






Research and Applications

SDoH-GPT: using large language models to extract social determinants of health

Bernardo Consoli, MS^{1,2}, Haoyang Wang, BMgt¹, Xizhi Wu, MS³, Song Wang , MS¹, Xinyu Zhao, MS⁴, Yanshan Wang, PhD³, Justin Rousseau , MD⁵, Tom Hartvigsen, PhD⁶, Li Shen , PhD⁷, Huanmei Wu, PhD⁸, Yifan Peng , PhD^{9,10}, Qi Long , PhD⁷, Tianlong Chen, PhD⁴, Ying Ding, PhD^{1,*}

¹School of Information, University of Texas at Austin, Austin, TX 78712, United States, ²School of Technology, Pontifical Catholic University of Rio Grande do Sul, Porto Alegre 90619-900, Brazil, ³Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA 15261, United States, ⁴Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, United States, ⁵Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX 75390, United States, ⁶School of Data Science, University of Virginia, Charlottesville, VA 22903, United States, ⁷Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, United States, ⁸College of Public Health, Temple University, Philadelphia, PA 19122, United States, ⁹Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, United States, ¹⁰Department of Radiology, Weill Cornell Medicine, New York, NY 10065, United States

B. Consoli and H. Wang share co-first authorship.

*Corresponding author. Ying Ding, PhD, School of Information, University of Texas at Austin, 1616 Guadalupe St, Austin, TX 78712, United States (ying.ding@school.utexas.edu)

Abstract

Objective: Extracting social determinants of health (SDoHs) from medical notes depends heavily on labor-intensive annotations, which are typically task-specific, hampering reusability and limiting sharing. Here, we introduce SDoH-GPT, a novel framework leveraging few-shot learning large language models (LLMs) to automate the extraction of SDoH from unstructured text, aiming to improve both efficiency and generalizability.

Materials and Methods: SDoH-GPT is a framework including the few-shot learning LLM methods to extract the SDoH from medical notes and the XGBoost classifiers which continue to classify SDoH using the annotations generated by the few-shot learning LLM methods as training datasets. The unique combination of the few-shot learning LLM methods with XGBoost utilizes the strength of LLMs as great few shot learners and the efficiency of XGBoost when the training dataset is sufficient. Therefore, SDoH-GPT can extract SDoH without relying on extensive medical annotations or costly human intervention.

Results: Our approach achieved tenfold and twentyfold reductions in time and cost, respectively, and superior consistency with human annotators measured by Cohen's kappa of up to 0.92. The innovative combination of LLM and XGBoost can ensure high accuracy and computational efficiency while consistently maintaining 0.90+ AUROC scores.

Discussion: This study has verified SDoH-GPT on three datasets and highlights the potential of leveraging LLM and XGBoost to revolutionize medical note classification, demonstrating its capability to achieve highly accurate classifications with significantly reduced time and cost.

Conclusion: The key contribution of this study is the integration of LLM with XGBoost, which enables cost-effective and high quality annotations of SDoH. This research sets the stage for SDoH can be more accessible, scalable, and impactful in driving future healthcare solutions.

Key words: large language models; social determinants of health; XGBoost classifier; few-shot learning.

Introduction

Social determinants of health (SDoHs) are defined as “non-medical factors that influence health outcomes, including the conditions in which people are born, grow, live, work, and age,” as per the World Health Organization.¹ SDoH have been extensively linked to the prevalence and progression of various diseases^{2–13} and contribute to an astonishing 80%–90% of health outcomes,^{14,15} with multiple factors significantly exacerbating health risks.^{16–19} Critical SDoHs are locked in unstructured clinical narratives.^{20–23} Methods for SDoH extraction using natural language processing (NLP) encompass rule-based (using keyword matching/counts or regular expression, such as¹⁰) tool-based (specialized, task-specific system tools, such as

Moonstone NLP, or cTAKES^{24,25}) and supervised/unsupervised learning approaches, relying on annotated data and lexicons constructed manually or semi-automatically.²⁶ This manual procedure depends extensively on guidelines that steer the annotation process,^{23,27} typically task-specific, resulting in poor reusability. Large language models (LLMs),²⁸ including pre-trained domain specific LLMs,²⁹ have demonstrated promising potential across various healthcare applications,^{30–34} especially showing the capability of reducing the cost and improving the quality of data labeling.^{35–39} Related studies have explored the application of LLMs in SDoH extraction with varying degrees of success.⁴⁰ employed a fine-tuned Flan-T5 model to classify SDoH categories, achieving mid-range F1 scores of 0.55 in

Received: August 30, 2024; Revised: March 13, 2025; Editorial Decision: May 20, 2025; Accepted: June 2, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. All rights reserved.

For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Economics and 0.44 in Community. Similarly, the model used in⁴¹ demonstrated the ability to extract SDoH data from medical notes with moderate success but required extensive fine-tuning and significant computational resources. This article presents a few-shot learning LLM method that uses contrastive examples and concise instructions for SDoH extraction, minimizing the need for extensive annotation guidelines and costly human input. Furthermore, we trained XGBoost classifiers using LLM annotated SDoH data to achieve optimal performance with affordable computational resources. XGBoost was preferred over transformer-based models like BERT due to its lower computational cost and effectiveness with larger datasets, making it a practical choice for handling heterogeneous medical notes and interinstitutional data challenges. The unique combination of LLM and XGBoost utilizes the strength of both methods: with a relative handful of annotations, a few-shot LLM can annotate enough samples to train an XGBoost, which is at least as accurate as, but much cheaper than, purely using LLM. The novelty of this article highlights this efficient combination of LLM and XGBoost as SDoH-GPT to achieve the cost effective and high quality SDoH annotations. The novelty of this article highlights this efficient combination of LLM and XGBoost as SDoH-GPT to achieve the cost-effective and high-quality SDoH annotations.

In this study, we developed SDoH-GPT using GPT-3.5 with few-shot learning and XGBoost to extract SDoH from three datasets: MIMIC-SBDH⁴²; the suicide notes⁴³; and the Sleep Notes⁴⁴ (See Materials and methods; Extended Data 1-3). SDoH-GPT achieved comparable accuracy to human annotations with a tenfold reduced time and a twentyfold decrease in cost, for 2048 sample annotations (Figure 1D). If more annotations had been performed, using SDoH-GPT would have become even cheaper and less time-consuming. When compared to human annotation, SDoH-GPT can become a thousand-fold cheaper and a hundred-fold faster while maintaining comparable accuracy (Extended Data 4) for large-scale clinical notes. With different few-shot learning strategies, SDoH-GPT reached superior consistency with human annotators measured by Cohen's kappa: 0.72-0.92 in MIMIC-SBDH, 0.71-0.88 in suicide notes, and 0.70-0.91 in Sleep Notes (Figure 2E). Through this study, we have identified several error patterns and ambiguity challenges in SDoH annotation and discussed the potential future directions.

Materials and methods

MIMIC SDoH data

MIMIC-III is a publicly available EHR dataset with 55 452 discharge summaries from 46 520 ICU patients (2001-2012).⁴⁵⁻⁴⁷ This database contains 55 452 discharge summary notes from 37 444 non-neonatal patients. The "Social History" section in the dataset containing SDoHs was identified by using regular expressions (RegEX),⁴² excluding neonates and those with missing values. This resulted in 44 566 social histories, divided into 7008 human-annotated entries from MIMIC-SBDH⁴² and 37 558 unannotated entries.

