT-4o with APO attains the best Micro-F1 across datasets (Synth 0.939, Mimic 0.938, PMC 0.947; average 0.941). GPT-4o without APO is much lower, so self prompt optimization helps. Fine-tuned open LLMs are also strong: Qwen2.5-7B-FT averages 0.922 and LLaMA-3.1-8B-FT averages 0.908. 3B models do well on Synth but drop on PMC, which contains longer notes. BERT encoders trained with our pipeline are very high on Synth (BioBERT 0.987) and competitive on average. See Appendix C.1.

RQ2. How well do models separate the seven *eviction* subtypes? This step is harder. Table 1 (middle) show two consistent effects. First, APO gives large gains for closed models: GPT-4o moves from Macro-F1 0.515 to 0.878 and from Micro-F1 0.613 to 0.886; GPT-4o-mini shows similar jumps. Second, fine-tuned open LLMs meet or pass GPT-4o-APO. Qwen2.5-7B-FT averages 0.888 Macro-F1 and 0.895 Micro-F1. LLaMA-3.1-8B-FT averages 0.888 and 0.897. 3B variants reach about 0.87 Micro-F1 on average, so they are practical under small compute. BioBERT and ClinicalBERT lag, especially on PMC. See Appendix C.2.

RQ3. How well do models classify the seven *non-eviction* SDoH subtypes? These labels are more common in public notes, and scores are higher for all models. Table 1 (bottom) show GPT-4o without APO is already strong (Macro-F1 0.797, Micro-F1 0.838), and APO lifts it to 0.873 and 0.889. Fine-tuned open LLMs are best or tied: Qwen2.5-7B-FT reaches 0.903 and 0.913 on average, and LLaMA-3.1-8B-FT reaches 0.883 and 0.915. 3B models outperform GPT-4o-mini-APO by a wide margin. BERT variants remain behind. See Appendix C.3.

RQ4. Which eviction subtypes are easy or hard? Table 5 reports class wise MCC and F1. *Eviction_absent* is trivial in Synth (all 1.00) and is removed from Mimic and PMC to avoid flooding negatives. *Hypothetical* is moderate and improves with fine tuning, for example Qwen2.5-7B-FT scores 0.98 to 1.00 on Synth and at least 0.92 on Mimic. *Mutual Rescission* is stable across models at about 0.90 or higher. *Pending* varies across corpora, and PMC is the noisiest. *Present_current* vs *Present_history* is the hardest pair because both rely on time cues; GPT-4o-APO is strong on the current case in PMC (about 0.93), while some open models drop. See Appendix C.4.

RQ5. Do reasoning traces in training help? Figure 3 compares labels only vs labels plus short rationales on the eviction multi class task. Small models gain the most. LLaMA-3.2-3B-FT improves from about 0.848 to about 0.886. Qwen2.5-3B-FT improves from about 0.855 to about 0.868. Large models change little. Reasoning thus acts as an explicit decision scaffold for compact LLMs. See Appendix C.5.

RQ6. How much data is enough for fine tuning? Figure 4 sweeps 200, 1k, 3k, 5k, and 10k training examples. Scores jump from 200 to 3k, and plateau by 5k. Some models slip at 10k, which suggests overfitting or domain drift. For Step 2, Qwen2.5-7B peaks near 5k (Macro-F1 about 0.912) and then drops at 10k. For Step 3, many models peak near 3k. A simple rule works in practice: use 3k to 5k examples for LLMs and use about 3k for BERT on Step 2; use about 3k for Step 3 for both. See Appendix C.6.

RQ7. Does mixing real long notes help transfer? Table 6 compares 100% synthetic with a 70% synthetic plus 30% PMC mix. The mixed set improves all three model families on Step 2 and Step 3. Gains are largest on PMC dev data, so long real notes improve cross domain robustness. See Appendix A.3.

Models	Synth	Mimic	PMC	Avg
Performance Results of Various Models for Step 1: Binary Classification (this is a balanced dataset, so we only report Micro-F1 here)				
GPT-4o-mini	0.761 (0.757-0.764) $p < 0.05$	0.735 (0.732-0.739) $p < 0.05$	0.889 (0.879-0.900) $p < 0.05$	0.795 (0.790-0.801) $p < 0.05$
GPT-4o-mini-APO	0.796 (0.781-0.811) $p < 0.05$	0.706 (0.653-0.759) $p < 0.05$	0.876 (0.860-0.892) $p < 0.05$	0.793 (0.770-0.816) $p < 0.05$
GPT-4o	0.880 (0.867-0.894) $p < 0.05$	0.860 (0.847-0.872) $p < 0.05$	0.839 (0.826-0.851) $p < 0.05$	0.859 (0.850-0.869) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.976 (0.961-0.991) $p < 0.01$	0.821 (0.780-0.861) $p < 0.05$	0.857 (0.836-0.878) $p < 0.05$	0.885 (0.870-0.899) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	**0.987 (0.979-0.995) $p < 0.01$**	0.916 (0.890-0.942)	0.892 (0.854-0.930) $p < 0.05$	0.931 (0.916-0.947)
SynthEHR-Eviction (Bio_ClinicalBERT)	0.978 (0.960-0.995) $p < 0.01$	0.756 (0.706-0.806) $p < 0.05$	0.805 (0.718-0.893) $p < 0.05$	0.846 (0.813-0.880) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	0.966 (0.956-0.976) $p < 0.01$	0.866 (0.850-0.882) $p < 0.05$	0.891 (0.875-0.907) $p < 0.05$	0.908 (0.897-0.919) $p < 0.05$
SynthEHR-Eviction (LLama-3.2-3B)	0.934 (0.929-0.938)	0.868 (0.858-0.877) $p < 0.05$	0.824 (0.813-0.834) $p < 0.05$	0.875 (0.871-0.879) $p < 0.05$
SynthEHR-Eviction (Qwen2.5-7B)	0.963 (0.955-0.981) $p < 0.01$	0.912 (0.904-0.920) $p < 0.05$	0.891 (0.882-0.900) $p < 0.05$	0.922 (0.918-0.926)
SynthEHR-Eviction (Qwen2.5-3B)	**0.987 (0.976-0.997) $p < 0.001$**	0.850 (0.840-0.859) $p < 0.05$	0.834 (0.820-0.848) $p < 0.05$	0.890 (0.884-0.896) $p < 0.05$
GPT-4o-APO (baseline)	0.939 (0.912-0.965)	**0.938 (0.926-0.949)**	**0.947 (0.936-0.958)**	**0.941 (0.931-0.951)**
Performance Results of Various Models for Step 2: Eviction Multi-Class (Macro-F1)				
GPT-4o-mini	0.337 (0.331-0.343) $p < 0.05$	0.238 (0.217-0.260) $p < 0.05$	0.243 (0.217-0.269) $p < 0.05$	0.273 (0.263-0.283) $p < 0.05$
GPT-4o-mini-APO	0.760 (0.696-0.823) $p < 0.05$	0.672 (0.652-0.692) $p < 0.05$	0.641 (0.547-0.735) $p < 0.05$	0.691 (0.640-0.742) $p < 0.05$
GPT-4o	0.670 (0.629-0.711) $p < 0.05$	0.521 (0.502-0.539) $p < 0.05$	0.352 (0.302-0.403) $p < 0.05$	0.515 (0.489-0.540) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.806 (0.790-0.822) $p < 0.05$	0.784 (0.734-0.834) $p < 0.05$	0.384 (0.248-0.519) $p < 0.05$	0.658 (0.604-0.712) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.811 (0.792-0.831) $p < 0.05$	0.668 (0.585-0.752) $p < 0.05$	0.342 (0.300-0.384) $p < 0.05$	0.607 (0.568-0.646) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.790 (0.762-0.818) $p < 0.05$	0.658 (0.630-0.686) $p < 0.05$	0.346 (0.321-0.371) $p < 0.05$	0.598 (0.584-0.611) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	0.881 (0.820-0.942)	0.935 (0.907-0.962) $p < 0.01$	**0.848 (0.821-0.876)**	**0.888 (0.872-0.904)**
SynthEHR-Eviction (LLama-3.2-3B)	0.909 (0.895-0.922)	0.812 (0.665-0.958)	0.801 (0.774-0.828)	0.840 (0.784-0.897)
SynthEHR-Eviction (Qwen2.5-7B)	**0.920 (0.903-0.938)**	**0.939 (0.933-0.944) $p < 0.001$**	0.804 (0.767-0.842)	**0.888 (0.869-0.907)**
SynthEHR-Eviction (Qwen2.5-3B)	0.889 (0.824-0.954)	0.850 (0.827-0.873)	0.825 (0.745-0.905)	0.855 (0.824-0.886)
GPT-4o-APO (baseline)	0.918 (0.912-0.925)	0.878 (0.846-0.911)	0.836 (0.750-0.923)	0.878 (0.842-0.914)
Performance Results of Various Models for Step 3: Non-Eviction Multi-Class (Macro-F1)				
GPT-4o-mini	0.830 (0.823-0.837) $p < 0.05$	0.672 (0.666-0.678) $p < 0.05$	0.645 (0.586-0.704) $p < 0.05$	0.716 (0.696-0.735) $p < 0.05$
GPT-4o-mini-APO	0.881 (0.875-0.887)	0.737 (0.716-0.757) $p < 0.05$	0.724 (0.695-0.754) $p < 0.05$	0.781 (0.771-0.791) $p < 0.05$
GPT-4o	0.850 (0.838-0.862) $p < 0.05$	0.795 (0.769-0.820) $p < 0.05$	0.746 (0.713-0.779) $p < 0.05$	0.797 (0.782-0.812) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.865 (0.858-0.873) $p < 0.05$	0.600 (0.486-0.714) $p < 0.05$	0.507 (0.303-0.711) $p < 0.05$	0.658 (0.581-0.735) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.863 (0.851-0.874) $p < 0.05$	0.623 (0.478-0.767) $p < 0.05$	0.564 (0.418-0.711) $p < 0.05$	0.683 (0.611-0.756) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.857 (0.847-0.867) $p < 0.05$	0.646 (0.601-0.692) $p < 0.05$	0.423 (0.395-0.451) $p < 0.05$	0.642 (0.631-0.654) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	**0.968 (0.962-0.973) $p < 0.001$**	0.847 (0.821-0.874)	0.833 (0.798-0.868)	0.883 (0.866-0.900)
SynthEHR-Eviction (LLama-3.2-3B)	0.914 (0.840-0.988)	0.642 (0.559-0.725) $p < 0.05$	0.782 (0.723-0.841) $p < 0.05$	0.780 (0.736-0.824) $p < 0.05$
SynthEHR-Eviction (Qwen2.5-7B)	0.965 (0.958-0.973) $p < 0.001$	**0.864 (0.852-0.877)**	**0.879 (0.835-0.923)**	**0.903 (0.888-0.918) $p < 0.05$**
SynthEHR-Eviction (Qwen2.5-3B)	0.950 (0.942-0.958) $p < 0.001$	0.727 (0.646-0.808) $p < 0.05$	0.785 (0.749-0.821) $p < 0.05$	0.821 (0.787-0.855) $p < 0.05$
GPT-4o-APO (baseline)	0.900 (0.877-0.923)	0.850 (0.837-0.863)	0.870 (0.827-0.913)	0.873 (0.852-0.894)

Table 1: Performance comparison of various models across all steps and datasets. Full table with Micro-F1 or Macro-F1 scores is in Appendix Table 4. In experiments, we report both Macro-F1 and Micro-F1 scores with 95% confidence intervals for Step 1 (Binary Classification), Step 2 (Eviction Multi-Class Classification), and Step 3 (Non-Eviction Multi-Class Classification). *SynthEHR-Eviction* refers to models fine-tuned using the SynthEHR-Eviction dataset created via our Augmenter + Annotator pipeline. *GPT-4o-APO* serves as the baseline for statistical comparisons in each subtask. Blue cells indicate scores that are significantly lower than the baseline, orange cells indicate scores that are not significantly different from the baseline (i.e., at a similar level as GPT-4o-APO), red cells indicate scores that are significantly higher than the baseline, and **bold underlined values** represent the best performance within each column (i.e., the highest Micro-F1 or Macro-F1 for that dataset/subtask).

6 Discussion

Eviction is a critical yet underrepresented SDoH, associated with cascading adverse outcomes such as homelessness, chronic stress, and emergency care overuse (Desmond and Kimbro, 2015; Desmond, 2016; Desmond and Gershenson, 2017; Tsai and Huang, 2019; Tsai, 2024; Yao et al., 2023b). While prior NLP efforts have focused on broader housing-related issues, such as homelessness or housing instability (Guevara et al., 2024; Patra et al., 2021), the status of eviction has been largely neglected in public clinical datasets and benchmarks. Only very recently has this issue received direct attention, for instance Yao et al. (2023) introduced an eviction classifier trained on a private Veterans Affairs dataset (Yao et al., 2023b). To help fill this gap, we present SynthEHR-Eviction: a scalable and clinically grounded data generation pipeline that produces high-fidelity, eviction-labeled clinical notes aligned with real-world documentation practices. The pipeline combines LLM-based data augmentation, automated prompt optimization, and human-in-the-loop validation to synthesize instruction-style notes paired with labels derived from ICD-10 Z59 codes (e.g., Z59.0 to Z59.9, covering housing and economic instability). This label schema enhances semantic precision and interoperability with structured health data systems (for Medicare & Medicaid Services , CMS; Guevara et al., 2024). Furthermore, our taxonomy goes beyond binary classification by capturing nuanced eviction subtypes, such as pending status, mutual rescission, and temporal distinctions (e.g., current vs. history), allowing for more accurate model training and more actionable risk detection. By releasing both the annotated dataset and labeling guidelines, we support the development and evaluation of eviction-aware NLP tools that generalize across institutions and reflect realistic documentation variability.

Efficiency, modularity, and CoT supervision. One of the most practical contributions of SynthEHR-Eviction lies in its efficiency: it dramatically reduces the human effort required to construct large-scale, high-quality annotated datasets. Manual annotation of complex SDoH categories, particularly those like eviction that involve temporal ambiguity, legal framing, and implicit cues, is notoriously labor-intensive. In our controlled study, full manual rewriting of 8,000 EHR notes took over 266 hours, and even direct annotation of labels consumed 33 hours. In contrast, our GPT-assisted, human-in-the-loop workflow, with targeted prompt engineering, batch validation, and automated reasoning trace generation, achieved comparable data quality with less than 6 hours of human effort, representing an over 80% reduction in annotation time. These findings align with recent work showing the scalability and reliability of weak supervision, prompt-based augmentation, and LLM-powered data generation in other domains (Hsu and Roberts, 2025; Zhang et al., 2022a; Zha et al., 2025; Wang et al., 2022). More importantly, our pipeline offers structured reasoning alongside each label, allowing smaller models to learn decision rationales explicitly, an advantage not typically found in weakly labeled or distant supervision setups. Because the pipeline is built on modular components (e.g., task-adaptive prompt optimization, reasoning-based CoT supervision, and human verification loops), it generalizes to other low-resource SDoH domains where fine-grained annotation is scarce but clinically significant (e.g., utility shutoff, food insecurity, or intimate partner violence). As clinical NLP shifts toward low-cost, explainable, and domain-adaptable solutions, SynthEHR-Eviction provides a blueprint for scaling up dataset construction without compromising semantic depth or real-world alignment (Guevara et al., 2024; Magister et al., 2022).

