# SocialStep: Fast Prediction of Social Determinants of Health

**Paul Landes**[†], **Adam Cross**[†], **Jimeng Sun**[♠♢]
Department of Pediatrics, University of Illinois College of Medicine Peoria[†]
Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign[♠]
Carle Illinois College of Medicine, University of Illinois Urbana-Champaign[♢]
{plande2, arcross}@uic.edu, jimeng@illinois.edu

**Abstract**

Given thousands of medical documents, how can we automatically uncover patients' social risk factors? Social Determinants of Health (SDoH) constitute a growing class of non-clinical risk factors that shape patient trajectories. While clinically significant, automatic detection of SDoH from free text remains understudied due to scarce and imbalanced training data. Current approaches often rely on monolithic large language models. We present SocialStep, a two-step hybrid pipeline that first uses a lightweight classifier to triage sentences and then applies a Large Language Model (LLM) for multilabel classification to the relevant subset. On the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, SocialStep improves macro F1 by 5 points over the state-of-the-art baseline while running 12.2× faster. These findings demonstrate that integrating compact neural encoders with large language models provides a scalable and highly accurate framework for clinical NLP tasks, including SDoH extraction. Notably, we also observe some unexpected patterns in LLM performance. SocialStep offers a practical blueprint for hybrid model deployment that identifies critical social risk factors without prohibitive computational cost.

**Keywords:** social determinants of health, encoder-only transformers, large language models

## 1. Introduction

Social Determinants of Health (SDoH) greatly affect patient health, well-being, and quality of life. Safe housing, job opportunities, discrimination, and environmental factors are just a few examples. Adverse SDoH have immediate and profound influence on patient health. For example, poor air quality may negatively affect respiratory function, or lack of transportation may lead to reduced healthcare services. Identifying adverse SDoH has been shown to be helpful in prediction of diabetes mellitus (Hill-Briggs et al., 2021), recurring diabetic keto-acidosis (Lyerla et al., 2021), and prolonged hospital stays (Keenan et al., 2002).

Although some electronic health record systems have incorporated SDoH as structured data, many systems have not, or represent the data with highly varying formats (Li et al., 2024; Wang et al., 2021; Gold et al., 2018). To address these challenges, many have turned to methods to automatically extract SDoH from clinical text. The current state-of-the-art for this type of extraction incorporates Large Language Model (LLMs) in either a few-shot learning or supervised-fine tuning settings (Guevara et al., 2024; Lituiev et al., 2023).

We build on the work of Guevara et al. (2024) by training a variety of models on their publicly released datasets. Our experiments include encoder-only models (RoBERTa), instruction-tuned LLMs (Llama), and a hybrid of the two (Liu et al., 2019; Touvron et al., 2023). We experiment with feature combinations in our encoder-only model and our LLM models with few-shot and supervised fine-tuning settings. The hybrid models are shown to exploit the best of both worlds as shown in Figure 1. This work's contributions include:

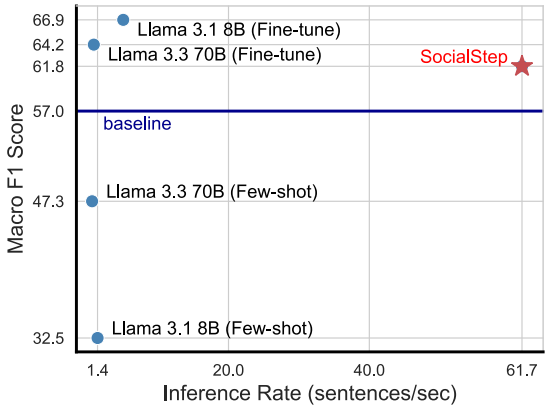


Figure 1: **Model performance to latency trade-off.** The fine-tuned LLMs perform well above the baseline, but are slow and expensive. SocialStep (our method) matches near LLM accuracy at 1/12 the cost (see Table 2).

1. Our method that leverages LLMs while striking a balance with cost for sparse datasets.
2. Our performance boosting findings on the effects using synthetic data.
3. Our supervised fine-tuned Llama embeddings and encoder-only models.
4. Our source code[1] to reproduce the results.

[1] https://github.com/sunlabuiuc/sdoh



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10802–10814
11–16 May 2026. 

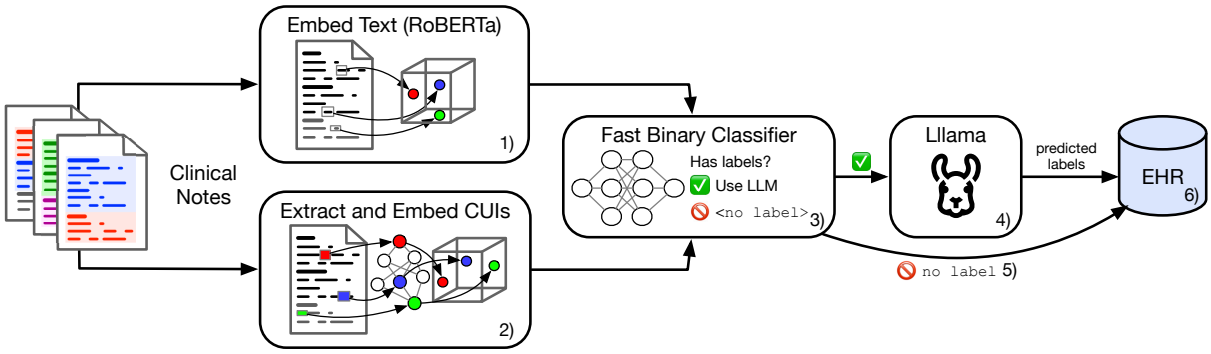


Figure 2: **Pipeline Overview.** The clinical text is embedded using RoBERTa (1). Medical concepts (concept unique identifiers (CUIs)) are extracted from the text and embedded (2). The embeddings are fine-tuned as the a fast binary classifier (3). The binary classifier uses Llama for multilabel prediction if SDoH labels were detected (4), otherwise it outputs `no label` to the electronic health record (5).

## 2. Related Work

The automatic extraction of SDoH from clinical text is a recent, but growing area of research with significant potential to inform healthcare interventions and improve patient outcomes. Although this task is relatively recent, several studies have examined it using diverse methodological approaches.

Segar et al. (2022) used random forests incorporating SDoH data (including zip code and demographic features) to enhance heart failure mortality prediction. They showed that traditional logistic regression models struggled to accurately predict in-hospital mortality for heart failure patients, often overlooking the crucial impact of SDoH.