MIMIC-SBDH provides annotations for Social and Behavioral Determinants of Health (SBDH) to 7025 randomly selected discharge summary notes from MIMIC-III.⁴¹ It encompasses annotations for four categories of SDoH: Community (including Community-Present and Community-Absent), Education, Economics, and Environment; and three categories of Behavioral Determinants of Health (BDoH):

Alcohol Use, Tobacco Use, and Drug Use. For clarity, we consider all seven determinants to be SDoH and do not divide them into SDoH and SBoH as MIMIC-SBDH does. Using our RegEX-based method, we finally identified 7,008 social history entries. SDoH-GPT was applied to the three categories with the highest lexical complexity⁴¹: Community, Economics, and Tobacco Use in MIMIC-SBDH. Lexical complexity refers to the variety and richness of vocabulary within a category. Categories with higher lexical complexity feature a broader range of unique terms appeared in this category, reflecting higher diversity in the concepts being discussed.⁴¹

Validation datasets

To evaluate the generalizability of SDoH-GPT, we tested it on two alternative datasets: Suicide Notes and Sleep Notes. The Suicide Notes dataset uses the National Violent Death Reporting System (NVDRS) dataset, which covers 500 072 incidents of suicide deaths across all 50 U.S. states, Puerto Rico, and the District of Columbia from 2003 to 2020.⁴³ For this study, we chose Job Problem as a typical crisis with 30 525 positive and 469,547 negative incidents. The Sleep Notes dataset comes from an Alzheimer's Disease dataset (AD) collected by the University of Pittsburgh Medical Center (UPMC) between January 2016 and December 2020.⁴⁴ This dataset has a cohort of 7266 patients associated with 379 120 clinical documents, 193 351 of which contained keywords related to sleep. For this study, we chose Sleep Apnea as a specific type of sleep disorder with 118 positive and 118 negative incidents.

For the Suicide Notes dataset, employment issues (eg, job-related stress, unemployment, and financial instability), as core SDoHs, have been demonstrated as pivotal factors influencing suicide behavior.^{48,49} Targeting employment-related SDoHs allows for the evaluation of SDoH-GPT's capacity to extract clinically and socially meaningful information from unstructured medical notes, emphasizing its potential as a valuable tool for addressing the challenge of suicide prevention through the identification of critical SDoHs.

The Sleep Notes dataset focuses on extracting information related to sleep apnea, which, while not a direct SDoH, serves as a proxy to evaluate SDoH-GPT's capacity to identify clinically relevant conditions from notes. Sleep apnea is recognized as an important comorbidity with Alzheimer's disease and other chronic health issues,⁵⁰ demonstrating its significance in clinical contexts. Prior research has validated the use of NLP methods to extract sleep apnea information from electronic health records.^{44,51} By addressing sleep apnea extraction, we aim to evaluate SDoH-GPT's broader utility in identifying health determinants and conditions relevant to public health.

These datasets thus represent two complementary dimensions of the capabilities of SDoH-GPT: the Suicide dataset is rooted in the extraction of SDoH for an important medical issue of suicide prevention, while the Sleep Notes dataset demonstrates the model's ability to generalize to tasks beyond SDoH, such as identifying clinically significant conditions like sleep apnea. Together, these analyses showcase SDoH-GPT's versatility in addressing both social and clinical dimensions of health data.

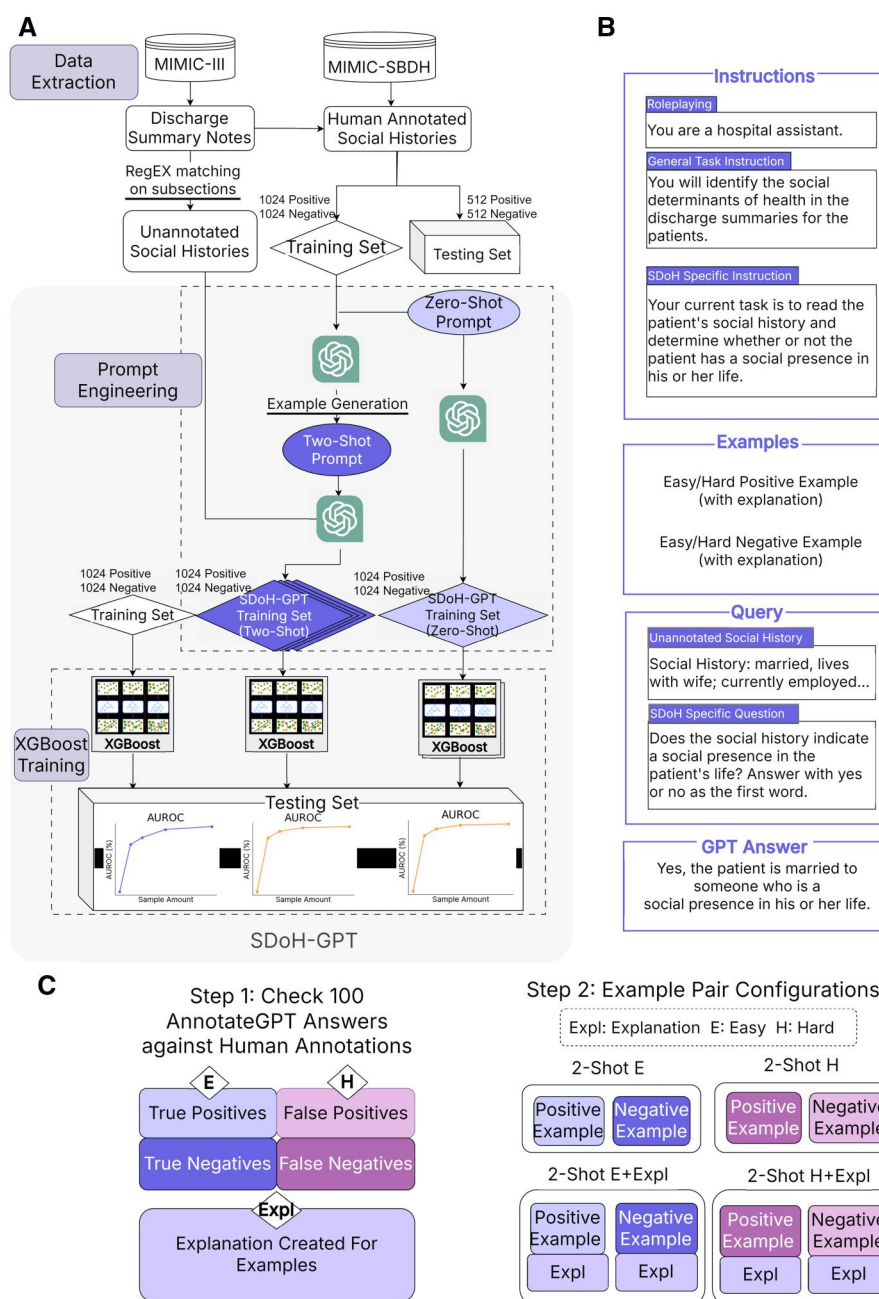


Figure 1. The workflow of SDoH-GPT. (A) The process of building SDoH-GPT, divided into three key sections: Data extraction, prompt engineering, and XGBoost training, for analyzing three MIMIC SDoH categories (Community, Economics, and Tobacco Use). (B) Three sections of SDoH-GPT prompt template: Instructions, Examples, and Query. See Extended sdata 7 and 8 for complete prompts. (C) Two steps on how to generate examples for 2-Shot SDoH-GPT.

