LLMs to Extract SDoH from Clinical Notes

Siddarath Vats

Abstract

Social Determinants of Health (SDOH)—such as housing instability, transportation barriers, employment status, and social support—are major drivers of health outcomes but are often buried in unstructured clinical notes. Our project proposes a prompt engineering approach using large language models (LLMs) to extract SDOH directly from clinical text, without the need for model training or fine-tuning. By designing structured prompts grounded in expert-defined annotation guidelines, we enable zero-shot SDOH classification across multiple categories. We applied our method to notes derived from the MIMIC-III dataset and evaluated consistency against manually annotated data. The results show that carefully crafted prompts can effectively extract relevant SDOH attributes and offer a scalable, interpretable, and low-resource solution for integrating social risk data into clinical workflows. This research highlights the potential of prompt-based NLP to surface actionable social factors in healthcare decision-making.

1 Introduction

Social Determinants of Health (SDOH) refer to the non-medical factors that influence health outcomes, including housing, transportation access, employment status, relationship status, and social support. These determinants play a critical role in shaping patient health, yet are rarely documented in structured fields of electronic health records (EHRs). Instead, they are often embedded within lengthy, unstructured clinical narratives—making them difficult to analyze at scale.

Traditional approaches for extracting SDOH from clinical notes rely on supervised learning models, which require large annotated datasets and considerable computational resources for training. Such methods are often costly to develop, difficult to generalize across institutions, and opaque in interpretability.

Recent advances in large language models (LLMs) like GPT-4 and DeepSeek-R1 offer a new paradigm: instead of training models from scratch, one can engineer prompts that guide pre-trained LLMs to perform specific tasks in a zero-shot or few-shot manner. Prompt-based learning reduces the dependence on labeled data and offers a more flexible, interpretable solution for domain-specific information extraction.

However, orchestrating multiple specialized LLM agents, managing prompt pipelines, and maintaining state across complex annotation tasks introduces significant engineering overhead. LangGraph, a framework for building stateful, multi-agent LLM applications, addresses these challenges by enabling modular, event-driven workflows where agents can communicate, share memory, and conditionally process data. LangGraph provides built-in support for persistence, namespaced memory, and streaming control flow—making it ideal for healthcare NLP pipelines that must be both secure and scalable.

In this paper, we introduce a LangGraph-powered prompt engineering approach for extracting SDOH from unstructured clinical text without any fine-tuning. Our method encodes domain-specific annotation guidelines into carefully designed prompts and leverages open-source LLMs served via local inference (e.g., Ollama) to annotate clinical notes from the MIMIC-III dataset. We implement a multi-agent system where each LangGraph node specializes in detecting specific SDOH categories (e.g., DeepSeek-R1 for employment and relationship status, LLaMA3 for parental status), and integrate the results into a structured format suitable for downstream analysis in a TimescaleDB.

This architecture not only demonstrates the feasibility of using lightweight, locally hosted LLMs for high-quality clinical information extraction, but also highlights LangGraph as a powerful backbone

for orchestrating secure, interpretable, and maintainable AI pipelines in real-world healthcare environments.

2 Related Work

A growing body of research has investigated methods for extracting social determinants of health (SDoH) from clinical texts, spanning from traditional statistical techniques to modern deep learning approaches. Several studies have focused on leveraging structured biomedical knowledge bases to guide SDoH extraction. (Martin et al., 2024), for example, employed a rule-based approach using terminologies like SNOMED CT, MeSH, and LOINC to identify indicator keywords and phrases within clinical notes via regular expressions (Martin et al., 2024). Similarly, (Yu et al., 2024) developed a keyword lexicon using a snowball strategy to identify notes that potentially contain SDoH information, illustrating the continued reliance on keyword-based retrieval in earlier methods (Yu et al., 2024).

The 2022 n2c2 NLP challenge marked a significant shift in methodology by fostering the adoption of transformer-based large language models (LLMs) for SDoH extraction tasks (Lybarger et al., 2023). This competition catalyzed the use of deep learning methods, especially pre-trained models such as BERT and RoBERTa, which have since been widely explored for their ability to capture contextual and semantic information in clinical texts (LeCun et al., 2015; Liu et al., 2019; Lample et al., 2016).

(Han et al., 2022) implemented three different neural architectures—CNN, LSTM, and BERT—to classify sentences into 13 manually annotated SDoH categories. Their findings demonstrated the effectiveness of BERT in capturing nuanced linguistic patterns in clinical narratives, although the need for high-quality annotated datasets remained a bottleneck. (Ahsan et al., 2021) further explored this domain by comparing traditional machine learning models (Random Forest and XGBoost) with domain-specific LLMs like Bio-ClinicalBERT. Their results, evaluated using both standard F1 scores and capability-centric evaluations like CheckList, highlighted model performance gaps—especially in generalization and robustness across different SDoH categories.