In the work of Lituiev et al. (2023), BERT-based models were trained on a dataset of clinical notes annotated with SDoH entities for token level classification. Their results highlight the effectiveness of pre-trained language models in capturing complex linguistic patterns relevant to SDoH identification. Similarly, Lybarger et al. (2021) developed a system for SDoH event extraction from social history discharge summary sections. They utilized active learning to optimize data annotation and achieved promising results on their SHAC corpus.

Guevara et al. (2024) present a novel approach to sentence-level SDoH extraction by leveraging the power of LLMs. They fine-tune BERT-based and seq2seq models on a curated dataset of clinical notes annotated with SDoH information, demonstrating significant improvements. Notably, their work highlights the importance of domain adaptation for LLMs, emphasizing the need to train models specifically on healthcare data and the challenges of sparse SDoH markers in clinical text.

## 3. Data

We start with the datasets by Guevara et al. (2024), which, to our knowledge, is the only publicly available SDoH sentence-level benchmark dataset for multilabel classification. The first dataset is a subset of the Medical Information Mart for Intensive Care III (MIMIC-III) corpus (Johnson et al., 2016), a large freely accessible hospital database of ICU data from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. This dataset consists of 5,355 sentences taken from MIMIC-III clinical notes that were annotated for zero or more SDoH. A synthetic dataset generated from LLMs was also released, containing one or more SDoH, with 588 sentences annotated with at least one SDoH. The synthetic dataset was intended to provide a larger training set to boost results on the test dataset (MIMIC-III).

The Guevara et al. (2024) publicly available MIMIC-III and synthetic datasets were used for all experiments. We also combined these two datasets, which we call the Amended dataset. Each data point in the datasets is a sentence with associated SDoH labels. The labels apply to sentences rather than tokens.

| Dataset | Split | Count | Portion |
|---|---|---|---|
| MIMIC-III (annotated) | Test | 1K | 20% |
| | Train | 3K | 60% |
| | Validation | 1K | 20% |
| | Total | 5K | 100% |
| Synthetic (generated) | Test | 117 | 20% |
| | Train | 352 | 60% |
| | Validation | 119 | 20% |
| | Total | 588 | 100% |
| Amended (combined) | Test | 1K | 20% |
| | Train | 4K | 60% |
| | Validation | 1K | 20% |
| | Total | 6K | 100% |

Table 1: **Stratified splits of the datasets.** The count is the number of label occurrences across all sentences. The Amended dataset is the original MIMIC-III dataset combined with the LLM generated sentences by Guevara et al. (2024).

We split both the MIMIC-III and Amended datasets each using a multilabel iterative stratification (Sechidis et al., 2011) library[2] across SDoH classes. Table 1 gives the splits of the MIMIC-III, synthetic, and Amended datasets by sentence label.

# 4. Methods

We elaborated on the models by Guevara et al. (2024) using their datasets. We also added a new model (SocialStep) that integrates both the encoder-only binary model for efficiency and a LLM for precision (see Figure 2).

## 4.1. Model Development

We trained two types of models: a binary model that determines if a SDoH exists, and a multilabel that classified zero or more SDoH. These models included:

- Supervised-fine-tuned LLM: classifies zero or more SDoH labels
- Encoder-only binary: predicts whether a sentence has at least one SDoH
- SocialEnc: encoder-only model multilabel classifies zero or more SDoH labels (see Section 4.3.3)
- SocialStep: integrates the encoder-only binary classifier with a LLM

The encoder-only model and LLM models were trained and developed on the same datasets.

## 4.2. Large Language Models

The Llama 3.1 8B Instruct and Llama 3.3 70B Instruct (Touvron et al., 2023) models were used for all LLM experiments. We used the Guevara et al. (2024) annotation guide[3] to engineer our prompts for two settings: few-shot (see Appendix A) and supervised-fine-tuned (see Appendix B). The few-shot prompts included the entire definition of each SDoH category from the annotation guide. The few-shot prompts also included examples of each of the six categories with a simple explanation of the special label (`No SDoH`) for missing SDoH. During fine-tuning, each prompt included a one- or two-sentence synopsis from the annotation guide for the corresponding category. All fine-tuned LLMs were trained using LoRA (Hu et al., 2021) with a rank of 64, a learning rate of $5 \times 10^{-5}$, and dropout of 10% for three epochs.

Our experiments included feature prompt injection (Kanayama et al., 2024) in both few-shot and supervised fine-tuned settings using the token-level SDoH feature (Lituiev et al., 2023) and the concept unique identifier (CUI) preferred name (see Section 4.3.2). However, only the few-shot configuration exhibited a marginal improvement; accordingly, these results are not reported.

## 4.3. Encoder-only Models

All encoder-only models were trained for 40 epochs, but the model with the lowest validation loss was used for evaluation. The SocialEnc model was trained with a learning rate of $1 \times 10^{-5}$ and the binary classifier with a learning rate of $6.5 \times 10^{-6}$.

### 4.3.1. Feature Engineering

The encoder-only models incorporated several combinations of features. The RoBERTa (Liu et al., 2019) base model transformer was used and enhanced to accommodate token-level features (see Section 4.3.3). The token-level linguistic features were extracted from the text using spaCy (Montani et al., 2023) and biomedical entities were extracted with scispaCy (Neumann et al., 2019). The Lituiev et al. (2023) model was used to add the prediction as the SDoH token feature.

These features were concatenated to the transformer's final layer output for fine-tuning. Medical concepts extracted from input sentences were particularly influential features (see Section 4.3.2). Feature combinations for the encoder-only models were selected based on validation-set macro-F1 ablations, as summarized in Figure 5. A list of features and their descriptions are given below:

- part-of-speech (POS): The token's Penn Treebank (Marcus et al., 1993) part of speech such as `run` (VB).
- dependency depth (Dep): The depth of the token in the sentence's dependency head tree.
- named entity (Ent): The named entity such as `person` or `organization`.
- medical entity (MedEnt): A biomedical entity such as `amino acid` (Pyysalo et al., 2013).
- CUI embedding (CuiEmb): A clinical trained SBERT (Reimers and Gurevych, 2019) embedding of the CUI (see Section 4.3.2).
- SDoH token (TokSDoH): A token-level SDoH from the Lituiev et al. (2023) model.

### 4.3.2. Concept Features

The Unified Medical Language System (UMLS) is a large graph based medical taxonomy and terminology data source (Bodenreider, 2004). Each node in the UMLS graph represents a concept unique

[2] https://github.com/trent-b/iterative-stratification
[3] https://github.com/AIM-Harvard/SDoH/blob/main/...

identifier (CUI). Each CUI has many properties, two of which are the *preferred name* (a common name for the concept) and a *definition* of the concept, which can be exploited to provide more context to the model for each token.