Few-Shot prompt

Developing a Zero-Shot prompt is the first step of SDoH-GPT. GPT-3.5 was employed using the Azure OpenAI which is HIPAA compliant (see [Supplementary Material](#) for model parameter settings). This prompt is composed of three instructions and a query. The instructions are as follows: a succinct roleplaying instruction to contextualize LLM, a General Task Instruction that explains the task, and SDoH Specific Instruction to state which information must be extracted. To facilitate straightforward responses from GPT-3.5, our queries were formulated as Yes/No questions. As an illustration, consider the query structure for Economics prompts: “Does the social history indicate

that the patient is currently unemployed or retired? Answer with yes or no as the first word.” This format prompts GPT-3.5 to respond with either “Yes” or “No.” The results were categorized into True Positive and True Negative groups based on gold standard human annotations. This process continued until True Positive and True Negative Zero-Shot prompt-annotated sample groups each comprised a minimum of 50 samples. These new balanced 100 LLM annotations are called the Prototype Annotated Dataset, which is further categorized into four distinct groups by employing 0-Shot SDoH-GPT: True Positives (positive samples correctly categorized by GPT-3.5), True Negatives (negative samples correctly categorized by GPT-3.5), False

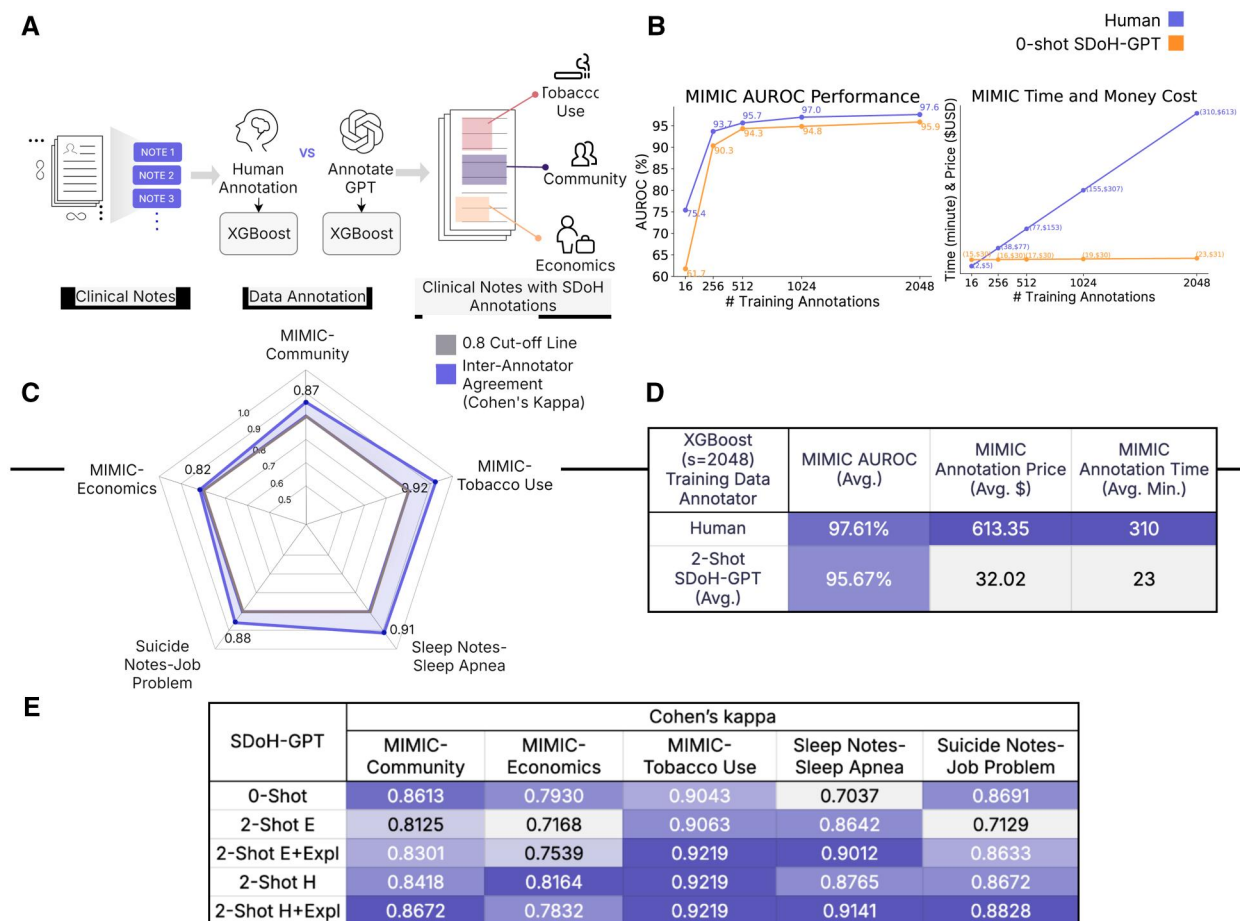


Figure 2. Overview of SDoH-GPT for SDoH annotation. (A) Flowchart illustrating the integration of SDoH-GPT with XGBoost to classify SDoH from clinical notes. (B) Comparison of the average performance, time, and cost of XGBoost classifiers trained on different numbers of human annotations or 0-Shot SDoH-GPT annotations for the three MIMIC SDoH categories. (C) Spider graph showing Cohen's Kappa values, representing consistency between human and SDoH-GPT annotations across MIMIC, suicide notes, and sleep notes. (D) Table summarizing XGBoost performance and cost when trained on 2048 human or 2-Shot SDoH-GPT annotations for MIMIC SDoH categories. (E) Breakdowns of Cohen's Kappa between human, 0-Shot SDoH-GPT, and each 2-Shot SDoH-GPT for 1024 annotated examples from MIMIC-SBDH and Suicide Notes, and 236 annotated examples from Sleep Notes (Extended data 6).

Positives (positive samples incorrectly categorized by GPT-3.5), and False Negatives (negative samples incorrectly categorized by GPT-3.5).

A randomly selected single sample from each group, together with its social history, human-annotated label, and explanation, are systematically organized into a Shot. The explanation refers to the human-created reasoning for the gold standard labeling. Four kinds of Two Shots are generated: the Easy pair (E), consisting of one True Positive example and one True Negative example (ie, 2-Shot E); the Easy-Explained pair (E+Expl), mirroring the examples in the Easy pair, but adding explanations (ie, 2-Shot E+Expl); the Hard pair (H), comprising one False Negative example and one False Positive example (ie, 2-Shot H); and the Hard-Explained pair (H+Expl), replicating the examples in the Hard pair, but including explanations (ie, 2-Shot H+Expl, see Figure 1C).

Our experiment with different shots, ranging from zero to eight, revealed that prompts with zero and two examples exhibit comparable accuracy to those with eight examples. Figure 1B presents the whole structured template of a Two-Shot prompt, consisting of Instructions, Examples, and Query sections.

XGBoost classifiers

We used two kinds of training datasets to train XGBoost classifiers: Human-annotated data from MIMIC-SBDH, and SDoH-GPT-annotated data. The SDoH-GPT-annotated training datasets were created using the Zero-Shot and four kinds of Two-Shot prompts. Each prompt was employed to annotate a balanced training dataset, for a total of five SDoH-GPT training datasets. These six training sets have 1024 positive and 1024 negative examples for each MIMIC SDoH category: Human-Annotated Training Set (From MIMIC-SBDH), 0-Shot SDoH-GPT Training Set, 2-Shot E SDoH-GPT Training Set, 2-Shot E+Expl SDoH-GPT Training Set, 2-Shot H SDoH-GPT Training Set, and 2-Shot H+Expl SDoH-GPT Training Set. The XGBoost classifier trained on human annotations is called XGBoost-Human, while the XGBoost classifier trained on SDoH-GPT annotations is called XGBoost-SDoH-GPT (see Supplementary Material section SDoH-GPT generated Training Datasets). The Testing dataset contains 512 positive and 512 negative examples for each MIMIC SDoH category.

The input data for XGBoost is a 3000-integer array, representing word frequencies in social history samples. The top 3000 words, excluding stop words, are selected from each

training dataset using SciKit Learn's "CountVectorizer"⁵² and NLTK's stop word list. Six balanced training sets of 2048 samples for each SDoH category, were further sub-sampled to smaller balanced datasets with 16, 32, 64, 128, 256, 512, 1024, and 2048 samples. In total, 48 XGBoost models were trained and tested on a 1024-sample balanced dataset. We calculated the AUROC for each model, and the results are presented in Extended Data 5 and 6. See Estimating Time and Money Costs section in [Supplementary Materials](#) for the evaluation of cost estimation. [Figure 1A](#) shows the comprehensive workflow of SDoH-GPT using the MIMIC-III dataset.