Generalization, domain shift, and model families. To systematically evaluate model generalization in eviction-related SDoH classification, we tested performance across three distinct clinical datasets—Synth, MIMIC, and PMC—that vary substantially in linguistic style, narrative density, and documentation patterns (Table 10). Our results reveal a consistent domain shift: models achieved their highest performance on the synthetic data (Synth), performed moderately on real-world EHRs (MIMIC), and worst on academic case reports (PMC). While all three datasets adopt instruction-style formatting, they differ in vocabulary, structure, and contextual framing of eviction signals. The Synth dataset, generated using our LLM-based pipeline, closely mirrors the model's generation style and served as the primary training source, explaining its strong in-domain performance. In contrast, MIMIC notes, written by clinicians, feature more heterogeneous phrasing (e.g., "forced to remove from the rented house") and exhibit subtle lexical shifts. The PMC notes pose the greatest challenge, as eviction information is often deeply embedded in long diagnostic narratives and may only appear incidentally in unrelated sections, requiring fine-grained relevance detection and temporal disambiguation (as shown in Table 10). This distributional mismatch explains the observed drop in generalization: for example, Qwen2.5-7B's Macro-F1 fell from 0.920 (Synth) to 0.804 (PMC). These findings are consistent with prior SDoH extraction studies documenting performance degradation when models are evaluated on free-text narratives (Keloth et al., 2025; Han et al., 2022; Mitra et al., 2023). Because SynthEHR-Eviction spans synthetic and real-world document styles, it enabled us to evaluate how different model families generalize across varied clinical documentation settings. On the challenging PMC multi-class classification task, a general-domain LLaMA-3.1-8B model achieved a Micro-F1 of 0.861, while the BERT-base-cased model fell to just 0.444, a gap over 40 points. This substantial difference underscores LLMs' ability to handle diverse and noisy clinical narratives, likely due to larger parameter counts, broader pretraining corpora, and emergent reasoning capabilities (Guevara et al., 2024). These findings are in line with benchmarks showing that state-of-the-art LLMs, such as GPT-4 and Flan-T5-XXL, outperform domain-specific biomedical models like BioBERT and ClinicalBERT across SDoH extraction and other fine-grained clinical NLP tasks (Majid et al., 2025; Guevara et al., 2024).

Temporal reasoning as a persistent challenge. Our fine-grained eviction label schema, distinguishing between categories such as `Eviction_present_current` and `Eviction_present_history`, allowed us to examine one of the most persistent challenges in clinical NLP: temporal reasoning errors arising from implicit event timelines in free-text notes.

Temporal ambiguity emerged as a consistent source of classification error during model evaluation. A closer inspection of eviction status classification errors (Cases 1 and 2, Table 13) reveals a critical challenge: the models' difficulty in distinguishing between historical and current eviction events due to implicit temporal ambiguities in clinical notes. In Case 1, the phrase "resides now" prompted the Llama-3.1-8B model to incorrectly classify a historical eviction (`Eviction_present_history`) as ongoing (`Eviction_present_current`). Conversely, in Case 2, the model incorrectly labeled a recent eviction ("two months ago") as historical, heavily influenced by the initial lexical cue "history," despite the explicit recent temporal context. These bidirectional error patterns emphasize temporal ambiguity, where historical events can have ongoing impacts or be conflated with current circumstances. Our observations align with known challenges in clinical NLP: reasoning about the timing of events in narratives is difficult for models to capture (Kougia et al., 2024). Even advanced LLMs have been reported to struggle with temporal relation extraction, performing worse than fine-tuned models and showing inconsistent handling of event order.

Who benefits from reasoning supervision. We also explored structured reasoning annotations, that is chain-of-thought explanations provided during training, to examine their utility in enhancing model performance and interpretability, particularly for smaller, resource-efficient models. Smaller models benefit disproportionately from this form of supervision: LLaMA-3.2-3B improved by +3.8 Macro-F1 and Qwen2.5-3B by +1.3, compared to their counterparts trained on labels alone. In contrast, larger models (Llama-3.1-8B, Qwen2.5-7B) showed minimal improvements under the same setup, suggesting that extensive pretraining had already endowed them with stronger reasoning abilities. These findings are consistent with prior observations on chain-of-thought prompting (Wei et al., 2022): smaller models often struggle to exhibit complex reasoning unless explicitly guided, while larger models benefit less. SynthEHR-Eviction operationalizes this insight by pairing each label with interpretable reasoning traces, allowing smaller models to learn not just what the correct answer is, but why it is correct. Qualitative analysis further supports this conclusion. In Case 3 (Table 13), the Llama-3.2-3B model trained with reasoning annotations correctly classified a complex eviction situation as `Eviction_pending`, clearly articulating the active and unresolved nature of the eviction process. In contrast, the same model trained without reasoning annotations misclassified the scenario as `Eviction_present_current`, failing to capture subtle temporal and procedural cues. This example illustrates how structured reasoning guidance can help smaller models develop more precise and interpretable decision boundaries in semantically challenging cases.

Limits of explanation faithfulness. While structured reasoning annotations improved transparency, our evaluation revealed limitations in reasoning fidelity, that is whether a model's explanation faithfully reflects its underlying decision process. Our case studies (Cases 4 to 6, Table 14) revealed interpretability challenges even when the final classification was correct. In several instances, the model produced an explanation that was partially incorrect or conceptually misaligned, despite outputting the right label. For example, the Llama-3.1-8B model correctly labeled a case as `Eviction_mr_current` but its explanation stated that the mutual lease rescission was unrelated to eviction, indicating a misunderstanding of the concept. In another case, the model correctly identified `Eviction_pending` but described it using language that suggests a historical context, which deviates from the annotation guidelines. Similarly, a note properly labeled as `Eviction_present_current` was explained as if the eviction process were still ongoing, failing to recognize it had been completed. These examples show that the presence of a reasoning chain does not guarantee fidelity and that plausible rationales may be inconsistent with the model's actual decision strategy.

7 Conclusion

SynthEHR-Eviction reframes eviction from a rare, underspecified signal into a tractable, fine-grained NLP task. A lightweight APO+DSPy+HITL loop yields a labeled corpus with interpretable reasoning at low cost, and small open models fine-tuned on this data match or exceed strong API baselines while remaining deployable. The approach extends to other underrepresented SDoH and supports transparent, workflow-ready outputs for equitable clinical NLP.

8 Limitations

Our study has several limitations. First, synthetic training notes, although grounded in EHR style and ICD-anchored labels, may not capture the full sociolinguistic variability across institutions, specialties, and populations; this can introduce distributional artifacts or over-reliance on template phrases. Second, our generation design emphasizes prompt-optimized augmentation and high-precision filtering. We did not exhaustively explore broader generator families or higher-recall strategies, which may shift trade-offs between coverage and noise. Third, temporal reasoning remains a bottleneck: models often confuse `present_current` vs. `present_history` and borderline `pending` cases. Future improvements may require explicit temporal tagging or leveraging note timestamps, potentially guided by time-aware training frameworks (Liu et al., 2025). Fourth, out-of-domain generalization still degrades on long, heterogeneous narratives such as PMC case reports. Although mixing 30% real notes improves transfer, further clinical validation is needed before deployment in routine care. Fifth, explanation faithfulness is imperfect. A correct label can co-occur with a partially incorrect rationale, which limits the use of CoT as a definitive audit trail. Sixth, performance gains tend to plateau around 3,000 to 5,000 training examples; larger sets can yield diminishing returns or mild overfitting when synthetic notes become repetitive. Finally, practical constraints matter. Inference latency, compute costs, and limited GPU infrastructure can hinder use in under-resourced health systems (Dennstädt et al., 2025). Hybrid strategies such as distillation, caching, dynamic inference control, or task-specific adapters may be necessary to strike a balance between accuracy and feasibility.

9 Ethical Considerations

Data governance and privacy. We follow access terms for MIMIC and PMC and keep credentialed raw text under original controls. Released materials contain synthetic notes, schema, and prompts without personally identifying information. API-backed annotation is configured without human review, consistent with credentialed-use guidance.

Intended use and risk. Models are for research and decision support. Outputs should not be used to deny care, housing, or benefits. Predictions should be reviewed by trained staff and paired with established social-service workflows.

Bias and fairness. SDoH content reflects structural inequities. Synthetic data and LLM priors may inherit or amplify bias. We recommend subgroup auditing where feasible, monitoring false positive and false negative rates for protected classes, and stress-testing temporal and legal variants.

Explainability and responsibility. Reasoning traces improve transparency but are not guaranteed to be faithful to internal computation. We recommend logging predictions and rationales, sampling for expert review, and documenting failure modes, especially temporal errors. Efforts to reduce hallucinations and improve factual alignment remain important (Kim et al., 2025; Ji et al., 2023; Mishra et al., 2024; Yao et al., 2023a). Collaboration with clinicians to refine reasoning templates and review outputs is essential (Yang et al., 2025; Jin et al., 2024). Any deployment should include governance for updates, rollback, appeals, and user feedback (Abgrall et al., 2024).

Broader social impact. Eviction detection can enable earlier referral to housing and social services, yet it also carries risks of profiling, stigma, and secondary use outside care delivery (for example by payers or landlords). To mitigate harm, we release only synthetic notes and code; we accompany releases with a data statement, a model card, and license terms that prohibit punitive or non-clinical uses. We recommend institutional IRB review for any prospective deployment, clinician-in-the-loop review for all flags, and auditing for disparate error rates. Deployments should frame outputs as offers of assistance, not as grounds for sanctions, and should provide clear appeal channels for affected patients and communities.

References

Gwénolé Abgrall, Andre L Holder, Zaineb Chelly Dagdia, Karine Zeitouni, and Xavier Monnet. 2024. Should ai models be explainable to clinicians? *Critical Care*, 28(1):301.

Barbara G Bokhour, Jolie N Haun, Justeen Hyde, Martin Charns, and Benjamin Kligler. 2020. Transforming the veterans affairs to a whole health system of care: time for action and research. *Medical Care*, 58(4):295–300.

Anusha Bompelli, Yanshan Wang, Ruyuan Wan, Esha Singh, Yuqi Zhou, Lin Xu, David Oniani, Bhavani Singh Agnikula Kshatriya, Joyce (Joy) E. Balls-Berry, and Rui Zhang. 2021. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. *Health Data Science*, 2021.

Katherine L Chen, Isomi M Miake-Lye, Meron M Begashaw, Frederick J Zimmerman, Jody Larkin, Emily L McGrath, and Paul G Shekelle. 2022. Association of promoting housing affordability and stability with improved health outcomes: a systematic review. *JAMA Network Open*, 5(11):e2239860–e2239860.

John Joon Young Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. *arXiv preprint arXiv:2306.04140*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *npj Digital Medicine*, 8(1):143.

Matthew Desmond. 2016. *Evicted: Poverty and Profit in the American City*. Crown, New York City, USA.

Matthew Desmond and Carl Gershenson. 2017. Who gets evicted? assessing individual, neighborhood, and network factors. *Social Science Research*, 62:362–377.

Matthew Desmond and Rachel Tolbert Kimbro. 2015. Eviction's fallout: housing, hardship, and health. *Social Forces*, 94(1):295–324.

Yihong Dong, Kangcheng Luo, Xue Jiang, Zhi Jin, and Ge Li. 2023. Pace: Improving prompt with actor-critic editing for large language model. *arXiv preprint arXiv:2308.10088*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Centers for Medicare & Medicaid Services (CMS). 2023. Improving the collection of social determinants of health (sdoh) data with icd-10-cm z codes. Retrieved from https://go.cms.gov/OMH.

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.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2023. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *arXiv preprint arXiv:2309.08532*.

Sifei Han, Robert F Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of Biomedical Informatics*, 127:103984.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Peter Hepburn, Renee Louis, and Matthew Desmond. 2020. Racial and gender disparities among evicted americans. *Sociological Science*, 7:649–662.

Carlyn M Hood, Keith P Gennuso, Geoffrey R Swain, and Bridget B Catlin. 2016. County health rankings: relationships between determinant factors and health outcomes. *American journal of preventive medicine*, 50(2):129–135.

Enshuo Hsu and Kirk Roberts. 2025. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific reports*, 15(1):8241.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, and 1 others. 2024. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *npj Digital Medicine*, 7(1):190.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, pages 49–55.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Vipina K Keloth, Salih Selek, Qingyu Chen, Christopher Gilman, Sunyang Fu, Yifang Dang, Xinghan Chen, Xinyue Hu, Yujia Zhou, Huan He, and 1 others. 2025. Social determinants of health extraction from clinical notes across institutions using large language models. *npj Digital Medicine*, 8(1):1–13.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, and 1 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Vasiliki Kougia, Anastasiia Sedova, Andreas Stephan, Klim Zaporojets, and Benjamin Roth. 2024. Analysing zero-shot temporal relation extraction on clinical notes using temporal consistency. *arXiv preprint arXiv:2406.11486*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, and Michael Bendersky. 2023a. Automatic prompt rewriting for personalized text generation. *arXiv preprint arXiv:2310.00152*.

Rumeng Li, Xun Wang, and Hong Yu. 2023b. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 7129. NIH Public Access.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zijia Liu, Peixuan Han, Haofei Yu, Haoru Li, and Jiaxuan You. 2025. Time-r1: Towards comprehensive temporal reasoning in llms. *arXiv preprint arXiv:2505.13508*.

Lisa L Losada-Rojas, Dimitra Pyrialakou, Brigitte S Waldorf, Jorge A Banda, and Konstantina Gkritza. 2024. Understanding transportation-related health determinants and their interrelationships in rural us. *Journal of Transport & Health*, 38:101887.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Iyad Majid, Vaibhav Mishra, Rohith Ravindranath, and Sophia Y Wang. 2025. Evaluating the performance of large language models for named entity recognition in ophthalmology clinical free-text notes. In *AMIA Annual Symposium Proceedings*, volume 2024, page 778.

Prakamya Mishra, Zonghai Yao, Parth Vashisht, Beining Wang, Vidhi Mody, Hong Yu, and 1 others. 2024. Synfac-edit: Synthetic imitation edit feedback for factual alignment in clinical summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20061–20083.

Avijit Mitra, Emily Druhl, Raelene Goodwin, and Hong Yu. 2024. Synth-sbdh: A synthetic dataset of social and behavioral determinants of health for clinical text. *arXiv preprint arXiv:2406.06056*.

Avijit Mitra, Richeek Pradhan, Rachel D Melamed, Kun Chen, David C Hoaglin, Katherine L Tucker, Joel I Reisman, Zhichao Yang, Weisong Liu, and et al. Tsai, Jack. 2023. Associations between natural language processing–enriched social determinants of health and suicide death among us veterans. *JAMA Network Open*, 6(3):e233079–e233079.

National Center for Complementary and Integrative Health. n.d. Whole person health: What it is and why it's important. https://www.nccih.nih.gov/health/whole-person-health-what-it-is-and-why-its-imp Accessed: 2025-03-17.

World Health Organization. n.d. Social determinants of health. http://www.who.int/social_determinants/sdh_definition/en/. Accessed: 2025-03-17.

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, and 1 others. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12):2716–2727.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models. *arXiv preprint arXiv:2203.07281*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Kiran Ramnath, Kang Zhou, Sheng Guan, Soumya Smruti Mishra, Xuan Qi, Zhengyuan Shen, Shuai Wang, Sangmin Woo, Sullam Jeoung, Yawei Wang, and

1 others. 2025. A systematic survey of automatic prompt optimization techniques. *arXiv preprint arXiv:2502.16923*.

Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Hong Sun, Xue Li, Yinchuan Xu, Youkow Homma, Qi Cao, Min Wu, Jian Jiao, and Denis Charles. 2023. Autohint: Automatic prompt optimization with hint generation. *arXiv preprint arXiv:2307.07415*.

Jack Tsai. 2024. Theorizing pathways between eviction filings and increased mortality risk. *JAMA*, 331(7):570–571.

Jack Tsai and Melissa Huang. 2019. Systematic review of psychosocial factors associated with evictions. *Health and Social Care in the Community*, 27(3):e1–e9.

Jack Tsai, Natalie Jones, Dorota Szymkowiak, and Robert A Rosenheck. 2021. Longitudinal study of the housing and mental health outcomes of tenants appearing in eviction court. *Social psychiatry and psychiatric epidemiology*, 56:1679–1686.

Jack Tsai, Dorota Szymkowiak, Hind Beydoun, and Leandro Dasilva. 2024. Transportation as a determinant of health: Evaluating the veterans affairs rideshare program for homeless patients: Article examines the veterans affairs rideshare program for homeless patients. *Health Affairs*, 43(8):1090–1099.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. Promda: Prompt-based data augmentation for low-resource nlu tasks. *arXiv preprint arXiv:2202.12499*.

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.

Wikipedia contributors. n.d. Eviction in the united states. https://en.wikipedia.org/wiki/Eviction_in_the_United_States. Accessed: 2025-03-17.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Gps: Genetic prompt search for efficient few-shot learning. *arXiv preprint arXiv:2210.17041*.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023a. Knowledge-infused prompting improves clinical text generation with large language models. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Weijia Xu, Andrzej Banburski-Fahey, and Nebojsa Jojic. 2023b. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. *arXiv preprint arXiv:2305.09993*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, David McManus, Dan Berlowitz, and Hong Yu. 2025. Unveiling gpt-4v's hidden challenges behind high accuracy on usmle questions: Observational study. *Journal of Medical Internet Research*, 27:e65146.

Zonghai Yao, Ahmed Jaafar, Beining Wang, Zhichao Yang, and Hong Yu. 2024. Do clinicians know how to prompt? the need for automatic prompt optimization help in clinical note generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 182–201.

Zonghai Yao, Benjamin Schloss, and Sai Selvaraj. 2023a. Improving summarization with human edits. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2604–2620.

Zonghai Yao, Jack Tsai, Weisong Liu, David A. Levy, Emily Druhl, Joel I. Reisman, and Hong Yu. 2023b. Automated identification of eviction status from electronic health record notes. *Journal of the American Medical Informatics Association*, 30(8):1429–1437.

Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, and 1 others. 2023. Ehr-safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine*, 6(1):141.

Zehao Yu, Xi Yang, Yi Guo, Jiang Bian, and Yonghui Wu. 2022. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Frontiers in Public Health*, 10:778463.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. 2022a. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. 2022b. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific data*, 10(1):909.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Appendix Method Details

As illustrated in Figure 1, our approach focuses on developing a highly effective and reusable pipeline for multi-class annotation, which can be adapted to other SDoH annotation tasks. In this study, eviction serves as the primary experimental focus, showcasing the pipeline's ability to generate high-quality annotations across multiple eviction-related categories. The system is divided into three primary components: a training process for DSPy, data augmentation pipeline, and annotation system. The training process utilizes a small-scale human-validated dataset to optimize the DSPy program; the data augmentation pipeline generates labeled samples from raw clinical records; and the annotation system applies the trained models to classify augmented data and provide reasoning.

A.1 Data Augmentation Pipeline

A.1.1 Raw Note Extraction

We selected over 30,000 discharge notes from the MIMIC-III and MIMIC-IV databases. The selection criteria focused on notes that contained a "social history" section. This was based on our observation that most descriptions of patients' social circumstances—such as housing, financial status, and other social determinants—were concentrated within this part of the discharge notes. We then extracted the "social history" section from each note, using it as the raw note input for further augmentation.

A.1.2 Data Construction

Based on SDoH data with ICD-10-CM Z Codes (for Medicare & Medicaid Services , CMS), we categorize eviction as a new SDoH, classified under Z59.89, which addresses problems related to housing and economic circumstances. This classification positions eviction alongside other related classes under the Z59 category, as illustrated in Figure 2.

We classify eviction-related categories into 7 specific ones based on two key factors: presence (Absent, Hypothetical, Present, Pending, Mutual Rescission (MR)) and period (Current, History, Future)(Yao et al., 2023b):

- Eviction Absent: No history or indication of eviction.
- Eviction Present History, Eviction Present Current: Eviction has already been completed. "Current" refers to ongoing eviction impacts within the current natural year, not related to eviction. "History" refers to past without a specific time or that occurred more than one year ago.
- Eviction Hypothetical: Refers to possible eviction based on hypothetical scenarios.
- Eviction Pending: Eviction has been noticed but is not yet completed.
- Eviction MR History, Eviction MR Current: MR is a legal agreement between the landlord and tenant to terminate a lease.

In addition to eviction-related categories, we included tier 1 categories: Homelessness (Z59.0), Inadequate housing (Z59.1) and Lack of adequate food (Z59.4). We also incorporated tier 2 categories including Housing instability (Z59.81), Transportation insecurity (Z59.82), Financial insecurity (Z59.86), and Material hardship (Z59.87).

In total, we defined 14 classes. For each class, experts provided precise definitions and a few-shot examples to guide data augmentation. The detailed definitions and examples for each category can be found in the Appendix Table 7 and 8.

A.1.3 Augmenter Construction Process

As shown in Algorithm 11, we systematically developed an augmenter for each SDoH class. An augmenter in our context is essentially an LLM enhanced with an optimized prompt. Once the prompt is fully optimized, the augmenter takes raw notes as input and generates augmented notes that accurately reflect the target SDoH class.

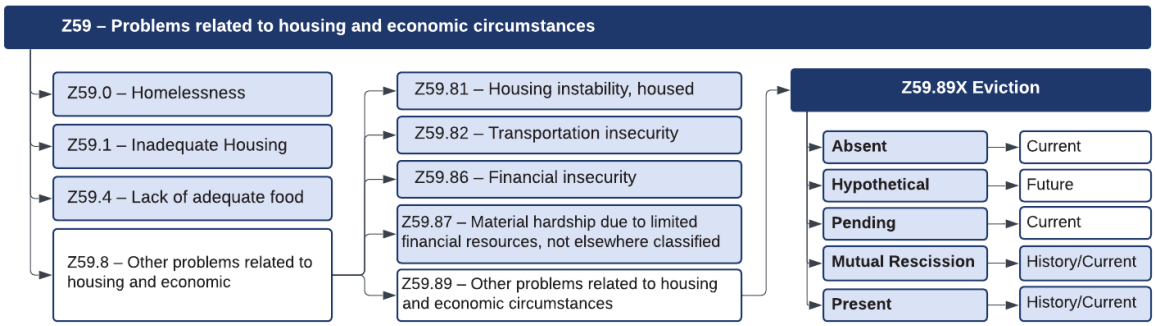

Figure 2: Relationships between Eviction and other housing and economic challenges, providing a comprehensive framework for understanding its broader implications within the SDoH context.

For every SDoH class, the process starts by defining the SDoH category and providing raw notes to guide the LLM in generating relevant augmented notes.

HITP Evaluation: The output generated by the augmenter in each iteration undergoes human verification. Human experts verify if each output note accurately reflects the target SDoH. Verified notes are added to the augmented dataset, while incorrect outputs are logged with feedback.

Assessment of Performance: Accuracy is calculated as the proportion of correct annotations out of total outputs. If accuracy exceeds 90%, the augmenter is deemed successful for that class. For augmenters not meeting the accuracy threshold, incorrect cases and feedback are returned to the LLM. The feedback is used to refine the prompt. The updated prompt is re-evaluated, and the cycle continues until the accuracy criterion is satisfied.

We conducted three optimization rounds to finalize augmenters for all 14 SDoH classes. For classes with clear distinctions and well-defined criteria (e.g., Tier 1 and Tier 2 SDoH categories), most augmenters reached the desired accuracy within a single iteration. For eviction-related classes, which involved finer granularity and ambiguous definitions, multiple iterations were required. This iterative process significantly improved the accuracy and reliability of the augmenters.

Each finalized augmenter demonstrated the capability to consistently generate augmented notes reflecting the target SDoH class. The optimized prompts for all classes are available in the Appendix.

A.2 DSPy Annotator Setting and Training Process

Inspired by the systematic survey of APO techniques (Ramnath et al., 2025), the practical implementation of APO via DSPy (Khattab et al., 2023), and our prior work demonstrating structured prompt optimization in healthcare settings (Yao et al., 2024), we design our annotator module as a critical component to ensure precise and efficient classification of notes into multiple SDoH categories.

A.2.1 Annotator Setup and Design

The dataset contains overlapping categories, particularly between Tier 1, Tier 2, and eviction-related classes. For instance, eviction can lead to homelessness or financial insecurity can lead to eviction, causing ambiguity in multi-class annotation. To address these overlaps, we first used a binary classification to determine whether content is related to eviction, and then divided the dataset into two groups, corresponding to three separate annotators, as outlined in Algorithm 11. The second annotator handles notes related to eviction-specific categories, including the following 7 labels: Eviction Absent, Eviction Present History, Eviction Present Current, Eviction Hypothetical, Eviction Pending, Eviction MR History, Eviction MR Current. The third annotator manages notes that do not directly pertain to eviction, covering the remaining 7 labels: Homelessness, Inadequate Housing, Lack of Adequate Food, Housing Instability, Transportation Insecurity, Financial Insecurity, and Material Hardship. This division ensures better class separation and significantly improves annotation accuracy.

A.2.2 DSPy Training Workflows

The training process (upper section of Figure 1) is founded on a small-scale human-validated dataset, annotated by domain experts and divided into training and development sets. Our DSPy program employs either GPT-4 or GPT-4o-mini as the underlying language models, implementing some key components shown as Table 2:

The output of the training process is an optimized DSPy model capable of accurately identifying potential eviction situations and providing corresponding reasoning. Any notes are input into this annotation tool to produce structured annotation results (annotation and detail reasoning), ensuring that the output dataset is high-quality, well-structured, and ready for downstream analysis.

A.2.3 Trainset and Devset

For each DSPy program, the trainset consists of data generated by the augmenter, specifically GPT-rewritten notes that have been verified by human evaluators. The trainset is carefully balanced, with 8 examples for each label category to ensure equal representation across all classification targets.

For the devset, we designed a more comprehensive evaluation to ensure the model's robustness and performance, with each label category represented by 48 examples. The devset includes three groups of data:

- Synth clinical notes: Similar to the trainset, this group consists of notes generated by GPT-4o-mini and verified by human experts, used to test the model's ability to generalize to data created by the augmenter.

- MIMIC-IV (Johnson et al., 2020) clinical notes: To assess the model's performance on real-world data, we included raw notes from the MIMIC dataset. These notes primarily cover tier 1 and tier 2 SDoH categories but do not initially contain any references to eviction. To introduce eviction-related content, we first used a set of keywords (listed in Table 12) to filter the raw notes. For cases where this filtering process did not yield sufficient results, we employed human experts to rewrite a subset of these raw notes. The rewriting process was done carefully, ensuring that the original structure and characteristics of the notes were preserved while selectively adding eviction information.

- PMC-Patient (Zhao et al., 2023) case reports: The third group comprises raw notes from PMC articles, following a similar strategy as the MIMIC raw notes. These notes were rewritten by humans to include targeted eviction-related scenarios without altering the inherent structure and characteristics of the original text.

By testing the model across these three distinct groups, we ensure that it can effectively handle both augmented and real-world data, as well as notes that have been selectively modified to introduce eviction scenarios. This diverse evaluation helps validate the model's performance and generalizability across various SDoH contexts.

A.3 Leveraging Augmenters and Annotators for Fine-Tuning Open-Source Models

Our fine-tuning process leverages the augmenter-annotator pipeline, which is instrumental in generating large volumes of high-quality labeled data. Label-specific augmenters were used to generate augmented notes tailored to the target labels. Each augmenter outputs synthetic data related to a specific SDoH category. The output from the augmenters was further annotated by the multi-label annotator. This annotator assigns the correct label along with corresponding reasoning.

In our fine-tuning process, we iteratively improved the model's performance through two phases. Initially, we exclusively used GPT-augmented sentences, which performed well on shorter texts but showed limited generalization on PMC long notes, highlighting the need for more diverse data. In the second phase, we enriched the training set by adding 30% PMC long notes, significantly enhancing the model's ability to handle complex inputs.

For the selection of open-source models, we conducted extensive experiments with multiple model architectures. We fine-tuned various models including BERT-based models (bert_base_cased-FT, biobert-v1.1-FT), domain-specific models (Bio_ClinicalBERT-FT), and their SBDH-adapted variants (sdh_bert_base_cased-FT, sdh_biobert-v1.1-FT, etc.). We also explored LLMs such as LLaMA-3.1-8B-FT, LLaMA-3.2-3B-FT, Qwen2.5-7B-FT, and Qwen2.5-3B-FT, along with their SBDH-adapted versions. While all models demonstrated improvements after fine-tuning, the larger models (particularly the 7B and 8B variants)

showed superior capacity for generalization and performance compared to the 3B-level models, which had more limited gains. This comprehensive approach allowed us to identify the most effective model architectures for our specific task.

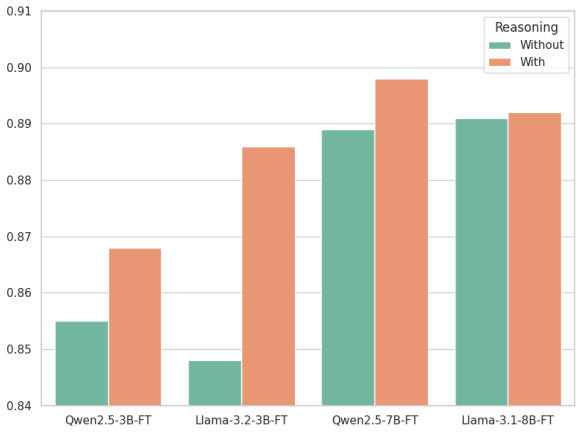

Figure 3: Performance comparison of LLMs trained with and without reasoning annotations on the Eviction Multi-Class task.

Training Set Composition Our initial training approach utilized purely synthetic data for model training. However, this configuration demonstrated suboptimal performance on PMC note evaluations, revealing limitations in the synthetic data's ability to capture the nuances of real-world medical documentation. To address this limitation, we modified our training set composition by replacing 30% of synthetic data with actual PMC notes. This hybrid approach yielded substantial improvements across different experimental settings. As shown in Table 6, the introduction of PMC data led to an increase of 0.237 in Micro-F1 score on the PMC development set in our DSPy-based experiments, while achieving an overall average improvement of 0.08 across all evaluation metrics. Similarly, in the LLM fine-tuning experiments, such as with Qwen2.5-7B, the hybrid training set demonstrated even more pronounced benefits, with performance on PMC-specific tasks improving by 0.302 while overall performance increased by 0.112.

Fine-tuneing Settings Our fine-tuning methodology was tailored to each model architecture to optimize performance while maintaining computational efficiency. For large language models (LLaMA and Qwen variants), we employed Unsloth *, a parameter-efficient fine-tuning framework that enables significantly faster training and inference compared to conventional methods.