We considered two methods of creating additional features using CUIs found in the clinical text. The first was simply to use embeddings generated from the CUI's preferred name and definition. We also considered adding cui2vec (Beam et al., 2020) embeddings, which are 500D vectors trained from clinical text using the word2vec algorithm (Mikolov et al., 2013a,b). However, cui2vec feature sparsity was a concern as its vocabulary is a subset of UMLS, which is then further conditioned by a less than perfect recall by the entity linker.

To make an informed decision, we computed a CUI and cui2vec in-vocabulary distributions from the MIMIC-III dataset. CUIs were extracted using The MedCAT (Kraljevic et al., 2021) entity linker and were then assigned a numerical count per sentence. 3,434 (65% of the dataset) sentences lack CUIs, and the cui2vec embeddings available for those that do exist, are even more scarce. However, given that 1,887 sentences (35% of the dataset) were found to have at least one CUI, we opted for a middle-ground solution to encode the CUI properties and leave the implementation of the cui2vec features for future work.

A fine-tuned SBERT model (Reimers and Gurevych, 2019) was used to encode and create CUI embeddings. We give the model a way to relate from medical concepts to SDoH semantically as SBERT models embed text in Euclidean space. Embedding the CUI with RoBERTa would add little or add redundancy in cases where the properties are close or identical to the text.

A clinically trained SBERT model (Deka et al., 2022) was used for the CUIs' embedding. Only the static embeddings from a forward inference were used due to memory constraints of fine-tuning two models (SBERT and RoBERTa) in parallel. Text in the form `<preferred name>:<definition>` was used as input to the SBERT model and repeated (stacked) the embeddings across all wordpieces (Wu et al., 2016) (token sub-units with associated vectors and provided by the model's tokenizer) tagged by the entity linker for each concept. Next, we explain how CUIs are vectorized and used with examples.

#### 4.3.3. SocialEnc Model

The SocialEnc and binary models used the same neural network architecture. Only the output layer differed: the SocialEnc model has one neuron for each SDoH label (6) and the binary model has a single output neuron.

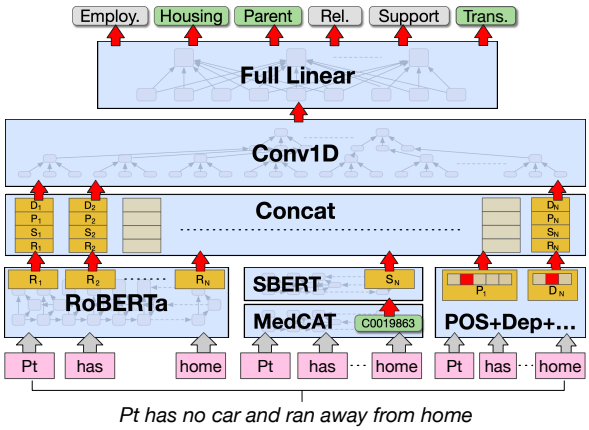


Figure 3: **Encoder-only Processing Example.** The clinical text input is encoded using RoBERTa Base embeddings, an entity linker and several linguistic taggers. Each of these components' output is concatenated, passed through a 1D convolutional neural network and decoded as zero or more labels.

The components in Figure 3 were build with DeepZensols (with biomedical extensions) and PyHealth (Landes et al., 2023; Wu et al., 2026):

- **RoBERTa (base) model's last layer**: The sentence was first tokenized and applied to the model.
- **CUI extraction**: First MedCAT links tokens to concepts. The CUI for "Homelessness" (`C0019863`) is linked to the tokens "ran away" using preferred name "Ran away, life event" (see Section 4.3.2). This text was then used as input to SBERT with frozen layers (no parameter updates).
- **One-hot encoded features**: Vectors were encoded from enumerated linguistic values, such as part-of-speech tags. One-hot encoded features were: `POS`, `Dep`, `Ent`, `MedEnt` and `TokSDoH`.

These components were used in parallel to enriched embeddings of wordpiece tokens using a single sentence as input for each inference. The components' output is then concatenated so that each wordpiece has the RoBERTa last layer embedding, SBERT last layer embedding, and the one-hot encode vectors. The concatenated tokens and features are then passed through a two layer 1D convolutional neural network. Finally, the last fully connected linear layer decodes the convolutional to SDoH label. A threshold for each neuron determines if the output is considered as present.

#### 4.3.4. SocialStep Model

The SocialStep model first uses the binary model to detect whether a sentence has one or more

| Dataset | Model | Size | F1 | Precision | Recall | Improve | Rate | Speed | Cost/Perf |
|---|---|---|---|---|---|---|---|---|---|
| MIMIC-III | Guevara et al. (2024) | 11B | 57.0 | - | - | - | - | | - |
| | Llama 3.1 | 8B | 66.9 | 65.0 | 77.9 | 10% | 5.1 | | 7.6 |
| | Llama 3.3 | 70B | 64.2 | 70.5 | 71.6 | 7% | 0.9 | | 1.4 |
| | SocialEnc | 110M | 52.9 | 60.1 | 48.8 | -4% | 371.0 | 73.5× | 700.7 |
| | SocialStep | 8.11B | 61.8 | 76.1 | 59.5 | 5% | 61.7 | 12.2× | 99.7 |
| Amended | Guevara et al. (2024) | 11B | 55.0 | - | - | -2% | - | | - |
| | Llama 3.1 | 8B | 77.9 | 86.3 | 73.9 | 21% | 1.3 | | 1.6 |
| | Llama 3.3 | 70B | 86.0 | 82.4 | 90.2 | 29% | 0.3 | | 0.4 |
| | SocialEnc | 110M | 88.1 | 91.0 | 85.9 | 31% | 203.1 | 158.1× | 230.6 |
| | SocialStep | 8.11B | 87.8 | 87.5 | 88.8 | 31% | 1.4 | 1.1× | 1.6 |

Table 2: **Fine-tuned LoRA Average.** Macro-average performance metrics of our multilabel models on the MIMIC-III (top) and Amended (bottom) datasets with the Guevara et al. (2024) baselines as the first row for each sub table. The (**Improve**)ment over the baseline, prediction rate in sentences per second, the (**Speed**)up and **Cost** to (**Perf**)ormance on the test set given in the last columns.

SDoH. For those that do, it then utilizes a larger more costly model with more precision for the SocialEnc classification[4] such as a LLM. Our results report the performance of the encoder-only binary model with the Llama 3.1 8B Instruct on the MIMIC-III dataset. We believe the results of the SocialStep on the Amended dataset would be higher, but leave this as a future work.