Results

SDoH-GPT: an effective SDoH classifier

Training an XGBoost⁵³ classifier using SDoH-GPT annotations can yield enhanced accuracy, consuming less computational resources ([Figure 2A](#)). By scaling LLMs, the SDoH-GPT framework can handle a broad range of tasks without needing task-specific fine-tuning, leveraging few-shot or zero-shot learning to outperform traditional fine-tuning approaches.^{37,54} [Figure 2B](#) shows the average performance of XGBoost classifiers for three MIMIC SDoH categories trained on 16 to 2048 annotated examples either by human annotators or SDoH-GPT with zero-shot learning, indicating that with a mere 256 examples, XGBoost-SDoH-GPT can reach ~0.90 AUROC. The discrepancy between 0-Shot SDoH-GPT and XGBoost-Human with more than 256 annotations was marginal, within a range of 0.014 to 0.022 AUROC ([Figure 2B](#)). In scenarios with high lexical diversity, such as Economics, XGBoost-Human trained on 256 annotations achieved 0.95 AUROC, while XGBoost-SDoH-GPT trained on 256 annotations by 2-Shot SDoH-GPT, consistently maintained above 0.90 AUROC scores for all four kinds of 2-Shot SDoH-GPT (Extended Data 5). For Community, XGBoost-Human trained on 512 annotations surpassed 0.95 AUROC, while XGBoost-SDoH-GPT on 1024 2-Shot H+Expl SDoH-GPT annotations can maintain the same AUROC score (Extended Data 5). In Tobacco Use, both XGBoost-Human and XGBoost-SDoH-GPT, trained with 128 annotations, reached above 0.90 AUROC. These results demonstrate the comparable performance of XGBoost-Human and XGBoost-SDoH-GPT (Extended Data 5). In [Figure 2D](#), the overall performance difference between XGBoost-Human and average 2-Shot SDoH-GPT trained on 2048 annotations is minimal (0.0194 AUROC), yet average 2-Shot SDoH-GPT is approximately 19 times more cost-effective and 13 times faster, respectively (Extended Data 5).

Moreover, annotated examples by human annotators and SDoH-GPT shared extremely high agreement measured by Cohen's kappa.⁵⁵ In our experiments, four kinds of 2-Shot SDoH-GPT achieved substantial agreement with human annotations, with Cohen's kappa scores of 0.87 for Community, 0.82 for Economics, and 0.92 for Tobacco Use in the MIMIC dataset, 0.88 for Suicide Notes, and 0.91 for Sleep Notes ([Figure 2E](#)). These results indicate strong agreement between human annotators and SDoH-GPT. For comparison, Cohen's kappa between two human annotators conducting similar SDoH annotations on clinic notes from Brigham and Women's Hospital/Dana-Farber Cancer Institute was 0.76 for

Employment status (akin to MIMIC Economics) and Social Support (akin to MIMIC community).⁴⁰ Setting the threshold at 0.8 Cohen's kappa, the results affirm SDoH-GPT's high consistency with human annotators, surpassing benchmarks across all SDoH categories evaluated ([Figure 2C](#)).

SDoH-GPT: Cost-effective with high accuracy

We evaluated SDoH-GPT on MIMIC-III discharge summaries to classify the top three highest lexically SDoH categories: Community, Economics, and Tobacco Use,⁴² using binary "Yes" or "No" classifications ([Figure 3A](#)). [Figure 3B](#) shows that the best-performing SDoH-GPT configurations were 2-Shot H+Expl for Community, 0-Shot SDoH-GPT for Economics, and 2-Shot E for Tobacco Use, based on F1 scores. Given the high expense and complexity inherent in human annotation, related studies typically have a limited number of SDoH annotations: 1000,⁵⁶ 1576⁵⁷ and 500.⁵⁸ Assuming only 512 human annotations are available for Community, 2-Shot H+Expl SDoH-GPT can efficiently generate an additional 2048 annotations at approximately one-fifth of the cost and one-third of the time required for human annotators. Furthermore, it achieves an AUROC score that is 0.01 higher than that of XGBoost trained on the same 512 human annotations (Extended Data 5). Notably, SDoH-GPT requires only 100 human annotations to select two examples for 2-Shot SDoH-GPT, demonstrating significant cost-effectiveness and efficiency. Ablation studies revealed minimal variance in AUROC scores between 0-Shot SDoH-GPT and various 2-Shot SDoH-GPT trained on 2048 annotations, suggesting that additional shots do not necessarily enhance performance ([Figure 3C](#)). Employing 2-Shot SDoH-GPT directly for annotating the Testing Set without XGBoost yielded an F1 score nearly equivalent to that achieved by using XGBoost trained on 2048 human annotations, reducing the need for extensive manual annotation by a factor of twenty to maintain comparable F1 scores ([Figure 3D](#)). While 0-Shot SDoH-GPT outperformed XGBoost trained on 2048 human annotations in F1 scores for Community and Tobacco Use categories, the Economics category posed greater challenges due to its highest lexical complexity in SBDH,⁴² resulting in slightly lower scores, as shown in [Figure 3D](#).

In this study, we attained a notably higher F1 score using 2-Shot SDoH-GPT: 0.905 in Economics, which is 0.102 higher than those from GPT-4 in Ramachandran et al⁴¹; 0.963 in Tobacco Use, surpassing 0.138 than those in Ramachandran et al⁴¹; and 0.926 in Community, a significant improvement of 0.336 over Living Status in Ramachandran et al.⁴¹ Several factors potentially contribute to the differences: (1) Variance in GPT prompt structure: Ramachandran et al's⁴¹ query in GPT prompts was to annotate discharge summaries in the BRAT standoff format, while ours were pure SDoH categorization; (2) Dataset differences: Ramachandran et al⁴¹ contains MIMIC III and an additional dataset from the University of Washington; while ours are MIMIC III and two other datasets; (3) Instruction guideline: Ramachandran et al's⁴¹ prompt included a lengthy instruction, exceeding 1000 characters; while our instruction employed a more concise instruction, limited to a few sentences; and (4) Two-shot learning: Ramachandran et al's⁴¹ prompt did not have two examples to facilitate two-shot learning. In comparison, Guevara et al,⁴⁰ used 200 manually annotated MIMIC-III notes to finetune Flan-T5 (18M parameters) with LoRA,