For LLaMA and Qwen models, we implemented the following configuration using Unsloth:

- Maximum sequence length of 1024 tokens with 4-bit quantization (load_in_4bit=True) to reduce memory usage
- LoRA (Low-Rank Adaptation) with rank (r) of 16 and alpha value of 16 Target modules included query, key, value projections ('q_proj', 'k_proj', 'v_proj', 'o_proj', 'gate_proj', 'up_proj', 'down_proj')
- LoRA dropout set to 0 and bias configuration set to 'none' for optimized performance
- Learning rate of 2e-4 with a linear scheduler
- Training for 2 epochs with gradient accumulation steps of 4
- Warmup steps of 5 and weight decay of 0.01
- Per-device batch size of 2 with AdamW 8-bit optimizer

Our training dataset was split into training and test sets with an 80/20 ratio (test_size=0.2), using a seed of 42 for reproducibility. For data formatting, we employed a chat template structure with a system prompt instructing the model to act as a healthcare annotator specifically focused on eviction-related content. This approach helped align the model's responses with our annotation requirements.

For BERT-based models (bert_base_cased, biobert, etc.), we utilized the standard fine-tuning approach from Hugging Face's Transformers library. These models were trained with the same learning rate (2e-4) and epochs (2) as the LLMs for consistency, but with appropriate modifications to accommodate their architectural differences. The fine-tuning was performed on a Tesla T4 GPU with 14.748 GB of memory.

The hyperparameters for each model were determined through preliminary experiments, with optimal configurations selected based on Micro-F1 scores for our target SBDH categories. This comprehensive approach allowed us to identify the most effective model architectures while maintaining computational efficiency through parameter-efficient fine-tuning techniques.

*https://github.com/unslothai/unsloth

Component	Description
Signature	Defines input-output behavior for DSPy modules. Input: Patient Social History Note, Output: SDoH Annotation.
Modules	Each built-in module abstracts a prompting technique to handle signature. Use **Chain-of-Thought (CoT) module** being step-by-step inference for accurate SDoH annotations.
Evaluation Metric	Uses **Answer-Exact-Match** to evaluate output correctness. Ensures that each predicted label aligns precisely with the ground truth.
Optimizer	A DSPy optimizer is an algorithm that can tune the parameters of a DSPy program (i.e., prompts and/or LM weights) to maximize the metrics. We use **BootstrapFewShotWithRandomSearch**, which is building datasets for our modules and using them to finetune the LM weights in our system to iteratively select and evaluate example sets to optimize prompts.

Table 2: Overview of DSPy Design in Annotation Pipeline

Label	DSPy-train (Synth)	DSPy-eval (Synth)	SFT-train+eval	Test-total	Test-Synth	Test-Mimic	Test-PMC
Eviction_absent	8	12	500	20	20		
Eviction_hypothetical	8	12	750	48	20	20	8
Eviction_mr_current	8	12	750	48	20	20	8
Eviction_mr_history	8	12	750	48	20	20	8
Eviction_pending	8	12	750	48	20	20	8
Eviction_present_current	8	12	750	48	20	20	8
Eviction_present_history	8	12	750	48	20	20	8
Eviction_total	56	84	5000	308	140	120	48
Homelessness	8	12	450	20	20		
InadequateHousing	8	12	450	48	20	20	8
LackOfAdequateFood	8	12	450	48	20	20	8
FinancialInsecurity	8	12	450	48	20	20	8
HousingInstability	8	12	450	48	20	20	8
MaterialHardship	8	12	450	48	20	20	8
TransportationInsecurity	8	12	300	48	20	20	8
Non-Eviction_total	56	84	3000	308	140	120	48

Table 3: Data statistics across training and evaluation splits for APO (for close-source models) and SFT (for open-source models).

B Experimental Settings

B.1 Main Experimental Setup

We conducted a comprehensive evaluation across three sequential steps (Step 1: Binary Classification, Step 2: Eviction Multi-Class Classification, and Step 3: Non-Eviction Multi-Class Classification). For each model, we performed 5 independent runs to calculate 95% confidence intervals, ensuring the statistical reliability of our results.

For language models like GPT-4o and GPT-4o-mini, we initially used temperature$= 0$ for the annotation task to minimize uncertainty. However, to compute reliable confidence intervals, we performed 4 additional runs with temperature$= 0.5$. For models utilizing dspy frameworks and other LM/LLM architectures, we maintained consistency by using different random seeds across the 5 runs. Throughout all experiments, we kept the fine-tuning process consistent, with identical fine-tuning parameters across all models and training set sizes. To ensure compliance with MIMIC-IV data usage requirements [†], we exclusively used the Azure OpenAI service for GPT-4o and GPT-4o-mini, with human review explicitly disabled, following PhysioNet's guidelines for responsible use of credentialed data. We introduce the detailed experimental settings in the appendix.

Each model was evaluated on three distinct development sets (Synth, Mimic, and PMC), with the average performance calculated by simply taking the arithmetic mean across these three datasets. This approach allows us to assess model generalization across different data distributions.

The overall performance metric was derived from the best-performing model configuration for each step. Importantly, we consider a prediction correct only when Step 1 (Binary Classification) is correct, and either Step 2 (Eviction Multi-Class Classification) or Step 3 (Non-Eviction Multi-Class

[†] https://physionet.org/news/post/gpt-responsible-use

Models	Synth	Mimic	PMC	Avg
Performance Results of Various Models for Step 1: Binary Classification (this is a balanced dataset, so we only report Micro-F1 here)				
GPT-4o-mini	0.761 (0.757-0.764) $p < 0.05$	0.735 (0.732-0.739) $p < 0.05$	0.889 (0.879-0.900) $p < 0.05$	0.795 (0.790-0.801) $p < 0.05$
GPT-4o-mini-APO	0.796 (0.781-0.811) $p < 0.05$	0.706 (0.653-0.759) $p < 0.05$	0.876 (0.860-0.892) $p < 0.05$	0.793 (0.770-0.816) $p < 0.05$
GPT-4o	0.880 (0.867-0.894) $p < 0.05$	0.860 (0.847-0.872) $p < 0.05$	0.839 (0.826-0.851) $p < 0.05$	0.859 (0.850-0.869) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.976 (0.961-0.991) $p < 0.01$	0.821 (0.780-0.861) $p < 0.05$	0.857 (0.836-0.878) $p < 0.05$	0.885 (0.870-0.899) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	**0.987 (0.979-0.995) $p < 0.01$**	0.916 (0.890-0.942)	0.892 (0.854-0.930) $p < 0.05$	0.931 (0.916-0.947)
SynthEHR-Eviction (Bio_ClinicalBERT)	0.978 (0.960-0.995) $p < 0.01$	0.756 (0.706-0.806) $p < 0.05$	0.805 (0.718-0.893) $p < 0.05$	0.846 (0.813-0.880) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	0.966 (0.956-0.976) $p < 0.01$	0.866 (0.850-0.882) $p < 0.05$	0.891 (0.875-0.907) $p < 0.05$	0.908 (0.897-0.919) $p < 0.05$
SynthEHR-Eviction (LLama-3.2-3B)	0.934 (0.929-0.938)	0.868 (0.858-0.877) $p < 0.05$	0.824 (0.813-0.834) $p < 0.05$	0.875 (0.871-0.879) $p < 0.05$
SynthEHR-Eviction (Qwen2.5-7B)	0.963 (0.955-0.981) $p < 0.01$	0.912 (0.904-0.920) $p < 0.05$	0.891 (0.882-0.900) $p < 0.05$	0.922 (0.918-0.926)
SynthEHR-Eviction (Qwen2.5-3B)	**0.987 (0.976-0.997) $p < 0.001$**	0.850 (0.840-0.859) $p < 0.05$	0.834 (0.820-0.848) $p < 0.05$	0.890 (0.884-0.896) $p < 0.05$
GPT-4o-APO (baseline)	0.939 (0.912-0.965)	**0.938 (0.926-0.949)**	**0.947 (0.936-0.958)**	**0.941 (0.931-0.951)**
Performance Results of Various Models for Step 2: Eviction Multi-Class (Macro-F1 shown above, Micro-F1 shown below)				
GPT-4o-mini	0.337 (0.331-0.343) $p < 0.05$	0.238 (0.217-0.260) $p < 0.05$	0.243 (0.217-0.269) $p < 0.05$	0.273 (0.263-0.283) $p < 0.05$
GPT-4o-mini-APO	0.760 (0.696-0.823) $p < 0.05$	0.672 (0.652-0.692) $p < 0.05$	0.641 (0.547-0.735) $p < 0.05$	0.691 (0.640-0.742) $p < 0.05$
GPT-4o	0.670 (0.629-0.711) $p < 0.05$	0.521 (0.502-0.539) $p < 0.05$	0.352 (0.302-0.403) $p < 0.05$	0.515 (0.489-0.540) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.806 (0.790-0.822) $p < 0.05$	0.784 (0.734-0.834) $p < 0.05$	0.384 (0.248-0.519) $p < 0.05$	0.658 (0.604-0.712) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.811 (0.792-0.831) $p < 0.05$	0.668 (0.585-0.752) $p < 0.05$	0.342 (0.300-0.384) $p < 0.05$	0.607 (0.568-0.646) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.790 (0.762-0.818) $p < 0.05$	0.658 (0.630-0.686) $p < 0.05$	0.346 (0.321-0.371) $p < 0.05$	0.598 (0.584-0.611) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	0.881 (0.820-0.942)	0.935 (0.907-0.962) $p < 0.01$	**0.848 (0.821-0.876)**	**0.888 (0.872-0.904)**
SynthEHR-Eviction (LLama-3.2-3B)	0.909 (0.895-0.922)	0.812 (0.665-0.958)	0.801 (0.774-0.828)	0.840 (0.784-0.897)
SynthEHR-Eviction (Qwen2.5-7B)	**0.920 (0.903-0.938)**	**0.939 (0.933-0.944) $p < 0.001$**	0.804 (0.767-0.842)	**0.888 (0.869-0.907)**
SynthEHR-Eviction (Qwen2.5-3B)	0.889 (0.824-0.954)	0.850 (0.827-0.873)	0.825 (0.745-0.905)	0.855 (0.824-0.886)
GPT-4o-APO (baseline)	0.918 (0.912-0.925)	0.878 (0.846-0.911)	0.836 (0.750-0.923)	0.878 (0.842-0.914)
GPT-4o-mini	0.504 (0.496-0.512) $p < 0.05$	0.357 (0.331-0.383) $p < 0.05$	0.438 (0.410-0.465) $p < 0.05$	0.433 (0.421-0.444) $p < 0.05$
GPT-4o-mini-APO	0.840 (0.791-0.889) $p < 0.05$	0.772 (0.752-0.792) $p < 0.05$	0.681 (0.609-0.754) $p < 0.05$	0.764 (0.722-0.807) $p < 0.05$
GPT-4o	0.754 (0.741-0.767) $p < 0.05$	0.583 (0.567-0.600) $p < 0.05$	0.502 (0.459-0.545) $p < 0.05$	0.613 (0.595-0.631) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.810 (0.795-0.825) $p < 0.05$	0.880 (0.842-0.918)	0.444 (0.285-0.603) $p < 0.05$	0.711 (0.647-0.775) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.829 (0.790-0.868) $p < 0.05$	0.745 (0.700-0.790) $p < 0.05$	0.434 (0.289-0.579) $p < 0.05$	0.669 (0.603-0.736) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.791 (0.765-0.818) $p < 0.05$	0.775 (0.730-0.820) $p < 0.05$	0.390 (0.343-0.437) $p < 0.05$	0.652 (0.623-0.682) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	0.906 (0.888-0.923)	0.925 (0.909-0.941) $p < 0.01$	**0.861 (0.834-0.888)**	**0.897 (0.882-0.913)**
SynthEHR-Eviction (LLama-3.2-3B)	0.908 (0.896-0.921)	0.882 (0.848-0.916)	0.824 (0.796-0.853)	0.871 (0.849-0.893)
SynthEHR-Eviction (Qwen2.5-7B)	**0.921(0.905-0.938)**	**0.938(0.932-0.945) $p < 0.001$**	0.824 (0.789-0.860)	0.895 (0.877-0.912)
SynthEHR-Eviction (Qwen2.5-3B)	0.911 (0.901-0.921)	0.853 (0.832-0.875) $p < 0.05$	0.857 (0.801-0.914)	0.874 (0.856-0.892)
GPT-4o-APO (baseline)	0.909(0.886-0.931)	0.883(0.856-0.911)	0.868(0.806-0.929)	0.886(0.858-0.915)
Performance Results of Various Models for Step 3: Non-Eviction Multi-Class (Macro-F1 shown above, Micro-F1 shown below)				
GPT-4o-mini	0.830 (0.823-0.837) $p < 0.05$	0.672 (0.666-0.678) $p < 0.05$	0.645 (0.586-0.704) $p < 0.05$	0.716 (0.696-0.735) $p < 0.05$
GPT-4o-mini-APO	0.881 (0.875-0.887)	0.737 (0.716-0.757) $p < 0.05$	0.724 (0.695-0.754) $p < 0.05$	0.781 (0.771-0.791) $p < 0.05$
GPT-4o	0.850 (0.838-0.862) $p < 0.05$	0.795 (0.769-0.820) $p < 0.05$	0.746 (0.713-0.779) $p < 0.05$	0.797 (0.782-0.812) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.865 (0.858-0.873) $p < 0.05$	0.600 (0.486-0.714) $p < 0.05$	0.507 (0.303-0.711) $p < 0.05$	0.658 (0.581-0.735) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.863 (0.851-0.874) $p < 0.05$	0.623 (0.478-0.767) $p < 0.05$	0.564 (0.418-0.711) $p < 0.05$	0.683 (0.611-0.756) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.857 (0.847-0.867) $p < 0.05$	0.646 (0.601-0.692) $p < 0.05$	0.423 (0.395-0.451) $p < 0.05$	0.642 (0.631-0.654) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	**0.968 (0.962-0.973) $p < 0.001$**	0.847 (0.821-0.874)	0.833 (0.798-0.868)	0.883 (0.866-0.900)
SynthEHR-Eviction (LLama-3.2-3B)	0.914 (0.840-0.988)	0.642 (0.559-0.725) $p < 0.05$	0.782 (0.723-0.841) $p < 0.05$	0.780 (0.736-0.824) $p < 0.05$
SynthEHR-Eviction (Qwen2.5-7B)	0.965 (0.958-0.973) $p < 0.001$	**0.864 (0.852-0.877)**	**0.879 (0.835-0.923)**	**0.903 (0.888-0.918) $p < 0.05$**
SynthEHR-Eviction (Qwen2.5-3B)	0.950 (0.942-0.958) $p < 0.001$	0.727 (0.646-0.808) $p < 0.05$	0.785 (0.749-0.821) $p < 0.05$	0.821 (0.787-0.855) $p < 0.05$
GPT-4o-APO (baseline)	0.900 (0.877-0.923)	0.850 (0.837-0.863)	0.870 (0.827-0.913)	0.873 (0.852-0.894)
GPT-4o-mini	0.836 (0.830-0.842) $p < 0.05$	0.703 (0.698-0.708) $p < 0.05$	0.685 (0.669-0.701) $p < 0.05$	0.741 (0.734-0.748) $p < 0.05$
GPT-4o-mini-APO	0.923 (0.904-0.942)	0.783 (0.725-0.840) $p < 0.05$	0.852 (0.806-0.898)	0.852 (0.827-0.878) $p < 0.05$
GPT-4o	0.858 (0.851-0.866) $p < 0.05$	0.851 (0.839-0.864) $p < 0.05$	0.804 (0.791-0.817) $p < 0.05$	0.838 (0.831-0.844) $p < 0.05$
SynthEHR-Eviction (bert_base_cased)	0.965 (0.958-0.973) $p < 0.01$	0.600 (0.486-0.714) $p < 0.05$	0.589 (0.415-0.763) $p < 0.05$	0.718 (0.621-0.815) $p < 0.05$
SynthEHR-Eviction (biobert-v1.1)	0.963 (0.951-0.974) $p < 0.01$	0.623 (0.478-0.767) $p < 0.05$	0.604 (0.443-0.765) $p < 0.05$	0.730 (0.627-0.833) $p < 0.05$
SynthEHR-Eviction (Bio_ClinicalBERT)	0.961 (0.956-0.966) $p < 0.01$	0.630 (0.612-0.648) $p < 0.05$	0.478 (0.459-0.497) $p < 0.05$	0.690 (0.683-0.696) $p < 0.05$
SynthEHR-Eviction (LLama-3.1-8B)	**0.968(0.963-0.973) $p < 0.01$**	0.850(0.820-0.880)	**0.926 (0.910-0.942) $p < 0.05$**	**0.915(0.901-0.929)**
SynthEHR-Eviction (LLama-3.2-3B)	0.939 (0.917-0.960)	0.817 (0.787-0.847) $p < 0.05$	0.874 (0.844-0.904)	0.876 (0.856-0.897)
SynthEHR-Eviction (Qwen2.5-7B)	0.965 (0.958-0.973) $p < 0.01$	0.866 (0.854-0.877)	0.907 (0.871-0.944)	0.913 (0.900-0.925) $p < 0.05$
SynthEHR-Eviction (Qwen2.5-3B)	0.951 (0.944-0.959) $p < 0.01$	0.800 (0.769-0.831) $p < 0.05$	0.837 (0.818-0.856) $p < 0.05$	0.863 (0.849-0.877) $p < 0.05$
GPT-4o-APO (baseline)	0.910(0.877-0.943)	**0.878(0.853-0.904)**	0.878(0.833-0.922)	0.889(0.867-0.911)