3 Data

(Guevara et al., 2023, 2024) has created MIMIC-SBDH, a publicly available dataset of EHR notes annotated for patients' SBDH status. They developed annotation guidelines for sentence-level annotation of SDoH that are not reliably available as structured data in the EHR: employment, housing, transportation, parental status, relationship, and social support. Sentences were labeled for both the presence of an SDoH mention and the presence of an adverse SDoH mention. As an external validation, they collected 200 notes from 183 patients in the MIMIC (Medical Information Mart for Intensive Care)-III database [24-25], which includes data associated with patients admitted to the critical care units at Beth Israel Deaconess Medical Center in Boston, Massachusetts, from 2001-2008. They manually annotated these 200 MIMC-III notes.

They also generated two files of synthetic sentences labeled with SDoH to enable an exploratory evaluation. Each file consists of 901 synthetic sentences and their SDoH label. These sentences and labels have not been manually verified to be correct and were used directly for data augmentation during model development.

In our experiments, we use the synthetic dataset for training our LLMs and use 200 manually annotated notes as our validation dataset. All our results are presented on the 200 manually annotated notes.

To evaluate our method, we used three annotated datasets:

- **SDOH_MIMICIII_physio_release.csv**: Contains manually annotated clinical sentences derived from the MIMIC-III dataset, serving as our primary gold-standard benchmark.

- **SyntheticSentences_Round1.csv** and **SyntheticSentences_Round2.csv**: Contain artificially constructed sentences representing varied and often nuanced SDOH contexts. These were used to assess how LLMs perform on controlled inputs with clearly defined ground truth.

All datasets were preprocessed to extract sentence-level inputs. Texts were cleaned, split on punctuation, and filtered to remove irrelevant fragments. Each sentence was then prepared for LLM-based evaluation.

3.1 SDoH Variables

The SDOH categories annotated in our study include:

- **Transportation**: `TRANSPORTATION_distance`, `TRANSPORTATION_resource`, `TRANSPORTATION_other`
- **Housing**: `HOUSING_poor`, `HOUSING_undomiciled`, `HOUSING_other`
- **RelationshipStatus**: `RELATIONSHIP_married`, `RELATIONSHIP_partnered`, `RELATIONSHIP_divorced`, `RELATIONSHIP_widowed`, `RELATIONSHIP_single`
- **Parental Status**: `PARENT`
- **EmploymentStatus**: `EMPLOYMENT_employed`, `EMPLOYMENT_under-employed`, `EMPLOYMENT_unemployed`, `EMPLOYMENT_disability`, `EMPLOYMENT_retired`, `EMPLOYMENT_student`
- **Social Support**: `SUPPORT_+`, `SUPPORT_-`

4 Method

Our approach centers on using prompt-engineered large language models (LLMs) to extract Social Determinants of Health (SDOH) from unstructured clinical text. We frame the problem as a zero-shot classification task, wherein a properly prompted LLM is expected to assign binary values (1 or 0) to each of the 20 SDOH attributes defined in clinical annotation guidelines, without the need for model fine-tuning or supervised training.

Each clinical note (or sentence) is annotated as a binary vector over these 20 attributes, with 1 indicating the presence of a given social factor and 0 indicating its absence.

4.1 Prompt Design

We designed structured prompts aligned with annotation guidelines provided in `SDOH_annotation_guidelines.pdf`. Each prompt included a brief instruction contextualizing the model as a clinical annotator, a clear list of all 20 SDOH categories along with their interpretation, and a directive to return only a single line of 20 binary values, comma-separated and in a fixed order. This structure enabled consistent output formatting and reduced the risk of hallucination from the model. An example output would look like: `0,0,1,0,1,0,0,0,1,0,0,1,0,0,0,0,1,0,1,0`.

4.2 Multi-Agent Framework via LangGraph

To improve modularity, interpretability, and scalability, we implemented a multi-agent network using `LangGraph`, a library for declaratively building agent-based computation graphs.

Each agent was responsible for inferring one or more SDOH categories. Specifically:

- **Agent 1 (DeepSeek-R1)**: Annotated the categories under *Transportation*, *Housing*, *Relationship Status*, and *Employment Status*.
- **Agent 2 (LLaMA3)**: Focused on *Parental Status* and *Social Support*.

Both models were hosted on local machines using `Ollama`, allowing for real-time, cost-effective inference without cloud API usage. Agents communicated via LangGraph nodes, enabling efficient orchestration and routing of clinical notes to the appropriate model.

4.3 Annotation Pipeline

The sentence-level clinical data was passed through the LangGraph system, with each sentence invoking both agents. Each agent produced partial binary outputs, which were then merged to form a complete 20-value SDOH annotation vector for that sentence.