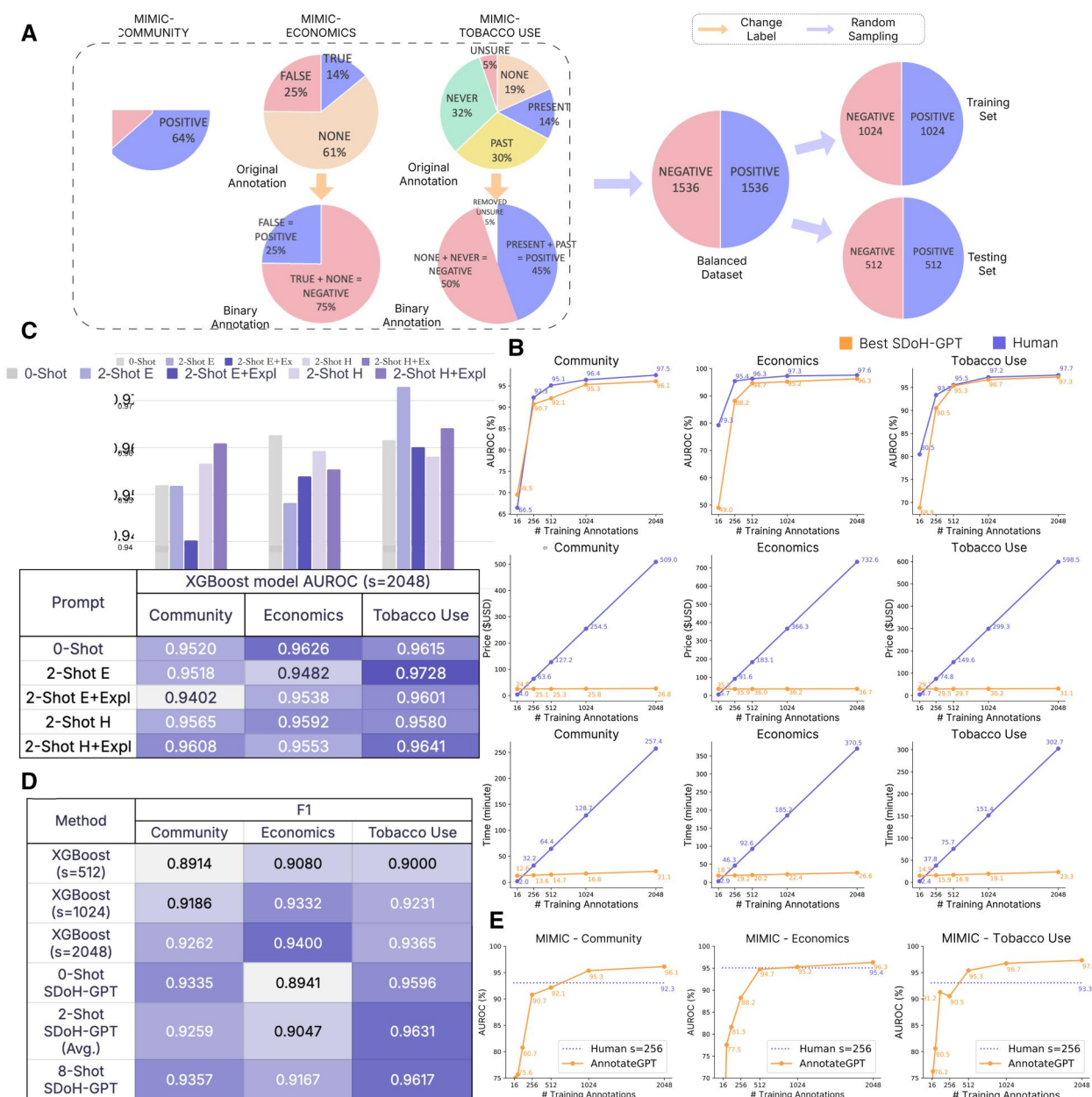


Figure 3. SDoH-GPT for MIMIC SDoH Classification. (A) Merging different values of each MIMIC SDoH category into balanced binary values to create training and testing sets. (B) The comparison results of AUROC, time and money cost for XGBoost classifiers trained on human annotations versus those trained on annotated examples from 2-Shot H+Expl SDoH-GPT for community, the 0-Shot SDoH-GPT for economics, and 2-Shot E SDoH-GPT for tobacco use (the best performing SDoH-GPT for their respective tasks). (C) Ablation studies to show the AUROC performance of XGBoost classifier trained on 2048 examples generated by 0-Shot SDoH-GPT and different 2-Shot SDoH-GPT. (D) Comparison of F1 measures on directly using 0-Shot and Few-Shot SDoH-GPT to classify three MIMIC SDoH categories of the Testing Set (1024 samples) without training XGBoost classifiers, against XGBoost classifiers directly trained on different numbers of human annotations. (E) Assuming only 256 human annotations are available (the dotted line), the number of 2-Shot H+Expl SDoH-GPT annotated examples are needed to reach the same level of accuracy for Community; same for Economics with 0-Shot SDoH-GPT and Tobacco Use with 2-Shot E SDoH-GPT.

achieving 0.44 F1 in Community and 0.55 in Economics. Our F1 scores for these categories were nearly double. With just 256 human annotations, SDoH-GPT can efficiently generate more annotations, boosting AUROC by up to 0.04 (Figure 3E).

Validating SDoH-GPT using suicide notes and sleep notes

SDoH-GPT was validated using two datasets: Sleep Notes and Suicide Notes. Sleep Notes were segmented into

sentences, while Suicide Notes were divided into two paragraphs. Each segment was processed using specific prompts to identify mentions of sleep apnea or associations with job problems (Figure 4A). A binary “Yes” or “No” classification balanced dataset was created through random sampling (Figure 4B). The resulting datasets were split into Training Set and Testing Set, containing equal positive and negative samples for model evaluation. Both 0-Shot and 2-Shot SDoH-GPT were employed to ascertain the association between job problems and the victim’s suicide ideation. Figure 4C presents the results in AUROC, time and cost. For

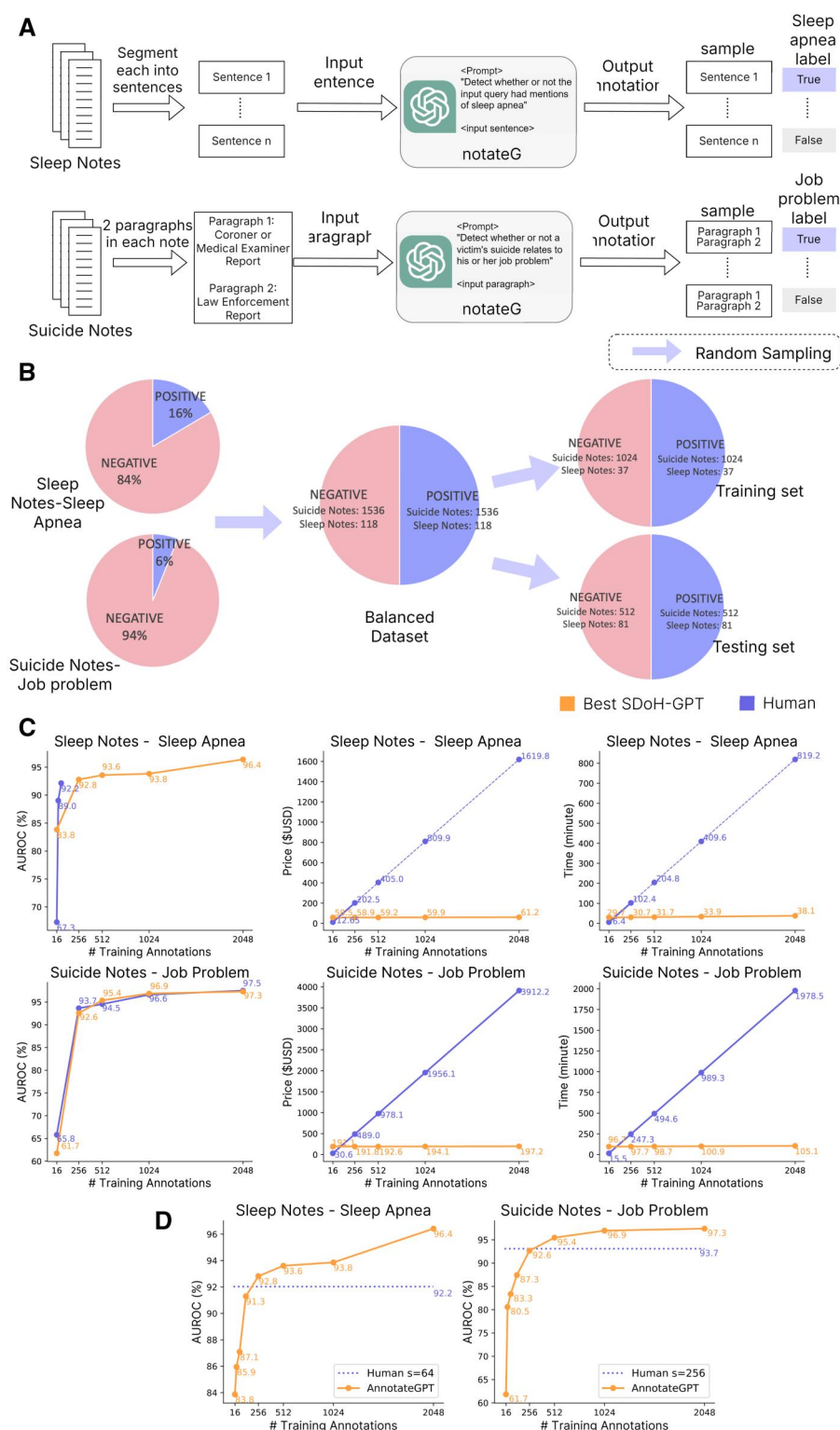


Figure 4. SDOH-GPT evaluation on sleep notes and suicide notes. (A) Data pre-processing. (B) Procedure to generate the training and testing sets. We took a large set as testing set for sleep notes to ensure stable performance scores to avoid huge variances due to the small samples of testing set. (C) Comparable performance of XGBoost-Human versus XGBoost-SDoH-GPT trained on different numbers of annotations generated by either human annotators or 2-Shot H SDOH-GPT (the best performing SDOH-GPT for both validation tasks). (D) Performance improvements with XGBoost-SDoH-GPT in cases of limited human annotations.