Table 4: Performance comparison of various models across all steps and datasets. We report both Macro-F1 and Micro-F1 scores with 95% confidence intervals for Step 1 (Binary Classification), Step 2 (Eviction Multi-Class Classification), and Step 3 (Non-Eviction Multi-Class Classification). *SynthEHR-Eviction* refers to models fine-tuned using the SynthEHR-Eviction dataset created via our Augmenter + Annotator pipeline. *GPT-4o-APO* serves as the baseline for statistical comparisons in each subtask. Blue cells indicate scores that are significantly lower than the baseline, orange cells indicate scores that are not significantly different from the baseline (i.e., at a similar level as GPT-4o-APO), red cells indicate scores that are significantly higher than the baseline, and **bold underlined values** represent the best performance within each column (i.e., the highest Micro-F1 or Macro-F1 for that dataset/subtask).

		Synth		Mimic		PMC		Avg	
Model	**Class**	**MCC**	**F1**	**MCC**	**F1**	**MCC**	**F1**	**MCC**	**F1**
GPT-4o-APO	Eviction_absent	1.00	1.00					1.00	1.00
	Eviction_hypothetical	1.00	1.00	0.75	0.75	0.83	0.84	0.86	0.86
	Eviction_mr_current	0.82	0.84	0.89	0.90	0.95	0.96	0.89	0.90
	Eviction_mr_history	0.81	0.81	0.88	0.89	0.89	0.89	0.86	0.86
	Eviction_pending	0.95	0.96	0.81	0.83	0.81	0.83	0.86	0.87
	Eviction_present_current	0.91	0.93	0.97	0.97	0.88	0.89	0.92	0.93
	Eviction_present_history	0.94	0.94	0.97	0.98	0.95	0.96	0.95	0.96
	Overall_(macro_f1 / micro_f1)	0.92	0.93 / 0.93	0.88	0.89 / 0.89	0.88	0.89 / 0.90	0.89	0.90 / 0.91
Llama-3.1-8B	Eviction_absent	1.00	1.00					1.00	1.00
	Eviction_hypothetical	0.97	0.98	0.94	0.95	0.91	0.92	0.94	0.95
	Eviction_mr_current	0.77	0.80	0.94	0.95	0.89	0.90	0.87	0.88
	Eviction_mr_history	0.87	0.89	0.94	0.95	0.76	0.80	0.86	0.88
	Eviction_pending	0.87	0.88	0.94	0.95	0.81	0.83	0.87	0.89
	Eviction_present_current	0.88	0.90	1.00	1.00	0.74	0.76	0.88	0.89
	Eviction_present_history	0.91	0.92	1.00	1.00	0.85	0.87	0.92	0.93
	Overall_(macro_f1 / micro_f1)	0.89	0.80 / 0.91	0.96	0.97 / 0.97	0.82	0.85 / 0.85	0.89	0.87 / 0.91
Llama-3.2-3B	Eviction_absent	1.00	1.00					1.00	1.00
	Eviction_hypothetical	1.00	1.00	0.94	0.95	0.94	0.95	0.96	0.97
	Eviction_mr_current	0.81	0.83	0.90	0.92	0.88	0.89	0.86	0.88
	Eviction_mr_history	0.87	0.89	0.92	0.93	0.79	0.82	0.86	0.88
	Eviction_pending	0.89	0.91	0.92	0.93	0.83	0.85	0.88	0.90
	Eviction_present_current	0.89	0.89	0.80	0.81	0.75	0.76	0.81	0.82
	Eviction_present_history	0.91	0.92	0.92	0.93	0.89	0.91	0.91	0.92
	Overall (macro_f1 / micro_f1)	0.91	0.92 / 0.92	0.90	0.91 / 0.92	0.84	0.86 / 0.87	0.89	0.90 / 0.90
Qwen2.5-7B	Eviction_absent	1.00	1.00					1.00	1.00
	Eviction_hypothetical	0.97	0.98	0.91	0.92	0.88	0.89	0.92	0.93
	Eviction_mr_current	0.87	0.88	0.94	0.95	0.94	0.95	0.91	0.93
	Eviction_mr_history	0.94	0.94	0.94	0.95	0.85	0.87	0.91	0.92
	Eviction_pending	0.89	0.91	0.87	0.89	0.77	0.80	0.85	0.87
	Eviction_present_current	0.91	0.93	0.94	0.94	0.71	0.73	0.85	0.87
	Eviction_present_history	0.91	0.92	1.00	1.00	0.88	0.90	0.93	0.94
	Overall (macro_f1 / micro_f1)	0.93	0.94 / 0.94	0.93	0.94 / 0.94	0.84	0.86 / 0.86	0.90	0.91 / 0.91
Qwen2.5-3B	Eviction_absent	1.00	1.00					1.00	1.00
	Eviction_hypothetical	0.91	0.93	0.81	0.82	0.88	0.89	0.87	0.88
	Eviction_mr_current	0.84	0.86	0.87	0.89	0.92	0.93	0.88	0.89
	Eviction_mr_history	0.87	0.88	0.88	0.90	0.85	0.86	0.87	0.88
	Eviction_pending	0.87	0.89	0.79	0.82	0.77	0.80	0.81	0.84
	Eviction_present_current	0.91	0.93	0.84	0.85	0.78	0.79	0.84	0.85
	Eviction_present_history	0.91	0.92	0.95	0.95	0.92	0.93	0.92	0.93
	Overall (macro_f1 / micro_f1)	0.90	0.91 / 0.91	0.86	0.87 / 0.88	0.85	0.87 / 0.87	0.87	0.88 /0.89

Table 5: Model performance comparison across different datasets (Synth, Mimic, PMC) with MCC and F1 scores (including Macro-F1 and Micro-F1 in Overall) for each eviction class. Results present four models: GPT-4o-APO, Llama-3.1-8B-FT, Llama-3.2-3B-FT, Qwen2.5-7B-FT, and Qwen2.5-3B-FT, with overall average metrics. Note that "Eviction_absent" appears exclusively in the Synth dataset as it functions as a critical negative class that establishes decision boundaries between eviction-relevant and irrelevant content during model training. Mimic and PMC datasets deliberately omit this class to focus evaluation specifically on discrimination capabilities between different positive eviction scenarios. Importantly, this omission is not due to data scarcity. On the contrary, most real-world clinical notes in MIMIC and PMC would fall under the "Eviction_absent" category. However, including such easy negatives would inflate overall accuracy and obscure the model's ability to distinguish among nuanced positive classes, so they were excluded by design.

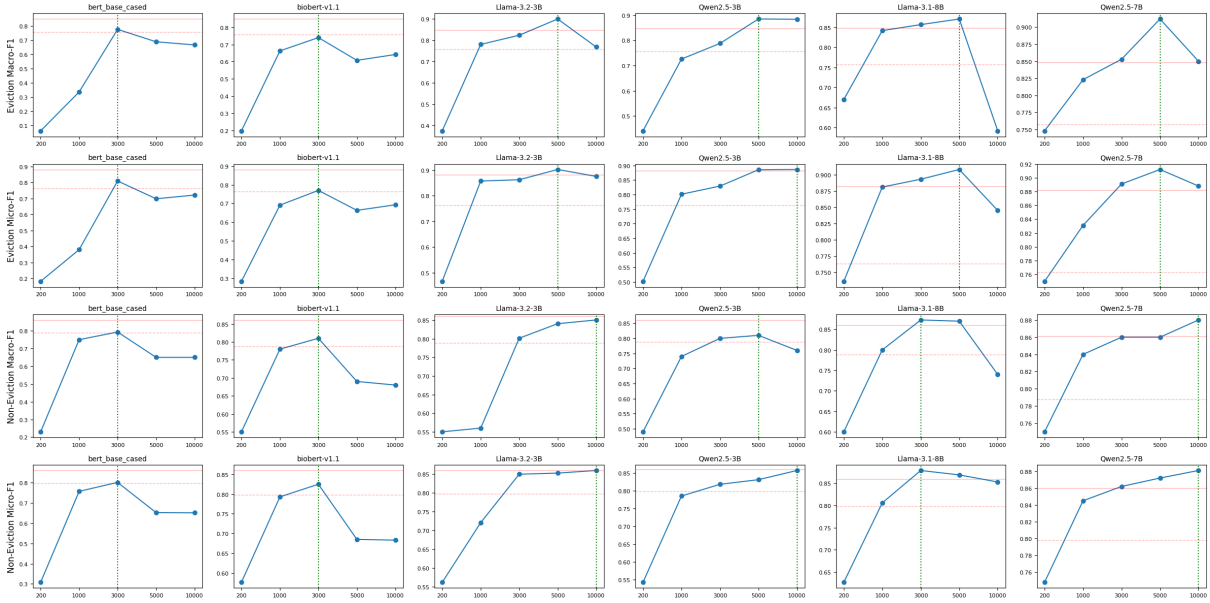

Figure 4: Impact of training set size on model performance across eviction and non-eviction classification tasks. Each subplot illustrates how model performance varies with different training set sizes (x-axis) across four classification scenarios: eviction (top two rows) and non-eviction (bottom two rows), evaluated using two metrics—Macro-F1 (first and third rows) and Micro-F1 (second and fourth rows). The y-axis represents the corresponding F1 score. Green vertical dashed lines indicate the training data size at which each model achieves its highest performance. Red solid and Red dashed horizontal lines denote the performance of GPT-4o-APO and GPT-4o-mini-APO, serving as a reference here.

Step 2: Eviction Multi-Class			
Training Set	GPT-4o-APO	Llama-3.1-8B-FT	Qwen2.5-7B-FT
100% Synth	0.835(0.692)	0.797(0.606)	0.800(0.558)
70% Synth 30% PMC	0.915(0.929)	0.913(0.888)	0.912(0.860)

Step 3: Non-Eviction Multi-Class		
GPT-4o-APO	Llama-3.1-8B-FT	Qwen2.5-7B-FT
0.863(0.789)	0.870(0.800)	0.864(0.793)
0.911(0.922)	0.929(0.942)	0.925(0.944)

Table 6: Comparison of model performance across different training data compositions.

Classification) is also correct for the same clinical note. This evaluation approach reflects the cascading nature of our classification task, where Step 1 determines whether a note contains eviction class, while Steps 2 and 3 classify non-overlapping subsets of notes.

This evaluation framework ensures that models must perform well across the entire classification pipeline to achieve high overall scores, providing a realistic assessment of their practical utility in clinical document analysis.

B.1.1 Training Data Size Impact Experiment

We tested a diverse set of models including traditional LMs (BERT variants) and LLMs (Llama and Qwen families) on both Eviction and Non-Eviction Multi-Class Classification. We used five different training set sizes (200, 1000, 3000, 5000, and 10000 samples) to establish clear performance trends. As the dataset size increased, the number of examples in each class increased proportionally. The datasets for BERT-based models contained only labels without reasoning, while LLM datasets included both labels and reasoning. Each training set was run only once, with consistent fine-tuning parameters across all experiments.

B.2 Reasoning in Trainset Experiment

We trained LLMs (Llama-3.1-8B, Llama-3.2-3B, Qwen2.5-7B, and Qwen2.5-3B) using two distinct training approaches: (1) With Reasoning: Models were trained on datasets containing both classification labels and explicit reasoning steps. (2) Without Reasoning: Models were trained on identical datasets but with only classification labels (reasoning steps removed). All models were evaluated on both Eviction and Non-Eviction Multi-Class tasks using Micro-F1 scores as the primary performance metric. All fine-tuning hyperparameters were kept consistent between the with-reasoning and without-reasoning training conditions to isolate the specific impact of reasoning annotations.

B.3 Metrics

Following Yao et al. (2023b), we calculate the confusion matrix C for both tasks, which are True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each class. We use both **Micro-F1** and **Macro-F1** in our evaluation. We also report the Matthews Correlation Coefficient (**MCC**) score, a metric that yields a high score only if the prediction yields good results in all four confusion matrix categories (TP, TN, FP, and FN).

C Full Results

Our experimental results demonstrate several key findings across different classification tasks.

C.1 Step 1: Main Results for Eviction Mention Detection via Binary Classification

As shown in Table 4, the binary classification task—distinguishing between eviction and non-eviction instances—was relatively straightforward across all model architectures. GPT-4o, when equipped with our APO pipeline, consistently achieved the highest F1 scores across all datasets—0.939 on Synth, 0.938 on MIMIC, and 0.947 on PMC—with an overall average of 0.941. This clearly outperformed the vanilla GPT-4o without APO (average: 0.859), highlighting the benefit of its self-prompt optimization. In contrast, GPT-4o-mini showed no clear gain from APO, suggesting that this smaller LLM may be less capable of benefiting from its self-reflective prompt adjustments for this task.

Open-source LLMs fine-tuned on the SynthEHR-Eviction dataset also demonstrated strong performance. Qwen2.5-7B maintained robust results across all datasets (Synth: 0.963, MIMIC: 0.912, PMC: 0.891), achieving an average F1 of 0.922. LLaMA-3.1-8B performed similarly well, with an average score of 0.908, and achieved particularly strong results on the Synth dataset (0.966). Although 3B-scale models (e.g., LLaMA-3.2-3B, Qwen2.5-3B) performed competitively on Synth, they lagged behind on PMC, indicating that smaller models may struggle with longer or more complex clinical texts.