The resulting structured annotations were appended to the original dataframe and stored in `TimescaleDB`, a PostgreSQL-compatible time-series database, enabling efficient downstream querying, analysis, and aggregation across patients, note types, or temporal windows.

4.4 Experimental Setup

We used the SDOH-annotated MIMIC-III dataset (`SDOH_MIMICIII_physio_release.csv`) containing clinical sentences with 20 SDOH attributes. The dataset consists of approximately 24,000 annotated sentences.

For benchmarking and training baselines, we focused on the four most frequently occurring SDOH labels based on label frequency analysis: *Housing*, *Employment Status*, *Social Support*, and *Parental Status*.

We split the dataset into 70% training, 15% validation, and 15% test:

- **Training set:** 16,800 sentences

- **Validation set:** 3,600 sentences
- **Test set:** 3,600 sentences

4.5 Baselines

We evaluated our zero-shot LLM-based approach against two baseline models:

- **BioClinicalBERT (Fine-tuned)**: A domain-specific transformer pre-trained on biomedical notes. We fine-tuned it on the top-4 SDOH attributes using a multi-label classification head with Focal Loss to handle class imbalance.
- **Vanilla BERT (Fine-tuned)**: We also experimented with the original BERT model, but results were consistently lower than BioClinicalBERT and are not reported in detail.

Each baseline model was evaluated using accuracy, micro-F1, and macro-F1 metrics. All results were benchmarked on the same held-out test set for fair comparison.

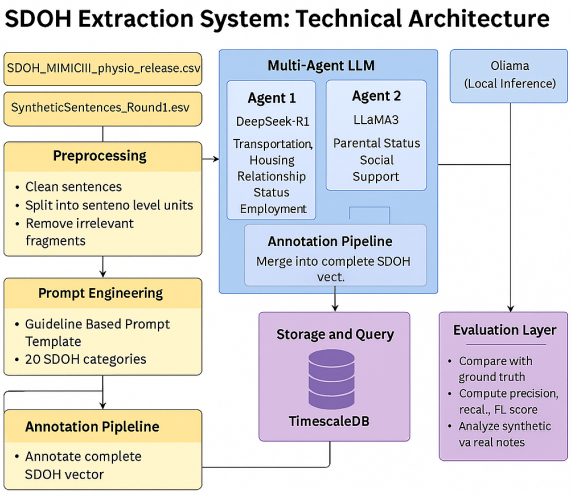

Figure 1: Technical Architecture of the SDOH Extraction System: showing data ingestion, preprocessing, multi-agent LangGraph inference, and structured storage.

4.6 Evaluation

Model predictions were compared against ground truth labels provided in the benchmark datasets. We computed classification metrics (precision, recall, and F1-score) for each SDOH attribute individually to understand the strengths and limitations of prompt-based zero-shot extraction. We also analyzed consistency across synthetic vs real-world clinical text.

References

Hiba Ahsan, Emmie Ohnuki, Avijit Mitra, and Hong You. 2021. Mimic-sbdh: a dataset for social and behavioral determinants of health. In *Machine Learning for Healthcare Conference*, pages 391–413. PMLR.

Marco Guevara, Shan Chen, Spencer Thomas, and Danielle Bitterman. 2023. Annotation dataset of social determinants of health from mimic-iii clinical care database. *Physionet*, 1(0):10–13026.

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.

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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

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 preprint arXiv:1907.11692*.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.

Elliot A Martin, Adam G D'Souza, Vineet Saini, Karen Tang, Hude Quan, and Cathy A Eastwood. 2024. Extracting social determinants of health from inpatient electronic medical records using natural language processing. *Journal of Epidemiology and Population Health*, 72(6):202791.

Zehao Yu, Cheng Peng, Xi Yang, Chong Dang, Prakash Adekkanattu, Braja Gopal Patra, Yifan Peng, Jyotishman Pathak, Debbie L Wilson, Ching-Yuan Chang, and 1 others. 2024. Identifying social determinants

Table 1: Model performance comparison on manually annotated RT dataset (Accuracy scores)

Model	Type	Avg Accuracy	Employment	Housing	Parent	Relationship	Support	Transportation
BioClinicalBERT	Fine-tuned	0.74	0.57	0.56	0.55	0.75	0.58	0.60
BERT-base	Fine-tuned	0.70	0.60	0.58	0.58	0.78	0.60	0.62
Single LLM	Prompt-based	0.78	0.64	0.61	0.58	0.78	0.64	0.64
LangGraph Multi-Agent	**Prompt-based**	**0.81**	**0.69**	**0.65**	**0.63**	**0.84**	**0.68**	**0.68**

Accuracy values for each SDoH category across models. LangGraph shows potential for higher consistency through agent modularity and local inference.

of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *Journal of biomedical informatics*, 153:104642.

A Example Appendix

This is an appendix.