Suicide Notes, XGBoost-SDoH-GPT trained on annotations from 2-Shot H SDOH-GPT performs comparably to XGBoost-Human with significantly reduced time and computational cost. If only 256 human annotations are available,

XGBoost-SDoH-GPT exhibited a substantial increase in AUROC score by 0.036 (Figure 4D). Moreover, with mere 128 annotations, both XGBoost-Human and XGBoost-SDoH-GPT with 0-Shot, 2-Shot E+Expl and 2-Shot H+Expl

attained 0.90 AUROC score. For Sleep Notes, where only 236 human annotations were available for sleep apnea (74 as Training Set and 162 as Testing Set), the peak AUROC achieved by XGBoost-Human is 0.922. By spending an extra \$2.68 dollars and 8.5 min, XGBoost-SDoH-GPT, trained 2048 annotations generated by 2-Shot H SDoH-GPT, can improve AUROC from 0.922 to 0.964 (Figure 4C and D). This demonstrates the substantial utility of SDoH-GPT in areas where human annotations are scarce and expensive to obtain. SDoH-GPT can markedly enhance performance with minimal additional effort.

Understanding SDoH-GPT annotation errors

Figure 5A shows the distribution of four categories of errors: Human error; SDoH-GPT error, Extraction error, and Ambiguity. We have summarized the following potential causes: (Figure 5B). (1) Abbreviations confuse SDoH-GPT. Replacing the abbreviation, such as “VNA” with its full form, “Visiting Nurse Association,” enabled SDoH-GPT to correctly identify relevant SDoHs; (2) Check-box style data are frequently used in clinical notes,⁵⁹ such as “Smoked no [x] yes [].” Converting them into plain language, such as “Never smokes,” improves the model’s accuracy. (3) Multiple SDoHs are in one single sentence. The example “Does not smoke and gave up drinking 10 years prior” includes two SDoHs—tobacco use and alcohol use—in a single sentence. By separating it into two sentences, “Does not smoke” and “Gave up drinking 10 years prior,” SDoH-GPT annotated the information correctly. (4) Unclear phrases: For instance, the phrase “Distant smoking history” where the term “distant” is broad and vague. Rewriting it as “Previously had a history of smoking but quit a long time ago” provides clearer information, enabling SDoH-GPT to classify the example correctly; (5). Hallucinations remain a major concern for LLMs.⁶⁰ Our example (Figure 5B) lacked explicit information about employment status. However, SDoH-GPT concluded that the patient was unemployed despite the absence of relevant details in the notes; and (6). Misunderstanding of multiple time frames: Errors can occur when notes contain mixed time frames referencing past, present, or future events. For example, in the Economics task, the phrase “Retired from his work as a manager in auto sales. He states he hopes to return to his previous work in the future” describes a past occupation and does not provide information about the individual’s current employment status. SDoH-GPT struggles to accurately differentiate between past occupations and current work status in this case. For the identified error patterns, improving the prompt design can significantly enhance the performance of SDoH-GPT (See the Refining SDoH-GPT Prompts section in the Supplementary Materials).

Discussion

SDoH-GPT demonstrated remarkable efficiency, achieving a tenfold reduction in annotation time and a twentyfold decrease in costs compared to traditional methods, while maintaining strong alignment with human annotators, as evidenced by Cohen’s kappa scores of up to 0.92. Our 2-Shot SDoH-GPT, a more effective approach compared to,⁴¹ achieved significantly higher F1 scores in Economics, Tobacco Use, and Community categories, avoiding the excessive complexity instructions and extensive fine-tuning. In comparison to,⁶¹ which utilized traditional classification

algorithms such as XGBoost, TextCNN, and Sentence-BERT, our approach combines the strengths of LLMs and XGBoost, thereby reducing the costs and time associated with creating large training annotation sets, while simultaneously achieving better performance.

SDoH data in medical notes are often brief and lack context, leading to ambiguous annotations. There are several pivotal issues concerning ambiguity (Figure 5C): (1) Contextual Misinterpretations: SDoH-GPT identified Case1 as community present, misinterpreting the context of “lost family”; (2) Evolving Status: SDoH-GPT’s annotation in Case2 was limited to the initial part of the sentence, leading to the annotation of community present; (3) Implicit Statements: Case3 suggests daily access to community services for the patient, yet this does not explicitly imply anything about community presence, while SDoH-GPT marked it as community present. Daily access to healthcare services cannot directly infer community presence.⁶² The frequency and duration of contact impact the patient’s subjective social support and overall well-being^{61–64}; and (4) Incomplete Information: It is unclear whether the patient in Case4 lives with his children or not. Nevertheless, SDoH-GPT classified this as community present, which can be questionable. Ambiguity in SDoH stems from its inherently complex and multifaceted nature, requiring a nuanced understanding and context-specific analysis to ensure precise and meaningful annotation.^{57,65,66} Effectively addressing these issues requires a comprehensive understanding of the medical domain, as well as the broader societal context pertinent to healthcare. By incorporating hard examples into SDoH-GPT prompts, we improve the model’s ability to handle complex scenarios.

While SDoH-GPT demonstrates competitive performance, certain limitations must be acknowledged for its applicability in broader clinical settings. First, our SDoH categorization is binary (Yes or No), which may not adequately capture clinical complexity. Future work should explore multi-label or hierarchical classification frameworks to capture richer and more actionable insights. Second, the reliance on the MIMIC-III dataset, which is primarily based on ICU visits, introduces potential constraints in capturing the variability of SDoH across different care settings and patient populations. Third, social determinants are highly context-dependent, varying significantly between individuals based on their unique circumstances. Future work should address these limitations by validating datasets encompassing a wider range of clinical settings and patient demographics. Fourth, our approach to SDoH annotation is confined to the categorical level and does not extend to sentence-level annotation of triggers and spans. This limits the model’s ability to capture context-specific details in real world clinical contexts. Additionally, real world EHRs are far more fragmented, heterogeneous, and complex, requiring tools capable of handling multiple data formats and sections. Previous studies have demonstrated that incorporating sentence-level annotations can enhance the extraction of SDoH performance.^{67,68} To address these challenges, future work should aim to enhance SDoH-GPT’s ability to handle this diversity while incorporating sentence-level annotation techniques for more precise and meaningful insights. Fifth, we acknowledge the LLM hallucination in our research. The combination of XGBoost and LLM can mitigate a certain level of hallucination by relying more on XGBoost for classification when the training dataset has been sufficiently built. Lastly, while SDoH-GPT