BERT-based models fine-tuned with our SynthEHR-Eviction pipeline also yielded strong results on the Synth dataset. Notably, bert_base_cased (0.976), biobert-v1.1 (0.987), and Bio_ClinicalBERT (0.978) all outperformed GPT-4o-APO on this dataset. While their performance was less consistent on MIMIC (ranging from 0.756 to 0.916) and generally lower than GPT-4o-APO on PMC, their overall averages remained competitive. For instance, biobert-v1.1 achieved an average F1 of 0.931—slightly surpassing Qwen2.5-7B. This strong performance is likely due to the relative simplicity of the binary classification task and the domain alignment of these biomedical encoders, which are well-suited for short-span eviction cue detection.

C.2 Step 2: Main Results for Multi-Class Classification of Eviction Subcategories

Eviction multi-class classification is substantially more challenging than binary classification because it requires distinguishing between seven fine-grained categories of eviction-related social risk. In this more challenging setup, two strategies demonstrated clear value: fine-tuning with the SynthEHR-Eviction dataset and DSPy-based automated prompt optimization.

First, APO brought significant improvements to GPT-4o and GPT-4o-mini. Since the eviction task is likely not covered in the pretraining corpus of these closed-source models, this represents a classic cold-start scenario. For GPT-4o, APO lifted the average Macro-F1 from 0.515 to 0.878 and Micro-F1 from 0.613 to 0.886. Similarly, GPT-4o-mini showed large gains—Macro-F1 improved from 0.273 to 0.691, and Micro-F1 from 0.433 to 0.764. These results confirm that prompt optimization enables better task alignment and more reliable multi-class reasoning, even in the absence of model fine-tuning.

Meanwhile, models fine-tuned on SynthEHR-Eviction achieved competitive and often stronger performance than the GPT-4o-APO baseline. Qwen2.5-7B reached a Macro-F1 of 0.920 and a Micro-F1 of 0.921 on the Synth dataset, slightly surpassing GPT-4o-APO's 0.918 and 0.909. On MIMIC, it scored 0.939 (Macro-F1) and 0.938 (Micro-F1), again marginally higher than GPT-4o-APO's 0.878 and 0.883. On PMC, Qwen2.5-7B's performance was more comparable, with a Macro-F1 of 0.804 vs. 0.836 and Micro-F1 of 0.824 vs. 0.868. Averaged across datasets, Qwen2.5-7B achieved 0.888 (Macro-F1) and 0.895 (Micro-F1), slightly exceeding GPT-4o-APO's 0.878 and 0.886. LLaMA-3.1-8B also performed robustly, with Macro-F1/Micro-F1 scores of 0.881/0.906 (Synth),

0.935/0.925 (MIMIC), and 0.848/0.861 (PMC), leading to averages of 0.888 and 0.897—again on par with or better than GPT-4o-APO. Importantly, 3B-scale models such as LLaMA-3.2-3B and Qwen2.5-3B performed competitively as well, suggesting that SynthEHR-Eviction enables scalable fine-tuning across model sizes. These 3B models achieved average Micro-F1 scores of 0.871 and 0.874, respectively, demonstrating their deployment potential in resource-constrained environments without sacrificing much performance.

In contrast, BERT-based biomedical models like BioBERT and Bio_ClinicalBERT struggled to generalize across datasets. Their average Macro-F1 and Micro-F1 scores remained below 0.70 and 0.72, respectively, with especially poor results on PMC. This suggests that while these models may encode biomedical terminology, they lack the representational flexibility needed for nuanced, fine-grained classification tasks like eviction status detection.

C.3 Step 3: Main Results for Multi-Class Classification of Non-Eviction Subcategories

Unlike the eviction-specific classification in Step 2, the Step 3 task encompasses a broader range of social risk categories, including homelessness, housing instability, food insecurity, transportation insecurity, and others. Many of these categories are commonly documented in public health and clinical records, making them more familiar to general-purpose LLMs. This is reflected in the relatively strong performance of GPT-4o even without APO, which achieved average scores of 0.797 Macro-F1 and 0.838 Micro-F1—significantly higher than its performance in Step 2.

Nonetheless, prompt optimization continued to enhance closed-source models. GPT-4o-APO reached 0.873 Macro-F1 and 0.889 Micro-F1, improving over the non-optimized version. The GPT-4o-mini also benefited from APO, with Macro-F1 increasing from 0.716 to 0.781, and Micro-F1 from 0.741 to 0.852.

Fine-tuned models based on SynthEHR-Eviction also demonstrated excellent performance. Most notably, Qwen2.5-7B achieved a Macro-F1 of 0.903 and a Micro-F1 of 0.913, both significantly outperforming the GPT-4o-APO baseline ($p < 0.05$). LLaMA-3.1-8B matched this strong performance with 0.883 and 0.915, respectively. These results highlight that even when tasks are not strictly "cold-start," fine-tuning with task-specific high-quality data remains valuable for enhancing performance.

Encouragingly, 3B-scale models also performed well. Qwen2.5-3B achieved 0.821 (Macro-F1) and 0.863 (Micro-F1), while LLaMA-3.2-3B followed with 0.780 and 0.876, respectively—both clearly surpassing GPT-4o-mini-APO. These results demonstrate that SynthEHR-Eviction facilitates efficient scaling, enabling smaller models to reach production-quality performance without reliance on massive human annotation resources.

Finally, the biomedical BERT variants again underperformed across metrics (Macro-F1 < 0.69, Micro-F1 < 0.74), with particularly weak results on MIMIC and PMC. These models may lack the generative flexibility and contextual reasoning needed for multi-class prediction across heterogeneous SDoH categories.

In summary, while Step 3 categories are more likely to have been seen during LLM pretraining, SynthEHR-Eviction still offers substantial fine-tuning benefits. When combined with its superior performance on unseen (eviction-specific) categories, this suggests that our solution serves as a robust and broadly applicable resource for improving SDoH modeling in both known and novel scenarios.

C.4 Class Performance Variations Across Eviction Categories

Table 5 reports the class-wise performance of five models—GPT-4o-APO, Llama-3.1-8B, Llama-3.2-3B, Qwen2.5-7B, and Qwen2.5-3B—on seven nuanced eviction-related categories across three datasets (Synth, Mimic, PMC). These classes vary in eviction status (e.g., `present`, `pending`, `hypothetical`), legal framing (e.g., `mutual rescission`), and temporal cues (e.g., `current` vs. `history`).

All models achieved perfect classification (F1 = 1.00) on **Eviction_absent** in the Synthetic dataset. This class was excluded from Mimic and PMC to prevent dominance of trivial negatives and to emphasize the model's ability to differentiate among positive cases.

Eviction_hypothetical, representing projected or threatened eviction, posed moderate challenges. GPT-4o-APO reached an average F1 of 0.86, while several fine-tuned open-source LLMs—specifically Qwen2.5-7B and Llama-3.2-3B—surpassed this benchmark with Synth scores of 0.98 and 1.00, and solid generalization in Mimic (0.92 and 0.95), highlighting the benefit of domain-specific fine-

tuning on SynthEHR-Eviction.

Eviction_mr_current and **Eviction_mr_history**, which capture mutually agreed lease terminations, were consistently well-handled. Qwen2.5-7B-FT achieved 0.93 and 0.92 on Macro-F1 and Micro-F1, respectively, slightly outperforming GPT-4o-APO (F1 = 0.90 and 0.86). Notably, fine-tuned 3B models also maintained high accuracy in these categories (F1 is about 0.88), suggesting strong transfer of structured legal signals even in smaller models after fine-tuning.

Eviction_pending, indicating formal but unresolved eviction processes, remained somewhat ambiguous. Qwen2.5-7B and Llama-3.1-8B achieved F1 = 0.91 and 0.95 in Synth and Mimic respectively, surpassing GPT-4o-APO's average of 0.87. However, smaller models showed more variance across datasets, especially in PMC.

Eviction_present_current and **Eviction_present_history**, which rely on subtle differences in recency and documentation of completed evictions, remained the most challenging. In PMC, for example, Llama-3.1-8B dropped to 0.76 on "present_current", while GPT-4o-APO maintained a stronger F1 of 0.93. These results highlight that while fine-tuned LLMs can excel in categories grounded in legal constructs, temporal ambiguity still presents difficulties.

In summary, class-level comparisons reveal that fine-tuned open-source LLMs—especially Qwen2.5-7B and Llama-3.2-3B—can match or exceed GPT-4o-APO in multiple eviction categories after training on SynthEHR-Eviction. Larger models (7B/8B) showed better generalization across datasets, while smaller models (3B) still offered strong performance on well-defined categories. GPT-4o-APO provided a competitive baseline with consistent performance across the board, though not uniformly superior to well-trained open models.

C.5 Impact of Reasoning in Training Data

To evaluate the effect of explicit reasoning annotations in training data, we prepared two versions of each model's training set: one with reasoning explanations and one without, keeping the labels identical. As shown in Figure 3, smaller models benefit more noticeably from reasoning annotations. For instance, Llama-3.2-3B-FT improved from 0.848 to 0.886 (+3.8), and Qwen2.5-3B-FT improved from 0.855 to 0.868 (+1.3). In contrast, larger models like Qwen2.5-7B-FT and Llama-3.1-8B-FT maintained stable performance with or without reasoning, with minimal differences (less than 0.01). These findings suggest that the reasoning annotations generated through the SynthEHR-Eviction pipeline can provide varying degrees of structured guidance, especially for smaller models that benefit from clearer decision logic in complex classification tasks. Additionally, models trained with these annotations are capable of generating their own reasoning during inference, improving transparency and interpretability. This makes them more amenable to expert review and downstream validation, moving beyond black-box eviciton prediction.

C.6 Training Data Size Impact on Finetuning

To examine how training data size affects model performance, we conducted systematic experiments on two traditional BERT-based models (BERT and BioBERT) and four open-source LLMs (two 3B-scale and two 7B/8B-scale LLMs). We evaluated both the Eviction Multi-Class and Non-Eviction Multi-Class tasks using training datasets of varying sizes (200 to 10,000 samples) and reported both Macro-F1 and Micro-F1 scores (Figure 4).

Across both tasks, models demonstrated substantial gains when the training data was increased from 200 to 3,000 samples. For instance, Qwen2.5-7B improved its Macro-F1 on the Eviction task from 0.748 to 0.853, and on the Non-Eviction task from 0.75 to 0.891. Similarly, LLaMA-3.1-8B increased from 0.670 to 0.857 (Eviction) and from 0.736 to 0.893 (Non-Eviction). Performance gains typically plateaued beyond 5,000 samples, and several models exhibited signs of overfitting at 10,000 samples. For example, Qwen2.5-7B reached its peak Macro-F1 of 0.912 on the Eviction task at 5,000 samples, then dropped to 0.850 at 10,000 samples.

Optimal dataset sizes varied by task. For Step 2 (Eviction), LLMs such as Qwen2.5-7B required larger datasets to reach optimal performance (e.g., 0.912 Macro-F1 at 5,000 samples). In Step 3 (Non-Eviction), most models converged earlier, reaching their peak performance around 3,000 samples (e.g., Qwen2.5-7B at 0.891).

For practical fine-tuning, training with 3,000–5,000 samples provides a strong balance between performance and efficiency, since several LLMs trained on just 3,000–5,000 samples reached performance levels comparable to GPT-4o-APO. Accordingly, we used 5,000 samples for LLMs

and 3,000 samples for traditional BERT models in Step 2. For Step 3, we used 3,000 samples for both model types, as performance had largely converged at this size.

Performance Variations Across Different Devsets Our comprehensive analysis compared performance across three distinct datasets: Synth, MIMIC, and PMC in the Eviction Multi-Class task (Figure 5). The Synth dataset consistently demonstrated the most stable performance characteristics, exhibiting high and uniform $F1$ scores across eviction classes. Similarly, the MIMIC dataset showed minimal variability, with stable and reliable performance across most classes. In contrast, the PMC dataset demonstrated the most significant performance heterogeneity, characterized by the widest performance ranges and lower median scores. This performance degradation aligns with the trends observed in Table 4, where models exhibited substantially lower micro-F1 scores on PMC data compared to Synth and MIMIC datasets. These results highlight notable inter-dataset variations, particularly emphasizing the challenges posed by PMC notes.

Interpretation of Inter-Dataset Variability The performance discrepancies between datasets, particularly the pronounced heterogeneity in PMC results, can be attributed to both structural differences in the clinical notes and model limitations. Specifically, PMC notes tend to be significantly longer and more detailed, leading to truncation due to our 512-token input limit, which likely eliminates crucial contextual information. Moreover, unlike the Synth and MIMIC datasets, where eviction-related information is often explicitly stated and follows a structured format, PMC notes embed relevant details within complex, narrative clinical discussions. This structural complexity complicates the model's ability to generalize, especially when it has been primarily trained on more structured datasets.