## 5. Experimental Setup

The scikit-learn[5] multilabel and metrics libraries were used to compute all metrics. The model evaluation metrics included weighted, micro and macro average performance metrics on the multilabel iterative stratified splits (Sechidis et al., 2011).

Our LLM experiments included two settings: few-shot and LoRA fine-tuning. These experiments are intentionally separated because they target different notions of model capability. The few-shot setting measures model adaptability under minimal supervision, whereas the fine-tuning setting measures performance after task-specific supervised adaptation. Presenting these separately avoids overstating direct comparability and makes clear whether gains arise from pretrained generalization or from parameter-efficient fine-tuning on labeled examples.

### 5.1. Few-shot Evaluation

Only the LLMs included experimentation in the few-shot setting using the publicly available checkpoints. The Guevara et al. (2024) MIMIC-III annotated dataset (all three splits shown in Table 1) was used to evaluate the few-shot models using a temperature of 0.1.

### 5.2. LoRA Fine-tuning Evaluation

For supervised experiments, we used the dataset splits shown in Table 1. The splits were used to evaluate the encoder-only, Llama 3.1 8B Instruct and Llama 3.3 70B Instruct fine-tuned models. The evaluation was done on the MIMIC-III and Amended datasets in a similar fashion to the Guevara et al. (2024) experiments. A 10-fold cross validation with 5 repeats was also used for evaluation on the encoder-only models (cross-validation on the LLMs was prohibitively expensive). The Amended SocialEnc classifier was also cross-validated with the same parameters.

## 6. Results

Our LLM fine-tuned performance metrics are based on a 20% held-out test set across the MIMIC-III and Amended datasets whereas Guevara et al. (2024) use the annotated MIMIC-III data as a test set. We compare our results to Guevara et al. (2024)[6] but only as a reference point and not as test-only datasets. We show that our SocialStep classifier is 4.8 macro F1 points more performant than the reference point baseline and 12.2 times faster than the smallest LLM (see Table 2). Furthermore, our models show improvement with the synthetic data added to the MIMIC-III data (we call this the Amended dataset) compared to their results, which show a decrease in performance.

Our SocialEnc model performed inference up to 73.5 times faster than our slowest LLM as shown in (see Table 2). Based on this observation, we adopted a new SocialStep model that performs sentence-level SDoH detection, followed by Llama-based multilabel classification for positive sentences.

[4]We joined the prediction data to measure performance and latency to simulate the classifier, but the implementation would be trivial.

[5]https://scikit-learn.org

[6]We cannot compare directly as Guevara et al. (2024) did not release their model, training dataset or prompts. In their work, the MIMIC-III and synthetic datasets were used for testing only.

| Model | Size | No SDoH (n=992) | Employ (n=13) | House (n=1) | Parent (n=5) | Relation (n=36) | Support (n=23) | Transport (n=1) |
|---|---|---|---|---|---|---|---|---|
| Guevara et al. (2024) | 11B | 98.0 | 65.0 | 0.0 | 63.0 | 91.0 | 32.0 | 50.0 |
| Llama 3.1 | 8B | 97.8 | **66.7** | **100.0** | 53.3 | 86.4 | 13.8 | 50.0 |
| Llama 3.3 | 70B | 98.0 | 35.3 | 40.0 | 80.0 | **88.3** | 7.7 | **100.0** |

Table 3: **Held-out Test-set Fine-tuned LoRA.** Per-label one-vs-rest F1 multilabel classification task across models on the MIMIC-III dataset. Each label was evaluated as a one-vs-rest binary comparison against the Guevara et al. (2024) baseline. Abbreviated prediction labels are (**Employ**)ment, (**Relation**)ship, and (**Transport**)ation. The highest per-label scores are **bolded** and header counts $n$ are the gold-positive test occurrences in the test set. Because some labels are rare (e.g., $n = 1$), these per-label estimates are unstable (see Section 6.1).

### 6.1. Metric Stability for Low-Support Labels

Table 3 reports per-label one-vs-rest metrics, which vary substantially for some labels. This is due to very low support[7] for minority labels in the test set, which is a result of the highly skewed Guevara et al. (2024) annotated dataset. The dataset was stratified across each sentence's multilabel annotation, so at least one example of each label was given in the training set. However, because each label was scored independently as a binary prediction task, false-positives can greatly change the resulting F1 score.

For example, the `housing` label had only one gold-positive test instance. This instance was correctly identified as a true positive, but the classifier also produced three additional false-positive `housing` predictions. This yields a precision of $1/4$, recall of $1/1$, and an F1 score of 40%. More generally, rare-label results for the few-shot LLMs were similarly sensitive to small numbers of false-positive predictions, and should therefore be interpreted cautiously.

The low test-set support that leads to instability is shown in Table 3. The few-shot classifier results in Table 4 show the same pattern as one-vs-rest sensitivity. In fact, an even larger shift is observed due to higher false-positive predictions on the larger dataset. Scores are substantially lower than those of other classifiers, which suggests that few-shot learning for SDoH prediction is lacking compared to the encoder-only model and fine-tuned LLMs.

### 6.2. SocialEnc Model

Table 2 gives the performance metric averages on the MIMIC-III and Amended datasets. The highest performing model on the MIMIC-III dataset by macro F1 (66.9) is the Llama 3.1 Instruct 8B model. However, the SocialEnc model shows the best F1 on the Amended dataset. The encoder-only model trails 13.9 points behind the Llama 3.1 8B Instruct model and 4.1 points behind the Guevara et al. (2024) baseline reference (57).

[7]Here, support denotes the number of gold-positive instances for a given label in relevant evaluation set.

In contrast, on the Amended dataset, the SocialEnc model achieves the highest macro-average F1 among all models. It is also the fastest model on the dataset with a speedup of 158.1× and the best cost-to-performance (sentences per second divided by F1) of 230.6. Figure 4 shows the classifier improves over every model on almost every label and comes close on all labels to the best.

Table 5 shows the results of the 10-fold cross-validation of the SocialEnc classifier on the MIMIC-III dataset. The weighted F1 score (99.8) is stable, but the macro F1 score (76.3) is relatively low. This statistic is illuminating as the classifier performs better by random variation on the test split (88.1 in Table 2).