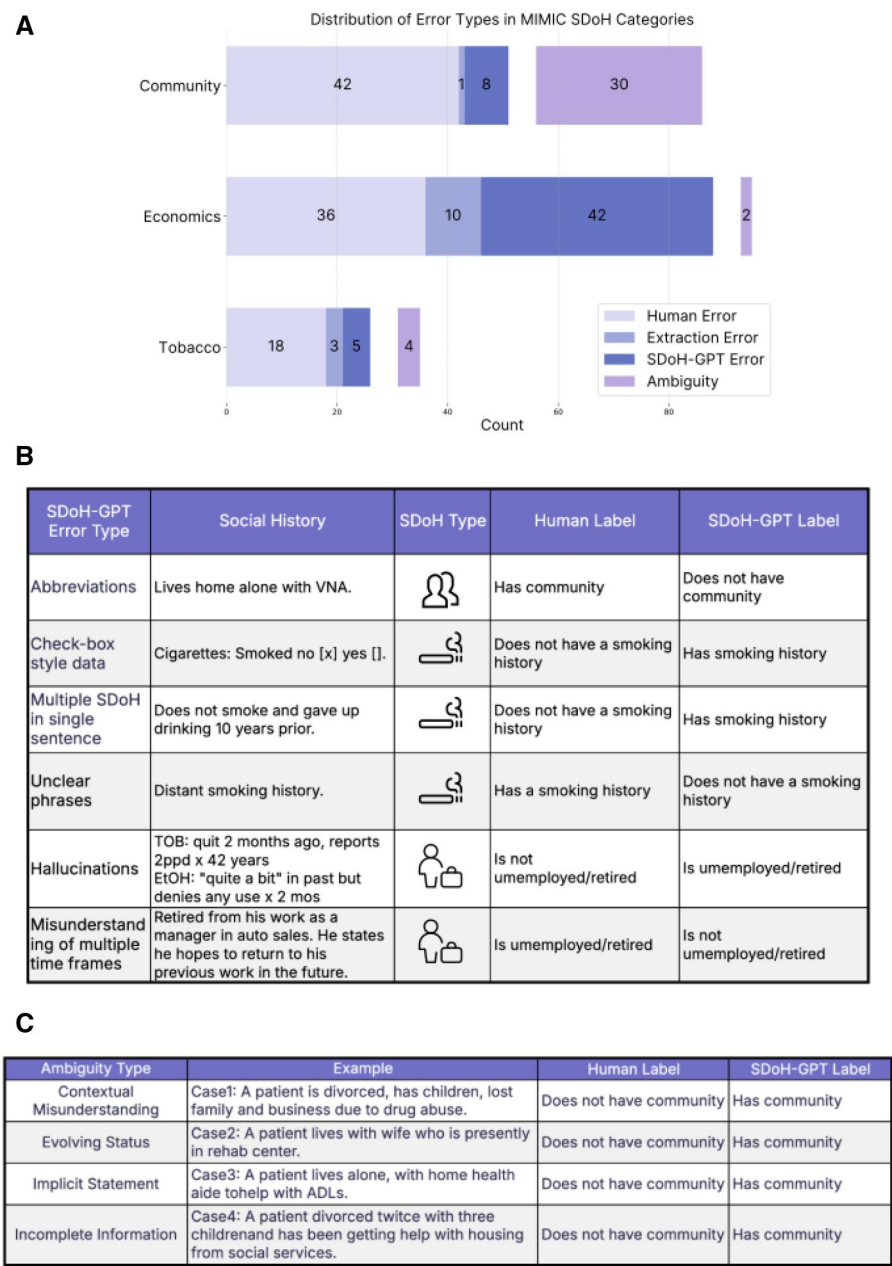


Figure 5. Error analysis of SDoH-GPT. (A) Distribution of error types in MIMIC SDoH categories: Human error (ie, errors in human annotations); SDoH-GPT error (ie, errors in SDoH-GPT annotations); Extraction error (ie, incorrect extractions of social histories from discharge summaries using Regular Expression algorithms), and Ambiguity (ie, hard to decide). The error analysis was conducted on a randomly selected sample of 1000 SBDH records. (B) SDoH-GPT errors were classified into six categories: Abbreviations, check-box style data, multiple SDoH in a single sentence, unclear phrases, hallucinations, and misunderstanding of multiple time frames. (C) Ambiguity cases and their corresponding human annotations and SDoH-GPT annotations.

demonstrates generalizability across different tasks, including applications in domains such as suicide and sleep apnea, further targeted analysis is required to fully explore its adaptability in different areas. Future research should also explore tailored adaptations of the model, such as optimizing prompts and integrating domain-specific knowledge for these unique contexts.

Conclusion

SDoH-GPT introduces a practical and efficient way to extract SDoH from unstructured clinical notes. We combined the

flexibility of few-shot LLM with the speed and simplicity of XGBoost, reducing the time and cost of annotation without sacrificing accuracy. This research sets the stage for a future where SDoH is more accessible, scalable, and impactful in driving future healthcare solutions.

Author contributions

Bernardo Consoli (Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—review & editing), Haoyang Wang (Software, Validation, Visualization,

Writing—review & editing), Xizhi Wu (Data curation, Formal analysis, Methodology, Resources, Software, Visualization), Song Wang (Data curation, Formal analysis, Resources, Software, Validation, Visualization), Xinyu Zhao (Data curation, Formal analysis, Resources, Software, Validation, Visualization, Writing—review & editing), Yanshan Wang (Conceptualization, Funding acquisition, Resources, Software, Supervision, Validation, Writing—review & editing), Justin Rousseau (Conceptualization, Data curation, Resources, Software, Supervision, Validation), Tom Hartvigsen (Resources, Software, Supervision, Validation, Writing—review & editing), Li Shen (Conceptualization, Resources, Software, Supervision, Writing—review & editing), Huanmei Wu (Supervision, Validation, Writing—review & editing), Yifan Peng (Conceptualization, Data curation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing—review & editing), Qi Long (Resources, Supervision, Validation, Writing—review & editing), Tianlong Chen (Resources, Software, Supervision, Validation, Visualization, Writing—review & editing), and Ying Ding (Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing—original draft, Writing—review & editing)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

We would like to acknowledge the following funding supports: NIH OTOD032581, NIH OTA-21-008, and NIH R01LM014306-01.

Conflicts of interest

None declared.

Data availability

MIMIC-III (<https://physionet.org/content/mimiciii/1.4/>) and NVDRS (<https://www.cdc.gov/violenceprevention/datasources/nvdrs/dataaccess.html>) are third-party datasets available for credentialed use. MIMIC-SBDH is a publicly available third-party dataset (<https://github.com/hibaahsan/MIMIC-SBDH>). The Sleep Notes dataset is protected by HIPAA law. It is restricted to research and healthcare use. Access can be granted by the University of Pittsburgh's Office of Sponsored Programs (osp@pitt.edu), which has a 2-6-month response timeframe. Our codebase is available upon request (corresponding author).