Table 7: Eviction-related Class Definitions

SDoH Class	Definition
Eviction_present_current	1. Eviction Present refers to the state where the eviction process has already been fully concluded, and the tenant has been legally removed from the property. All legal proceedings, such as notices and hearings, have been completed, and the tenant no longer has access to the property. - Consideration: When generating the note, it should mention that the eviction has already taken place, and the tenant is no longer living at or has access to the property. 2. Current refers to events related to eviction that happened currently or recently, with specificity about the timing. This includes cases where the eviction event occurred in the current year (e.g., "this year," "last month", "last week", "several months ago," or "a few months ago", "recently"...). - Consideration: The note should mention that the event was in the current year, providing some specific time reference.
Eviction_present_history	1. Eviction Present refers to the state where the eviction process has already been fully concluded, and the tenant has been legally removed from the property. All legal proceedings, such as notices and hearings, have been completed, and the tenant no longer has access to the property. - Consideration: When generating the note, it should mention that the eviction has already taken place, and the tenant is no longer living at or has access to the property. 2. History refers to events related to eviction that happened in the past, but with less specificity about the timing. This includes cases where the eviction event occurred in the distant past (e.g., "last year," "several years ago," or "a few years ago"), or where the mutual rescission agreement or eviction action itself isn't explicitly stated. - Consideration: The note should mention that the event was in the past, providing some general time reference, but without specifying the exact date or events.
Eviction_pending	Eviction Pending refers to a situation where eviction proceedings have been initiated but are not yet complete. The tenant is still in the property, and there is still an opportunity for negotiation, remediation, or resolution before a final court decision or physical removal occurs. This state indicates that while the eviction process has started, the outcome is still undecided, and there is a potential for the tenant to address the issue and avoid eviction. - Distinction from Completed Eviction: Unlike a completed eviction where the tenant has already been removed, eviction-pending indicates that the tenant has received a notice, but there has been no final court order or physical removal. The tenant may negotiate, pay overdue rent, or comply with other conditions to potentially stop the eviction process. Few-Shot Examples: - "The tenant received an eviction notice recently, but negotiations with the landlord to pay overdue rent are still ongoing." - "A few months ago, the landlord filed for eviction due to nonpayment, but the case is still pending in court, giving the tenant a chance to settle." - "Currently, the tenant is under an eviction notice but is working with a housing advocate to resolve the issue before the court date."
Eviction_hypothetical	Eviction Hypothetical refers to situations where the eviction process is anticipated or expected to occur but has not yet been initiated. The landlord has expressed an intention to proceed with eviction but has not issued any notice. The tenant is still in the property, and the eviction is expected to occur in the future, though the exact timeline may vary based on the tenant's actions or further legal procedures. - Time Frame: "hypothetical" in this context refers to actions or plans related to eviction that are expected to occur soon. This could include statements like "in the coming weeks," "next month," or "soon," indicating that the eviction is planned but not yet happend. Few-Shot Examples: - "The landlord has given the tenant a final warning, and eviction proceedings are expected to start next month if the rent is not paid." - "The tenant has been notified that they must vacate the premises in the coming weeks due to repeated violations of the lease." - "Eviction is planned for the near future, as the landlord has expressed intent to reclaim the property due to nonpayment." - "The landlord mentioned that they will file for eviction soon if the tenant does not comply with the notice to remedy the situation." - "Next week, the landlord plans to serve an eviction notice due to ongoing illegal activities on the property."
Eviction_mr_current	1. Mutual Rescission (mr) refers to a specific legal agreement in which both the landlord and tenant agree to terminate the lease early. This occurs after eviction proceedings have started, but the eviction process has not yet reached its final stage. As a result of this agreement, the tenant voluntarily vacates the rental property, and the eviction process is stopped before completion. All necessary legal proceedings are concluded, and the tenant no longer has access to the property. - Consideration: When generating the note, it should mention that the Mutual Rescission has already taken place, the eviction process had already stopped, and the tenant is no longer living at or has access to the property. 2. Current means that the agreement or action related to mutual rescission is mentioned within a recent period (e.g., "a few months ago," "recently," "this year").
Eviction_mr_history	1. Mutual Rescission (mr) refers to a specific legal agreement in which both the landlord and tenant agree to terminate the lease early. This occurs after eviction proceedings have started, but the eviction process has not yet reached its final stage. As a result of this agreement, the tenant voluntarily vacates the rental property, and the eviction process is stopped before completion. All necessary legal proceedings are concluded, and the tenant no longer has access to the property. - Consideration: When generating the note, it should mention that the Mutual Rescission has already taken place, the eviction process had already stopped, and the tenant is no longer living at or has access to the property. 2. History refers to events related to eviction that happened in the past, but with less specificity about the timing. This includes cases where the Mutual Rescission occurred in the distant past (e.g., "last year," "several years ago," or "a few years ago"), or where the mutual rescission agreement or eviction action itself isn't explicitly stated. - Consideration: The note should mention that the event was in the past, providing some general time reference, but without specifying the exact date or events.
Eviction_absent	The text clearly states "never evicted" or "no history of eviction".

Table 8: Non-Eviction-related Class Definition

SDoH	Definition (/Examples)
Homelessness	An individual or family who lacks a fixed, regular, and adequate nighttime residence, such as those living in emergency shelters, transitional housing, or places not meant for habitation. Here are some few-shot examples: "...is homeless and lives in a shelter. . . ", "...contact one of the homeless shelters....", "...found it difficult to secure housing and ended up living in his car. . . ", "...relying on friends and temporary shelters for support. . . ", "...is actively seeking employment and more permanent housing but has faced numerous obstacles. . . ", "living on the streets. . . ", "...living in a homeless encampment...", "...couch surfing..."
InadequateHousing	Inadequate housing is defined as an occupied housing unit that has moderate or severe physical problems (e.g., deficiencies in plumbing, heating, electricity, hallways, and upkeep). Examples of moderate physical problems in a unit include two or more breakdowns of the toilets that lasted more than 6 months, unvented primary heating equipment, or lack of a complete kitchen facility in the unit. Severe physical problems include lack of running hot or cold water, lack of a working toilet, and exposed wiring. Here are some few-shot examples: "...lives in an old apartment building with severe structural issues...", "...live in an apartment that lacks a functioning heating system. . . ", "...a family, consisting of six members spanning three generations, lives in a cramped two-bedroom apartment. . . ", "...unsafe housing situation. . . ", "...unsanitary living conditions. . . ", "...polluted living environment. . . ", "...lead and toxic exposures in home. . . "
LackOfAdequateFood	Food insecurity is the limited or inadequate access to food because of insufficient money and other resources for food. Food security, at the individual, household, national, regional, and global levels [is achieved] when all people, at all times, have physical and economic access to sufficient, safe, and nutritious food to meet their dietary needs and food preferences for an active and healthy life. This definition suggests that food insecurity is the absence of one or more of these conditions. Here are some few-shot examples: "...lack the variety and nutrients... ", "...frequently goes hungry or eats whatever is available...", "...no supermarkets...", "...difficult to access better food sources outside the neighborhood...", "...uses food pantries/soup kitchens for food. . . ", "...does not have stable food sources. . . ", "...lives in a food desert. . . "
FinancialInsecurity	Economic insecurity can be defined as "the anxiety produced by the possible exposure to adverse economic events and by the anticipation of the difficulty to recover from them". Examples could include a fear of unemployment, an expectation of a worsening financial situation, money mismanagement, or being financially exploited or a victim of financial scam. Here are some few-shot examples: "...has been having a lot of stress recently due to financial concerns.", "...was very concerned about the financial burden of hospitalization, medications and potential surgery.", "Rising living costs, including healthcare and housing, have made it difficult for them to cover their monthly expenses.", "The irregular income and lack of benefits make it difficult to budget and plan for the future.", "...has experienced several months of financial difficulty due to job loss.", "...not managing money well. . . ", "...lacks financial literacy. . . ", "...mismanaging funds.", "...lack of stable income. . . "
HousingInstability	Housing instability is variably defined as having difficulty paying rent, spending more than 50% of household income on housing, having frequent moves, living in overcrowded conditions, or doubling up with friends and relatives. Unstably housed, housing insecure, or in a temporary housing situation. At risk of being homeless or at imminent risk of being homeless. Here are some few-shot examples: "...has moved three times...", ". . . fell behind on his rent payments...", "...live together with three families to share...", "...couch surfing. . . ", "...temporarily staying with friends/family. . . ", "...staying in a motel/hotel. . . ", "...in a temporary housing situation. . . ", "...at risk of losing their housing. . . "
MaterialHardship	Material hardships, defined as difficulty meeting basic needs such as food, housing or medical care, are common among low-income households. Here are some few-shot examples: "Their electricity was cut off because they couldn't make the payments.", "...cannot afford to buy winter coats or shoes that fit correctly.", "...cannot afford the necessary school supplies and textbooks for their three children.", "...cannot afford to participate in health and wellness activities such as exercise classes, which are essential for managing his chronic conditions."
TransportationInsecurity	Transportation insecurity occurs when a person is regularly unable to get from place to place in a safe or timely manner because of a lack of resources. This can limit a person's access to work, school, medical care, social activities, and more. Here are some few-shot examples: ". . . lives in a rural area where there are no public transportation options. The nearest town with essential services such as grocery stores, schools, and healthcare facilities is 20 miles away.", "Buses and trains are often delayed or overcrowded, making it difficult for her to get to work on time.", "...can't afford the transportation fare, he and his children have to walk long distances, often in unsafe conditions.", "...uses a wheelchair and often finds that public transportation in her city is not fully accessible.", "...does not own a car. . . ", "...does not have bus passes. . . "

Table 9: Human Instructions / LLM Prompt

Instruction / Prompt	Detail
Augmenting Prompt	You are tasked with rewriting the raw notes to become related to some SDoH label. Here is the raw note: raw_notes And the specific label: $label$ The definition of the label: $definition$ Note: 1. Do NOT include the definitions or examples provided in your rewrited content. 2. Focus exclusively on the social history, disregarding sentences related to family history or other topics. 3. The augmented notes should clearly reflect the label context, be contextually coherent, with varied and diverse expressions. 4. The augmented note should be a detailed description of a specific patient case that illustrates the application of the SDoH label. Focus on the unique circumstances, events, and actions related to this individual case. Avoid using general or broad descriptions of processes or procedures; instead, provide concrete details and examples that are directly relevant to the patient's situation. 5. Your output should not exceed 100 words. Augmented Notes: input_variables=["raw_notes", "label", "definition"]
Annotation Prompt in Step 1: Binary Classification	Go through each sentence of the patient note. If a sentence reflects eviction-related social determinants of health (SDoH), assign the label "Yes", else annotate as label "No"
Annotation Prompt in Step 2: Eviction Multi-Class	Go through each sentence of the patient note. If a sentence reflects eviction-related social determinants of health (SDoH), assign the most appropriate label from the following list: "t3_Eviction_absent", "t3_Eviction_present_history", "t3_Eviction_present_current", "t3_Eviction_pending", "t3_Eviction_mr_history", "t3_Eviction_mr_current", "Other". For status part, if no eviction in the history and in the future: "absent"; if eviction is completed: "present"; if eviction noticed but not completed: "pending"; if eviction might be happend in the future: "hypothetical"; if mutual rescission: "mr". For timeframe part when "present" or "mr" status, if it is happened within this natural year: "current". If not shown specific time or noticed a time before this natural year: "history".
Annotation Prompt in Step 3: Non-Eviction Multi-Class	Choose the most approperate label from "t1_Homelessness", "t1_InadequateHousing", "t1_LackOfAdequateFood", "t2_FinancialInsecurity", "t2_HousingInstability", "t2_MaterialHardship", "t2_TransportationInsecurity", "Other". 't1_Homelessness': An individual or family who lacks a fixed, regular, and adequate nighttime residence, such as those living in emergency shelters, transitional housing, or places not meant for habitation. 't1_InadequateHousing': an occupied housing unit that has moderate or severe physical problems (e.g., deficiencies in plumbing, heating, electricity, hallways, and upkeep) 't1_LackOfAdequateFood': is the limited or inadequate access to food because of insufficient money and other resources for food. 't2_FinancialInsecurity': the anxiety produced by the possible exposure to adverse economic events and by the anticipation of the difficulty to recover from them. Examples could include a fear of unemployment, an expectation of a worsening financial situation, money mismanagement, or being financially exploited or a victim of financial scam. 't2_HousingInstability': having difficulty paying rent, spending more than 50% of household income on housing, having frequent moves, living in overcrowded conditions, or doubling up with friends and relatives. 't2_MaterialHardship': difficulty meeting basic needs such as food, housing or medical care, are common among low-income households. 't2_TransportationInsecurity': occurs when a person is regularly unable to get from place to place in a safe or timely manner because of a lack of resources. This can limit a person's access to work, school, medical care, social activities, and more.
Human Annotation for Augmentation Period	We need to check the quality of the GPT-generated notes to refine our prompts, and this requires manual verification—your role in this project. You'll review the entries, marking them 'True' if correct, and giving feedback if false. We have 14 labels, (including "t1_Homelessness", "t1_InadequateHousing", "t1_LackOfAdequateFood", "t2_FinancialInsecurity", "t2_HousingInstability", "t2_MaterialHardship", "t2_TransportationInsecurity", "t3_Eviction_absent", "t3_Eviction_present_history", "t3_Eviction_present_current", "t3_Eviction_pending", "t3_Eviction_mr_history", "t3_Eviction_mr_current"), which make up our initial data set, with 20 entries for each label. We've engaged three experts to ensure accuracy by cross-verifying the data. Each label is annotated by two different experts, so each expert will annotate over 200 entries. If you are not sure the definitions of some label, you can check the document (Same as Table 7 and 8).
Manual Rewrite Notes for Augmentation	We need your assistance to rewrite clinical notes from MIMIC to reflect specific social determinants of health (SDoH) classes related to eviction. For each note and assigned class/label, please modify the content to clearly demonstrate the corresponding eviction situation while preserving the original MIMIC documentation style. Your rewrites should maintain the clinical tone, formatting, and structure of the original notes (including headers, abbreviations, and documentation patterns typical in MIMIC), but incorporate eviction-related circumstances that align with the specified class (e.g., t3_Eviction_present_history, t3_Eviction_pending, etc.). The goal is to create authentic-looking clinical documentation that can be used to augment our training data for eviction classification models. If you are not sure the definitions of some label, you can check the document (Same as Table 7 and 8).
Manual Double-Check Rewritten Notes	We have some data entries related to eviction that were rewritten by medical students. However, since they are not experts, we need your help to double-check their correctness. This is the same as the task you've worked on before—just review and mark each entry as either True or False.

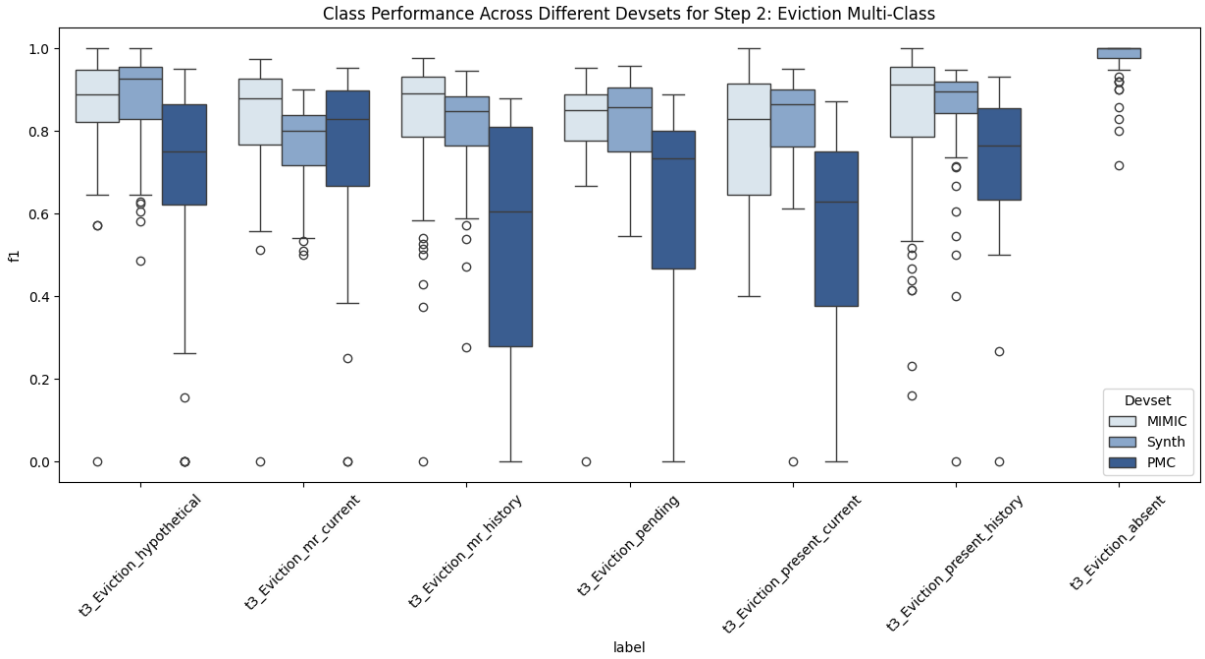

Figure 5: Class Performance Across Different Devsets for Step 2: Eviction Multi-Class. Box plot revealing $F1$ scores across three datasets (Synth, MIMIC, and PMC) for all evaluated models, including GPT-4o-APO, GPT-4o-mini-APO, fine-tuned Qwen, Llama, SBDH-LLM, BERT, Clinical-Bert and SDoH-BERT. Each model was evaluated across 5 independent runs, revealing both inter-dataset variations and substantial performance differences between eviction-related labels