### 6.3. SocialStep Model

The SocialStep classifier performs within a weighted F1 point of the best model (RoBERTa) and within 5 points of the best macro F1 score (Llama 3.1 8B Instruct) as shown in Table 2. The classifier uses the encoder-only model to detect

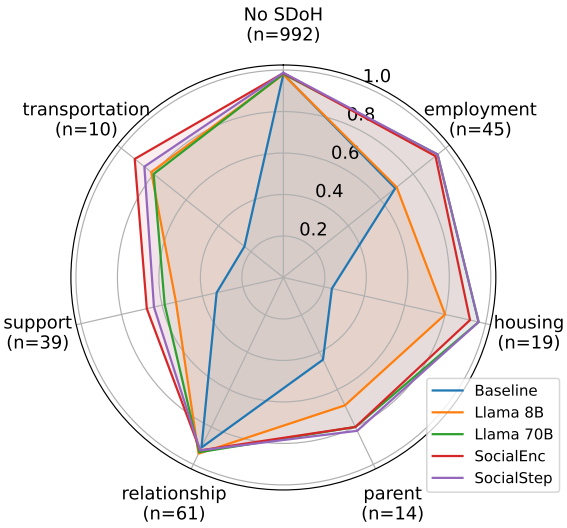


Figure 4: **Held-out Test-set Fine-tuned LoRA.** Results by label for fine-tuned models on the Amended dataset against the Guevara et al. (2024) baseline.

| Dataset | Model | Size | No SDoH | Employ | Hous | Parent | Relat | Supp | Tran | Macro-Ave |
|---|---|---|---|---|---|---|---|---|---|---|
| MIMIC-III | | | (n=4961) | (n=65) | (n=3) | (n=27) | (n=180) | (n=116) | (n=3) | |
| | Llama 3.1 | 8B | 87.6 | 41.0 | 0.8 | 26.3 | 45.8 | 23.3 | 2.6 | 32.5 |
| | Llama 3.3 | 70B | 95.1 | 76.5 | 5.8 | 33.3 | 76.9 | 24.1 | 19.0 | 47.3 |
| Amended | | | (n=4961) | (n=201) | (n=72) | (n=94) | (n=332) | (n=219) | (n=64) | |
| | Llama 3.1 | 8B | 87.0 | 58.1 | 14.3 | 56.6 | 62.1 | 37.9 | 52.8 | 52.7 |
| | Llama 3.3 | 70B | 94.8 | 83.4 | 49.3 | 64.2 | 84.8 | 36.2 | 76.1 | 69.8 |

Table 4: **Full-dataset Descriptive Few-shot.** Per-label One-vs-rest F1 multilabel classification task on LLMs evaluated as a one-vs-rest binary. These results are measured across the full MIMIC-III and Amended datasets and are descriptive and not directly comparable to held-out fine-tuned results in Table 3. Predicted labels are (**Employ**)ment, (**Hou**)sing, Parent, (**Relat**)ionship, (**Supp**)ort, and (**Tran**)sportation. The header counts $n$ are the gold-positive occurrences in the full dataset. Because some labels are rare (e.g., $n = 1$), these per-label estimates are unstable (see Section 6.1).

the negative label (`No SDoH`), but its recall is substantially lower than its precision. This shows that it struggles with false negatives at the binary level and also reflected in the comparatively low macro recall from the cross-validated results in Table 5. Because the model uses a gated ensemble design, missed positives at the binary stage upper-bound the classifier's end-to-end multilabel performance.

## 6.4. Error Analysis

Our SocialEnc classifier failed for labels `Housing` and `Transportation` on the MIMIC-III dataset, which is not surprising given these labels have only three occurrences. The SocialStep classifier gets the single `Housing` instance correct, which means the binary encoder-only model learned that it was SDoH positive and the LLM correctly classified it, but still failed on the `Transportation` label. However, all of our LLMs achieved a non-zero score on all labels and outperformed the Guevara et al. (2024) reference models for most labels. Furthermore, our models perform better on every label on the Amended dataset.

The method of parsing of the LLM output might negatively affect performance. A complex regular expression was needed to parse the noisy LLM output. Models hallucinated in the few-shot setting (output on the MIMIC-III dataset was more noisy than on the Amended), but were more consistent on the fine-tuned models. Of course, this was not an issue for the encoder-only model as its output layer directly predicted each label. However, coupling traditional models with LLMs can have consequences, such as with the SocialStep classifier.

The SocialStep classifier uses the binary classifier to predict if a SDoH is present in a sentence. When it predicts the presence of one or more SDoH, it uses the Llama 3.1 8B Instruct model to assign labels. The binary classifier has a lower recall than precision, but the Llama 3.1 8B Instruct model's recall is significantly higher as shown in Table 2. This could be propagation error from false negative SDoH classifications. However, Table 5 shows a very similar macro precision and recall, so the LLM might not assign any labels since the classifier was trained with `No SDoH` as a label. The SocialStep classifier may perform better using a LLM trained without negative SDoH labels.

| | SocialEnc | | | | Binary | | | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **Min** | **Max** | $\mu$ | $\sigma$ | **Min** | **Max** | $\mu$ | $\sigma$ |
| **wF1** | 94.6 | 100 | 99.8 | 0.8 | 96.2 | 100 | 99.9 | 0.5 |
| **wP** | 95.4 | 100 | 99.8 | 0.7 | 96.2 | 100 | 99.9 | 0.5 |
| **wR** | 94.9 | 100 | 99.8 | 0.7 | 96.2 | 100 | 99.9 | 0.5 |
| **mF1** | 95.6 | 100 | 99.8 | 0.6 | 96.2 | 100 | 99.9 | 0.5 |
| **mP** | 96.3 | 100 | 99.9 | 0.5 | 96.2 | 100 | 99.9 | 0.5 |
| **mR** | 94.9 | 100 | 99.8 | 0.7 | 96.2 | 100 | 99.9 | 0.5 |
| **MF1** | 45.6 | 100 | 76.3 | 9.7 | 85.1 | 100 | 99.7 | 2.1 |
| **MP** | 62.2 | 100 | 76.9 | 9.1 | 85.1 | 100 | 99.7 | 2.1 |
| **MR** | 40.2 | 100 | 76.0 | 10.1 | 85.1 | 100 | 99.7 | 2.1 |
| **acc** | 98.7 | 100 | 100.0 | 0.2 | 96.2 | 100 | 99.9 | 0.5 |

Table 5: SocialEnc and binary classifiers cross-validation results on the MIMIC-III dataset. The 10-Fold validation (5 repeats) metrics includes (**w**)eighted, (**m**)icro and (**M**)acro metric averages.

## 6.5. Ablation Studies

Our ablation studies include feature combinations on the encoder-only binary classifier. Each feature combination is a model that learns jointly with the RoBERTa embeddings (see Section 4.3.3). Figure 5 shows the ablation of the features as macro average F1 performance over epochs of training the models on the validation set. The test set performance for each feature combination is displayed with triangles and reported in parentheses in the legend.

We see very high variance of the SocialEnc feature combinations' test set scores across model type. Part-of-speech tags, head dependency features, and named entities were the most useful on the MIMIC-III dataset, but the Lituiev et al. (2023) token-level SDoH feature was the most helpful for the Amended dataset.

The binary models on the Amended dataset are much less sensitive to the choice of feature set,

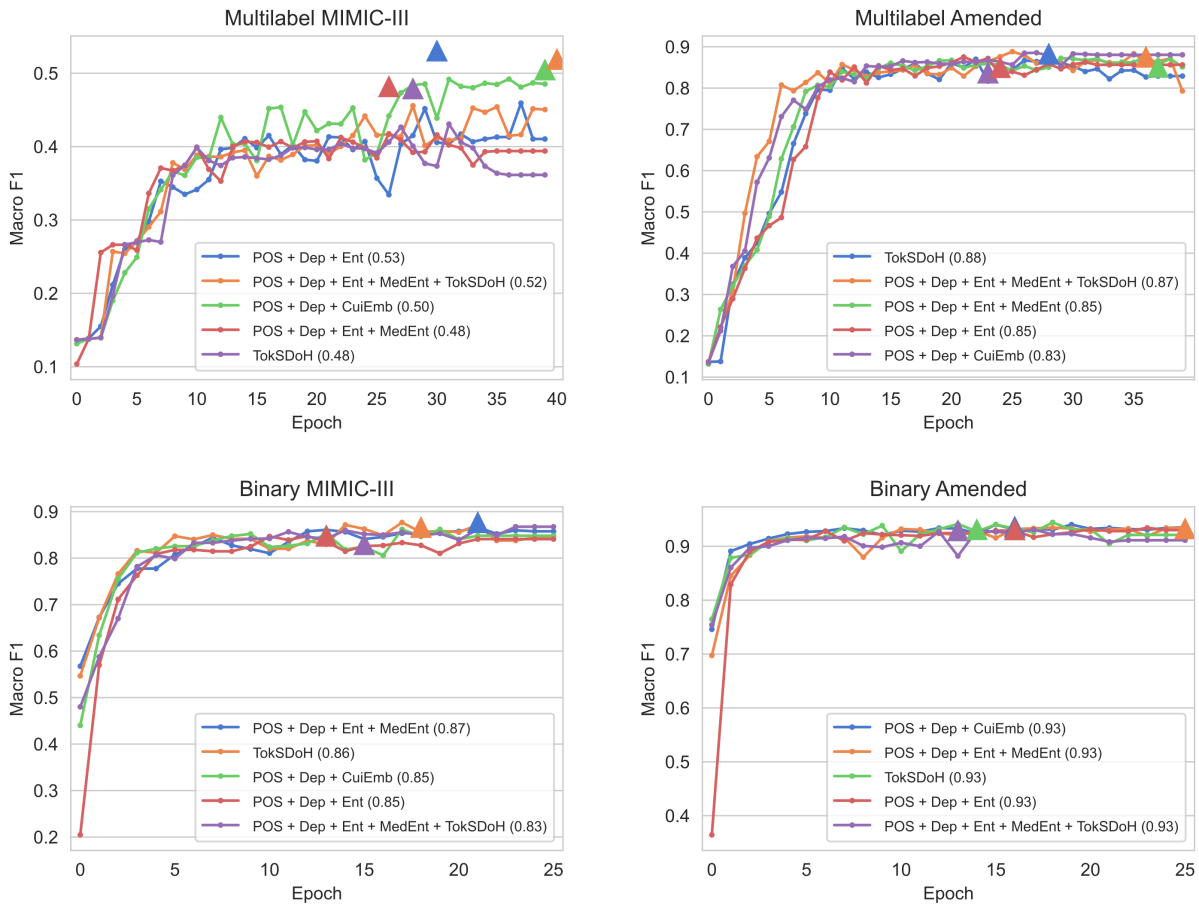


Figure 5: The multilabel and binary model feature ablations by macro F1 performance are shown for the MIMIC-III and Amended validation sets. The triangles represent the best performing macro F1 on the training set (also listed in the legend) and converged epochs on the X-axis. Features are (**P**)art-(**O**)of-(**S**)each tag, (**Dep**)endency head tree depth, named (**Ent**)ity, (**Med**)ical named (**Ent**)ity, (**Cui Emb**)bedding and (**Tok**)en-level(**SDoH**).

implying the RoBERTa embeddings are leveraged to take advantage of the added synthetic data. Adding features appears to negatively affect models on the MIMIC-III dataset for some combinations. The binary model illustrates how adding features may lead to worse performance. The feature combination that includes all features performs five points lower on the MIMIC-III dataset compared to the Amended dataset.

## 7. Discussion

Table 2 gives the performance of the fine-tuned Llama 3.1 8B Instruct classifier, and shows a significant improvement over the Guevara et al. (2024) baseline on the MIMIC-III dataset. The LLM models perform better on the Amended dataset over the MIMIC-III dataset, which demonstrates that the models learn from the synthetic data. The MIMIC-III dataset's leader model was the Llama 3.1 8B Instruct, but the encoder-only models matched or exceeded the LLMs on the Amended dataset.

The synthetic data provided in the Amended dataset led to models that outperform the LLMs. We believe better label-balanced datasets was a major contributor to the encoder-only models improvements (only 10% of the MIMIC-III dataset has SDoH labels). Furthermore, minority labels with extremely low support should be regarded as tentative, and the per-label results for these categories are best interpreted as descriptive rather than definitive (see Section 6.1).

The SocialEnc model was trained using the same features as the binary model. While the SocialEnc classifier performs well overall, it manages very poorly on the minority labels. The classifier under performed on our LLMs by 14 macro F1 points, but outperformed all LLMs in the weighted F1. Even though the model has difficulty with the label imbalance given it fails in predicting the `Housing` and `Transportation` labels as shown in Table 3 its results pale in comparison to the fine-tuned LLMs that have a non-zero performance with all labels.

We observed the encoder-only model's ability at detecting SDoH with a higher macro recall (49)

than precision (60). However, the converse is true with all LLMs on the MIMIC-III dataset (this dataset includes sentences with no SDoH). These observations motivate a binary classifier that predicts whether a sentence contains any SDoH. The strong weighted F1 and recall further motivated the SocialStep model.