Training Set	Sample Note
Synth	The patient, a widowed immigrant living with a relative, faces potential housing instability. Having lived in the country for about 50 years, her current housing situation may be threatened, necessitating discussions about mutual rescission to ensure she can voluntarily vacate her rental early if needed, which reflects her vulnerabilities related to social determinants of health.
MIMIC	Social History: + Tob, 1.5 ppy X many years, no EtOH, was forced to remove from the rented house, has a child in their 20s
PMC-Patient	Our patient is a girl aged 10–14 years with an unremarkable medical history, yet her family history is significant for allergies and asthma. She initially presented, to another institute, with multiple pruritic facial skin lesions and a pruritic left intranasal lump. Apart from having two erythematous pruritic plaques in the left suborbital region and a yellowish pruritic lump occupying the left nasal vestibule, her physical examination proved to be insignificant without any lymphadenopathies or salivary gland enlargements. Consequently, a laboratory workup was conducted in addition to an excisional biopsy of one dermatologic lesion and a needle biopsy of the nasal lesion. Both biopsies exhibited nonspecific inflammation with granulation, necrosis, and no signs of malignancy. No specific diagnosis was made. The patient was started on hydrogen peroxide treatment for the skin lesions, which resolved completely with no recurrence. Simultaneously, another lump started growing in the right nasal vestibule. Suspecting an inflammatory etiology, she was started on oral prednisolone 1 mg/kg/day by mouth twice a day. Despite therapy, these nasal lumps continued growing. Hence, prednisolone was discontinued after 10 days of therapy to start an indomethacin trial of 1.5 mg/kg/day by mouth twice a day for 14 days, which was also discontinued due to its inefficacy. Then a decision to perform surgery was taken, and the presurgical computed tomography (CT) scan revealed bilateral soft tissue masses arising from the right and left nasal vestibules. Although she had undergone many surgical attempts to remove the lumps, none of them succeeded and both lumps flared in size. After a couple of months, our patient presented to our institution with bilateral nasovestibular lumps; they were massive in size, occluding nasal entrance and protruding outside the nose (Fig.). We did an extensive laboratory workup to exclude any comorbidities (Table). We did a fine-needle aspiration (FNA) biopsy of the lesion, which was diagnostic of ALHE (Fig.). Our following surgical attempt included complete mass resection. Despite surgery and postsurgical treatment with topical steroid creams, the lesion recurred. Consequently, we started the patient on intralesional prednisolone twice a month and topical 0.1% tacrolimus ointment twice daily. This latter regimen seemed to slightly control the lesion's growth, causing a limited regression in size after 4 months of treatment (Fig.). A timeline of the patient's case can be seen in (Fig.). A family member reported her full adherence to treatment. The patient herself reported decreased quality of life and impaired social interactions due to the disfiguring lesions. She also reported marked fear and distress because of the ineffectiveness of multiple therapeutic regimens and surgeries. A relative reported severe financial burden due to the high costs of the treatments. The patient was evicted from the rented house due to a series of late rent payments recently.

Table 10: Sample Patient Notes Across Different Training Datasets (Age ranges and family relations anonymized)

Table 11: Pipelines for Augmenters and Annotators

Algorithm 1: Pipeline for Each Augmenter

Require: Definition of SDoH category D, Raw note N
Ensure: Optimized prompt, Verified augmented notes

```
Model: LLM
input_prompt ← D                                   ▷ Initialize with category definition
while True do
    augmented_notes ← {}
    output ← LLM(input_prompt)                     ▷ Generate notes based on D and N
    for each note n in output do
        verification, feedback ← HumanVerify(n)
        if verification == True then
            augmented_notes ← augmented_notes ∪ {n}
        else
            Save feedback
        end if
    end for
    accuracy ← Number of True verifications / Total number of notes
    if accuracy ≥ 0.90 then
        break
    else
        input_prompt ← OptimizePrompt(input_prompt, feedback)
    end if
end while
return input_prompt, augmented_notes
```

Algorithm 2: Pipeline for Annotators

Require: Training set $(notes, labels)$, Development set $(notes, labels)$
Ensure: Trained DSPy models, Classification of SDoH categories

```
Model: DSPy
Step 1: Binary Classification
trainset_1 ← (notes, labels)                       ▷ Eviction-related (Yes/No)
model_1 ← TrainDSPy(trainset_1)
accuracy_1 ← ValidateModel(model_1, devset)
Step 2: Eviction Multi-Class
trainset_2 ← (notes, labels)                       ▷ 7 eviction-related classes
model_2 ← TrainDSPy(trainset_2)
accuracy_2 ← ValidateModel(model_2, devset)
Step 3: Non-Eviction Multi-Class
trainset_3 ← (notes, labels)                       ▷ 7 non-eviction classes
model_3 ← TrainDSPy(trainset_3)
accuracy_3 ← ValidateModel(model_3, devset)
return model_1, model_2, model_3
```

Table 12: Keywords

SDoH	Keywords
t1_Homelessness	homeless, homelessness, shelter, transitional housing, living in car, living on streets, couch surfing, lacks housing, no fixed residence, no permanent home
t1_InadequateHousing	inadequate housing, structural issues, structural problems, deficiencies in plumbing, heating problems, no heating, electrical problems, lack of running water, broken toilet, no toilet, no kitchen, cramped apartment, overcrowded, unsafe housing, unsanitary living, polluted environment, lead exposure, toxic exposure, mold
t1_LackOfAdequateFood	food insecurity, limited access to food, insufficient food, lacks variety, lacks nutrients, no supermarkets, difficult to access food, unstable food sources, cannot afford food, malnutrition, undernourished, skipping meals, reliance on food assistance
t2_FinancialInsecurity	financial insecurity, economic insecurity, financial concerns, financial stress, financial burden, affordability issues, rising living costs, difficulty covering expenses, budget difficulties, lacks financial literacy, no stable income, debt, financial hardship
t2_HousingInstability	housing instability, difficulty paying rent, frequent moves, multiple moves, families sharing housing
t2_MaterialHardship	material hardship, difficulty meeting basic needs, utilities cut off, cannot afford clothing, winter coats, cannot afford school supplies, cannot afford health activities, limited resources for essentials, unable to afford medications, basic needs not met, lacks essential household items
t2_TransportationInsecurity	transportation insecurity, lack of transportation, no public transportation, transportation issues, cannot afford transportation fare, long walking distances, inaccessible transportation, no car, unreliable transportation, limited mobility, transportation costs prohibitive

Table 13: Error Analysis - Case 1-3

Case	Analysis
Case 1	**Physician Note**: The patient, a retired pathologist, experienced eviction due to ongoing financial strain as a relative battles a chronic illness. Despite his extensive career, their medical expenses have caused significant challenges, so he resides in a group home at [**Hospital1 3494**] now. **Ground Truth**: t3_Eviction_present_history **Prediction**(Llama-3.1-8B-FT): t3_Eviction_present_current **Analysis**: This misclassification likely stems from temporal ambiguity in the text. While the eviction explicitly occurred with no specific timeframe mentioned (indicating a historical event), the phrase 'resides in a group home at [Hospital1 3494] now' may have caused the model to associate the current housing situation with ongoing eviction issues. This highlights the challenge in distinguishing between historical events and currently active situations, particularly when the specific timing of the eviction is not mentioned and 'now' appears in the context.
Case 2	**Physician Note**: The patient reports a history of housing instability, having faced an eviction process stemming from unpaid rent two months ago. This event has exacerbated her ongoing health issues, such as bilateral lower extremity edema and nocturnal dyspnea, highlighting the interplay between her living situation and overall well-being. **Ground Truth**: t3_Eviction_present_current **Prediction**(Llama-3.1-8B-FT): t3_Eviction_present_history **Analysis**: In this case, the model misclassified the eviction situation as `present_history` rather than `present_current`. This error likely stems from multiple factors in the text. First, the explicit mention of "reports a history of housing instability" may have primed the model to categorize the entire situation as historical, despite the more recent "two months ago" timeframe which falls within our one-year threshold for `present_current` classification. The lexical cue "history" appears early in the sentence, potentially biasing the model's interpretation of subsequent temporal information. Additionally, phrases like "having faced an eviction process" use past-tense constructions that might further reinforce a historical framing, even though the event occurred recently. This demonstrates how models can be overly influenced by specific terminology like "history" rather than properly processing the actual temporal indicators ("two months ago") in relation to the annotation schema. This challenge highlights the need for models to prioritize explicit timeframes over potentially misleading lexical cues when making temporal classifications according to specific guidelines.
Case 3	**Physician Note**: The patient lives with a relative, who has been recently ill, leading to challenges in meal preparation and unintended weight loss. This situation contributes to financial strain, leading to a current eviction process initiated by their landlord. **Ground Truth**: t3_Eviction_pending (rationale='produce the answer. We need to identify any mention of eviction-related social determinants of health (SDoH) in the patient's note. The note describes a situation where the patient and a relative are currently undergoing an eviction process initiated by their landlord due to financial strain. This indicates that the eviction process is active and ongoing. Therefore, the most appropriate label for this situation is "t3_Eviction_pending".') **Prediction**(Llama-3.2-3B-FT): **Training set with Reasoning**: t3_Eviction_pending. Output_Reasoning: determine the appropriate label for the patient's eviction status. The fact states that the patient is currently facing an eviction process initiated by their landlord, indicating that eviction proceedings are ongoing. Since the eviction is not yet completed and the patient is still residing in the property while dealing with financial strain, we categorize this situation as "pending." **Training set without Reasoning**: t3_Eviction_present_current **Analysis**: This case offers valuable insights into how the presence of reasoning in training data affects both prediction accuracy and reasoning quality. When trained with reasoning annotations, the Llama-3.2-3B-FT model correctly identifies the eviction status as "t3_Eviction_pending." However, when the same model is trained without reasoning, it incorrectly classifies the situation as "t3_Eviction_present_current," failing to distinguish between an active ongoing eviction process (pending) and a completed recent eviction (present_current). Comparing the reasoning produced by the model trained with reasoning to the ground truth reasoning, we observe important similarities and differences. Both correctly identify the key evidence: "eviction process initiated by their landlord." Both reasoning paths also correctly emphasize that this indicates an ongoing, active eviction situation rather than a completed one. However, the model's reasoning is more concise and adds the specific detail that "the patient is still residing in the property," which is implied but not explicitly stated in the ground truth reasoning.

Table 14: Error Analysis - Case 4-6

Case	Analysis
Case 4	**Physician Note**: The individual does not currently use tobacco and has not faced eviction. He drinks up to 10 beers every weekend. He has used cocaine several times a week for the past several years, but has not used any for a month. He faces potential housing instability after mutually agreeing to terminate his lease. Previously retired and supporting troubled youth, the individual now navigates social pressures and community resources while managing his lifestyle choices, including alcohol consumption and past smoking history, which may impact his wellbeing. **Ground Truth**: t3_Eviction_mr_current (rationale='produce the answer. We need to identify any mention of eviction-related social determinants of health (SDoH) in the individual's note. The note describes a situation where the individual has mutually agreed to terminate her lease, which indicates a mutual rescission of the lease. This suggests that the eviction process is not adversarial but rather a mutual agreement to end the lease. Therefore, the most appropriate label for this situation is "t3_Eviction_mr_current" as the mutual rescission is happening currently.') **Prediction**(Llama-3.1-8B-FT): t3_Eviction_mr_current (Reasoning: determine the appropriate label for the patient's eviction status. The individual has mutually agreed to terminate her lease, indicating that she is currently in a situation where she is not facing eviction in the traditional sense, but rather has made a decision to vacate the property. This mutual rescission of the lease suggests that she is actively navigating her housing situation, but it does not indicate an eviction that has already occurred or is currently in progress. Therefore, we categorize this situation as a mutual rescission of the lease.) **GPT-as-judge**: False. The reasoning provided does not fully align with the ground truth. The ground truth clearly states that the situation should be labeled as 't3_Eviction_mr_current' because the mutual rescission of the lease is happening currently. However, the reasoning suggests that the individual is not facing eviction in the traditional sense and does not indicate an eviction that has already occurred or is currently in progress. This implies a misunderstanding of the label 't3_Eviction_mr_current,' which is meant to capture the current mutual rescission of the lease. The reasoning fails to recognize that a mutual rescission is a form of eviction-related SDoH and should be labeled as such. Therefore, there is a gap in understanding the categorization of the situation as per the ground truth.
Case 5	**Physician Note**: The patient lives at home with family members, one of whom is ESRD on HD and another provides care for the patient. They are in process of eviction because the landlord issues a notice due to missed rent payments stemming from financial strain. **Ground Truth**: t3_Eviction_pending (rationale='produce the answer. We need to identify any mention of eviction-related social determinants of health (SDoH) in the patient's note. The note describes a situation where the patient and their family are in the process of eviction due to missed rent payments caused by financial strain. This indicates that eviction proceedings have started but are not yet completed. Therefore, the most appropriate label for this situation is "t3_Eviction_pending".') **Prediction**(Llama-3.1-8B-FT): t3_Eviction_pending (Reasoning: produce the answer. The facts indicate that the patient is currently facing eviction due to missed rent payments, which have resulted in a notice from the landlord. Since the eviction process has not yet been completed, but is in progress, the status is "pending." The timeframe is not specified as occurring within the current year, so we will label it as "history." Therefore, the appropriate label will reflect that the eviction is not yet finalized, but is actively being pursued.) **GPT-as-judge**: False. The reasoning provided does not logically support the ground truth. The ground truth clearly identifies the situation as 't3_Eviction_pending' based on the fact that eviction proceedings have started but are not yet completed. However, the reasoning introduces an additional and incorrect element by labeling the situation as 'history' due to unspecified timing. This contradicts the ground truth, which does not consider the timeframe as a factor for labeling. The correct label should solely reflect the current status of the eviction process, which is 'pending,' without introducing the notion of 'history.' Therefore, the reasoning is incorrect as it deviates from the ground truth by adding an unnecessary and unsupported condition.
Case 6	**Physician Note**: Social History: Lives alone with PCA 20 hours per day. The individual recently faced financial strain, which led to eviction from her rental home due to unpaid rent. Fortunately, a relative lives nearby, so she moved in with the relative. **Ground Truth**: t3_Eviction_present_current (rationale='produce the answer. We need to identify any mention of eviction-related social determinants of health (SDoH) in the patient's social history. The note indicates that the individual recently faced financial strain, leading to an eviction from her rental home due to unpaid rent. She moved to her relative's home, which suggests that the eviction process was completed, and she found an alternative living arrangement. Therefore, the most appropriate label for this situation is "t3_Eviction_present_current") **Prediction**(Llama-3.1-8B-FT): t3_Eviction_present_current (Reasoning: determine the appropriate label for the patient's eviction status. The fact states that the individual is currently facing financial strain and has not paid rent, which has led to an eviction notice. Since the eviction process has not yet been completed but is in progress, we categorize this situation as "present." Additionally, the individual is currently living with a relative, indicating a temporary housing solution, but the ongoing eviction status suggests that the situation is still precarious.) **GPT-as-judge**: False. The reasoning provided does not logically support the ground truth. The ground truth indicates that the eviction process was completed, as the individual moved to a relative's home after being evicted due to unpaid rent. This situation is labeled as 't3_Eviction_present_current,' which implies that the eviction has already occurred and the individual is currently dealing with its consequences. However, the reasoning suggests that the eviction process is still in progress and categorizes the situation as 'present,' which contradicts the ground truth. The reasoning fails to acknowledge that the eviction has been completed and misinterprets the current living situation as temporary rather than a result of completed eviction.