### 7.1. Inference Latency

The models differ greatly in inference latencies, particularly between the encoder-only models and LLMs. Pipeline processing bottlenecks arise as the latency of a classifier grows with the input size, such as with longitudinal clinical notes.

Our binary classifier performs as well as the best Guevara et al. (2024) model on the MIMIC-III dataset. However, it performs inference in a fraction of the time of the LLMs as the model is a fraction of the size. The encoder-only classifier is able to predict up to 371 sentences per second on average compared to the fastest LLM that predicts 5.1 sentences per second (see Table 2). The SocialStep classifier is not far behind with speedup of 12.2×, which translates to a prediction rate of 61.7 sentences per second.

### 7.2. SocialStep and Binary Classifiers

The binary classifier achieved a F1 of 1 on our test split of the MIMIC-III dataset. The stratified dataset contains 992 negative SDoH labels and 73 positive labels. This dataset imbalance explains the lower F1 score of 76.7 on the positive labels. However, it is much faster compared to any LLM fine-tuned model.

Our experiments show that the binary classifier used as the first component in the SocialStep classifier is relatively close in performance to the fine-tuned LLMs. The SocialStep classifier yields a macro F1 of 61.8, which is only two macro F1 points lower than the Llama 3.3 70B Instruct model (see Table 2). Considering the SocialStep classifier is 12.2 times faster than the Llama 3.1 8B Instruct classifier, we believe it shows the best cost-to-performance trade-off for real-world clinical application.

## 8. Conclusion

Our study compares the advantages and disadvantages of encoder-only models with state-of-the-art LLMs. We found that customized clinical models are significantly more efficient. Our custom deep learning binary model predicted up to 12.2 times faster, and our SocialEnc model predicted 158.1 times faster than comparative LLMs. This is an important finding given that only 10% of clinical text have any SDoH markers.

Surprisingly, our encoder-only models outperformed all LLMs on SDoH labeled synthetic data generated by LLMs. The SocialEnc model was also much more cost effective and nimble, particularly compared to the Llama 3.3 70B Instruct model. Our takeaway is that there is no one-size-fits-all for a model given constraints and datasets. When given a small training dataset LLMs will perform better but are costly. Larger training datasets with domain specific features (such as CUIs and linguistic features) yield both performant and cost effective models.

## 9. Ethics Statement

We believe that the study and automatic detection of SDoH will not only aid in diagnosing illness but also promote greater health equity for underrepresented populations. This work benefits not only clinicians but also public health advocates, social workers, citizen scientists, and patients. By improving our understanding of the social root causes underlying health issues, physicians and social workers can more effectively and equitably address the needs of underserved communities. Automated SDoH detection should be deployed with safeguards, human oversight and caution, as they can be biased or context-dependent.

## 10. Limitations

Our experiments utilized the datasets provided by Guevara et al. (2024). Our models may perform differently based on other datasets or clinical text. Furthermore, we cannot compare our results directly to the Guevara et al. (2024) baseline as they did not release their model, training dataset or prompts. In their work, the MIMIC-III and synthetic datasets were used for testing only and their models were trained on datasets that are not publicly available.

Our latency measurements reflect the evaluated sentence-level inference setup and do not include deployment-specific optimizations. Our models and findings may not transfer to other hospital systems. Experiments used the Guevara et al. (2024) dataset as a benchmark, where each sentence has one or more SDoH labels. This enables comparison of our findings with those from their earlier work. However, some SDoH evidence may require paragraph- or note-level context, which we leave for future work.

## 11. Acknowledgements

This research was conducted with support from the AI.Health4All Center at the University of Illinois at Chicago College of Medicine.

## 12. Bibliographical References


Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:295–306.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.

Pritam Deka, Anna Jurek-Loughrey, and Deepak Padmanabhan. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–505.

Rachel Gold, Arwen Bunce, Stuart Cowburn, Katie Dambrun, Marla Dearing, Mary Middendorf, Ned Mossman, Celine Hollombe, Peter Mahr, Gerardo Melgar, James Davis, Laura Gottlieb, and Erika Cottrell. 2018. Adoption of Social Determinants of Health EHR Tools by Community Health Centers. *The Annals of Family Medicine*, 16(5):399–407.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L. Chaunzwa, Idalid Franco, Benjamin H. Kann, Shalini Moningi, Jack M. Qian, Madeleine Goldstein, Susan Harper, Hugo J. W. L. Aerts, Paul J. Catalano, Guergana K. Savova, Raymond H. Mak, and Danielle S. Bitterman. 2024. Large language models to identify social determinants of health in electronic health records. *npj Digital Medicine*, 7(1):1–14.

Felicia Hill-Briggs, Nancy E. Adler, Seth A. Berkowitz, Marshall H. Chin, Tiffany L. Gary-Webb, Ana Navas-Acien, Pamela L. Thornton, and Debra Haire-Joshu. 2021. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*, 44(1):258–279.

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. In *International Conference on Learning Representations*.

Alistair E. W. 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.

Hiroshi Kanayama, Yang Zhao, Ran Iwamoto, and Takuya Ohko. 2024. Incorporating Syntax and Lexical Knowledge to Multilingual Sentiment Classification on Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4810–4817, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Heather Keenan, Carol Foster, and Susan Bratton. 2002. Social factors associated with prolonged hospitalization among diabetic children. *Pediatrics*, 109(1).

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J. B. Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083.

Paul Landes, Barbara Di Eugenio, and Cornelia Caragea. 2023. DeepZensols: A Deep Learning Natural Language Processing Framework for Experimentation and Reproducibility. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 141–146, Singapore, Singapore. Association for Computational Linguistics.

Chenyu Li, Danielle L. Mowery, Xiaomeng Ma, Rui Yang, Ugurcan Vurgun, Sy Hwang, Hayoung K. Donnelly, Harsh Bandhey, Yalini Senathirajah, Shyam Visweswaran, Eugene M. Sadhu, Zohaib Akhtar, Emily Getzen, Philip J. Freda, Qi Long, and Michael J. Becich. 2024. Realizing the potential of social determinants data in EHR systems: A scoping review of approaches for screening, linkage, extraction, analysis, and interventions. *Journal of Clinical and Translational Science*, 8(1):e147.

Dmytro S. Lituiev, Benjamin Lacar, Sang Pak, Peter L. Abramowitsch, Emilia H. De Marchis, and Thomas A. Peterson. 2023. Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients. *Journal of the American Medical Informatics Association : JAMIA*, 30(8):1438.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631.

Ryan Lyerla, Brianna Johnson-Rabbett, Almoutaz Shakally, Rekha Magar, Hind Alameddine, and Lisa Fish. 2021. Recurrent DKA results in high societal costs - a retrospective study identifying social predictors of recurrence for potential future intervention. *Clinical Diabetes and Endocrinology*, 7(1):13.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. Explosion/spaCy: V3.7.2: Fixes for APIs and requirements. Zenodo.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, and Sophia Ananiadou. 2013. Overview of the Cancer Genetics (CG) task of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 58–66, Sofia, Bulgaria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the Stratification of Multi-label Data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6913, pages 145–158. Springer Berlin Heidelberg, Berlin, Heidelberg.

Matthew W. Segar, Jennifer L. Hall, Pardeep S. Jhund, Tiffany M. Powell-Wiley, Alanna A. Morris, David Kao, Gregg C. Fonarow, Rosalba Hernandez, Nasrien E. Ibrahim, Christine Rutan, Ann Marie Navar, Laura M. Stevens, and Ambarish Pandey. 2022. Machine Learning–Based Models Incorporating Social Determinants of Health vs Traditional Models for Predicting In-Hospital Mortality in Patients With Heart Failure. *JAMA Cardiology*, 7(8):844–854.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Michael Wang, Matthew S Pantell, Laura M Gottlieb, and Julia Adler-Milstein. 2021. Documentation and review of social determinants of health data in the EHR: Measures and associated insights. *Journal of the American Medical Informatics Association*, 28(12):2608–2616.

John Wu, Yongda Fan, Zhenbang Wu, Paul Landes, Eric Schrock, Sayeed Sajjad Razin, Arjun Chatterjee, Naveen Baskaran, Joshua Steier, Andrea Fitzpatrick, Bilal Arif, Rian Atri, Jathurshan Pradeepkumar, Siddhartha Laghuvarapu, Junyi Gao, Adam R. Cross, and Jimeng Sun. 2026. PyHealth 2.0: A Comprehensive Open-Source Toolkit for Accessible and Reproducible Clinical Deep Learning.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*.

## A. Few-shot Prompt

```
Classify sentences for social determinants of health (SDOH). Definitions SDOHs are
given in the below list:

*'housing': The status of a patient's housing is a critical SDOH, known to affect
the outcome of treatment. For the purposes of this annotation task, a sentence will
be annotated as housing if it expresses a challenge relating to the place of
residence of the patient. Please note that references to cities and towns, without
mention of specific housing should NOT be considered an SDOH annotation. Attributes
are Poor, Undomiciled, Other.

*'transportation': This SDOH pertains to a patient's inability to get to/from their
healthcare visits.. A patient being present at the treatment location, even if
explicitly textually represented, or discussions of transportation unrelated to
adequacy of transportation access, should NOT be considered an instance of
Transportation SDOH. However, if there is a case of explicit textual representation
that a patient is absent for treatment and that absence is due to transportation
issues, then this IS considered an instance of Transportation SDOH. Attributes are
Distance, Resource, Other.

*'relationship': Whether or not a patient is in a partnered relationship is an
abundant SDOH in the clinical notes. A sentence represents relationship status if
it expresses evidence that a patient is married, in a partnership,
divorced/separated, single, or widowed. Attributes are Married, Partnered,
Divorced, Widowed, Single.

*'parent': This SDOH should be used for descriptions of a patient being a parent to
at least one child who is a minor (under the age of 18 years old). Tthe only
evidence necessary for this SDOH is the existence of a patient's child under the
age of 18. For the purposes of this task, "teenage children" can be considered
minors. This SDOH category is binary and has no attributes, so the full annotation
will just be the SDOH.

*'employment': This SDOH pertains to expressions of a patient's employment status.
A sentence should be annotated as an Employment Status SDOH if it expresses if the
patient is employed (a paid job), unemployed, retired, or a current student.
Attributes are Employed, Unemployed, Under-Employed, Disability, Retired, Student.

*'support': This SDOH is a sentence describes a patient that is actively receiving
care support, such as emotional, health, financial support. This support comes from
family and friends but not health care professionals. The sentence must describe an
act of care, participation in the patient's care, or an explicit statement that the
person in the patient's life is "supportive", "caring for them", etc. In these
cases, we wish to capture a patient's Social Support with this annotation.

Here are some examples of "Sentence" input and "SDOH labels" you output:

### Sentence:Pt lives in Arlington.
### SDOH labels:'''housing'''

### Sentence:Pt lives 30mi away from hospital and and complains about needing to
transfer three times each way.
### SDOH labels:'''transportation'''

### Sentence:Pt and her husband came into my office today.
### SDOH labels:'''relationship'''

### Sentence:Pt has 2 children ages 9 and 13.
### SDOH labels:'''parent'''

### Sentence:Pt works as an electrician in Rockland.
### SDOH labels:'''employment'''
```

```
### Sentence:Here today is Pt, her daughter, and supportive wife
### SDOH labels:'''support'''

Now classify the sentence with a comma-separated list of labels that are mostly
likely to be present. Only output the labels (or '''-''' for no SDOH found)
surrounded by three back ticks.

### Sentence:{{ text }}
### SDOH labels:
```

Figure 6: **Few-shot prompt.** Our prompt used for SDoH prediction with definitions and examples take fromthe Guevara et al. annotation guide.

## B. Training Prompt

```
Classify sentences for social determinants of health (SDOH).

Definitions SDOHs are given with labels in back ticks:

*'housing': The status of a patient's housing is a critical SDOH, known to affect
the outcome of treatment.

*'transportation': This SDOH pertains to a patient's inability to get to/from their
healthcare visits.

*'relationship': Whether or not a patient is in a partnered relationship is an
abundant SDOH in the clinical notes.

*'parent': This SDOH should be used for descriptions of a patient being a parent to
at least one child who is a minor (under the age of 18 years old).

*'employment': This SDOH pertains to expressions of a patient's employment status.
A sentence should be annotated as an Employment Status SDOH if it expresses if the
patient is employed (a paid job), unemployed, retired, or a current student.

*'support': This SDOH is a sentence describes a patient that is actively receiving
care support, such as emotional, health, financial support. This support comes from
family and friends but not health care professionals.

*'-': If no SDOH is found.

Now classify sentences for social determinants of health (SDOH) as a list labels in
three back ticks. The sentence can be a member of multiple classes so output the
labels that are mostly likely to be present.

### Sentence: {{ text }}
### SDOH labels: '''{{ labels }}'''
```

Figure 7: **Training prompt.** Our prompt used for supervised fine-tuned training of SDoH predictionwith examples take from the Guevara et al. annotation guide.