References

- World Health Organization. Social determinants of health; 2025. Accessed March 7, 2025. <https://www.who.int/health-topics/social-determinants-of-health>
- Virnig BA, Baxter NN, Habermann E, et al. A matter of race: early- versus late-stage cancer diagnosis. *Health Aff Proj Hope*. 2009;28:160-168. <https://doi.org/10.1377/hlthaff.28.1.160>
- Özdemir BC, Dotto G-P. Racial differences in cancer susceptibility and survival: more than the color of the skin? *Trends Cancer*. 2017;3:181-197. <https://doi.org/10.1016/j.trecan.2017.02.002>
- Cancer Facts and Figures 2020; 2020. Accessed January 17, 2025. <https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2020.html>
- Burgard SA, Ailshire JA, Kalousova L. The great recession and health: people, populations, and disparities. *Ann Am Acad Pol Soc Sci*. 2013;650:194-213. <https://doi.org/10.1177/0002716213500212>
- Draper CE, Grobler L, Micklesfield LK, et al. Impact of social norms and social support on diet, physical activity and sedentary behaviour of adolescents: a scoping review. *Child Care Health Dev*. 2015;41:654-667. <https://doi.org/10.1111/cch.12241>
- Szreter S, Woolcock M. Health by association? Social capital, social theory, and the political economy of public health. *Int J Epidemiol*. 2004;33:650-667. <https://doi.org/10.1093/ije/dyh013>
- Williams DR, Mohammed SA. Discrimination and racial disparities in health: evidence and needed research. *J Behav Med*. 2009;32:20-47. <https://doi.org/10.1007/s10865-008-9185-0>
- Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: a systematic review. *J Am Med Inform Assoc JAMIA*. 2020;27:1764-1773. <https://doi.org/10.1093/jamia/ocaa143>
- Hatef E, Rouhizadeh M, Tia I, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Med Inform*. 2019;7:e13802. <https://doi.org/10.2196/13802>
- Jilani MH, Javed Z, Yahya T, et al. Social determinants of health and cardiovascular disease: current state and future directions towards healthcare equity. *Curr Atheroscler Rep*. 2021;23:55. <https://doi.org/10.1007/s11883-021-00949-w>
- Zettler ME, Feinberg BA, Jeune-Smith Y, et al. Impact of social determinants of health on cancer care: a survey of community oncologists. *BMJ Open*. 2021;11:e049259. <https://doi.org/10.1136/bmjopen-2021-049259>
- Ding X, Kharrazi H, Nishimura A. Assessing the impact of social determinants of health on diabetes severity and management. *JAMIA Open*. 2024;7:ooae107. <https://doi.org/10.1093/jamiaopen/ooae107>
- Magnan S. Social determinants of health 101 for health care: five plus five. *NAM Perspect*. 2017;7. <https://doi.org/10.31478/201710c>
- Hood CM, Gennuso KP, Swain GR, et al. County health rankings: relationships between determinant factors and health outcomes. *Am J Prev Med*. 2016;50:129-135. <https://doi.org/10.1016/j.amepre.2015.08.024>
- Bejan CA, Angiolillo J, Conway D, et al. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J Am Med Inform Assoc JAMIA*. 2018;25:61-71. <https://doi.org/10.1093/jamia/ocx059>
- Schroff P, Gamboa CM, Durant RW, et al. Vulnerabilities to health disparities and statin use in the REGARDS (reasons for geographic and racial differences in stroke) study. *J Am Heart Assoc*. 2017;6:e005449. <https://doi.org/10.1161/JAHA.116.005449>
- Reshetnyak E, Ntamatungiro M, Pinheiro LC, et al. Impact of multiple social determinants of health on incident stroke. *Stroke*. 2020;51:2445-2453. <https://doi.org/10.1161/STROKEAHA.120.028530>
- Pinheiro LC, Reshetnyak E, Sterling MR, et al. Multiple vulnerabilities to health disparities and incident heart failure hospitalization in the REGARDS study. *Circ Cardiovasc Qual Outcomes*. 2020;13:e006438. <https://doi.org/10.1161/CIRCOUTCOMES.119.006438>
- Wang M, Pantell MS, Gottlieb LM, et al. Documentation and review of social determinants of health data in the EHR: measures and associated insights. *J Am Med Inform Assoc JAMIA*. 2021;28:2608-2616. <https://doi.org/10.1093/jamia/ocab194>

21. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform*. 2020;8:e17984. <https://doi.org/10.2196/17984>
22. Shrank WH, Rogstad TL, Parekh N. Waste in the US health care system: estimated costs and potential for savings. *JAMA*. 2019;322:1501-1509. <https://doi.org/10.1001/jama.2019.13978>
23. Wei Q, Franklin A, Cohen T, et al. Clinical text annotation—what factors are associated with the cost of time? *AMIA Annu Symp Proc*. 2018;2018:1552-1560.
24. Conway M, Keyhani S, Christensen L, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semant*. 2019;10:6. <https://doi.org/10.1186/s13326-019-0198-0>
25. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc JAMIA*. 2010;17:507-513. <https://doi.org/10.1136/jamia.2009.001560>
26. Patra BG, Sharma MM, Vekaria V, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc JAMIA*. 2021;28:2716-2727. <https://doi.org/10.1093/jamia/ocab170>
27. Fort K, Nazarenko A, Rosset S. Modeling the complexity of manual annotation tasks: a grid of analysis. In: Kay M, Boitet C, eds. *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee; 2012:895-910.
28. Open AI, Achiam J, Adler S, et al. GPT-4 Technical Report; 2024.
29. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172-180. <https://doi.org/10.1038/s41586-023-06291-2>
30. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619:357-362. <https://doi.org/10.1038/s41586-023-06160-y>
31. Nazario-Johnson L, Zaki HA, Tung GA. Use of large language models to predict neuroimaging. *J Am Coll Radiol JACR*. 2023;20:1004-1009. <https://doi.org/10.1016/j.jacr.2023.06.008>
32. Sorin V, Barash Y, Konen E, et al. Large language models for oncological applications. *J Cancer Res Clin Oncol*. 2023;149:9505-9508. <https://doi.org/10.1007/s00432-023-04824-w>
33. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. <https://doi.org/10.2196/45312>
34. Brin D, Sorin V, Vaid A, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep*. 2023;13:16492. <https://doi.org/10.1038/s41598-023-43436-9>
35. Ding B, Qin C, Liu L, et al. Is GPT-3 a Good Data Annotator? 2023.
36. Guo X, Chen Y. Generative AI for synthetic data generation: methods, challenges and the future; 2024.
37. Agrawal M, Heggelmann S, Lang H, et al. Large language models are few-shot clinical information extractors; 2022.
38. Sushil M, Kennedy VE, Mandair D, et al. CORAL: expert-curated oncology reports to advance language model inference. *Nejm AI*. 2024;1:aidbp2300110. <https://doi.org/10.1056/AIdbp2300110>
39. Goel A, Gueta A, Gilon O, et al. LLMs accelerate annotation for medical information extraction. *Proceedings of the 3rd Machine Learning for Health Symposium*. PMLR; 2023:82-100.
40. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. *Npj Digit Med*. 2024;7:6. <https://doi.org/10.1038/s41746-023-00970-0>
41. Ramachandran GK, Fu Y, Han B, et al. Prompt-based extraction of social determinants of health using few-shot learning. In: Naumann T, Ben Abacha A, Bethard S, et al., eds. *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics; 2023:385-393.
42. Ahsan H, Ohnuki E, Mitra A, et al. MIMIC-SBDH: a dataset for social and behavioral determinants of health. *Proc Mach Learn Res*. 2021;149:391-413.
43. Nguyen BL, Lyons BH, Forsberg K, et al. Surveillance for violent deaths—national violent death reporting system, 48 States, the District of Columbia, and Puerto Rico, 2021. *MMWR Surveill Summ*. 2024;73:1-44. <https://doi.org/10.15585/mmwr.ss7305a1>
44. Sivarakumar S, Tam TYC, Mohammad HA, et al. Extraction of sleep information from clinical notes of Alzheimer's disease patients using natural language processing. *J Am Med Inform Assoc JAMIA*. 2024;31:2217-2227. <https://doi.org/10.1093/jamia/ocae177>
45. Johnson A, Pollard T, Mark R. MIMIC-III clinical database; 2015.
46. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016;3:160035. <https://doi.org/10.1038/sdata.2016.35>
47. Goldberger AL, Amaral LA, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101:e215-e220. <https://doi.org/10.1161/01.cir.101.23.e215>
48. Wang S, Dang Y, Sun Z, et al. An NLP approach to identify SDoH-related circumstance and suicide crisis from death investigation narratives. *J Am Med Inform Assoc*. 2023;30:1408-1417. <https://doi.org/10.1093/jamia/ocad068>
49. Beseran E, Pericàs JM, Cash-Gibson L, et al. Deaths of despair: a scoping review on the social determinants of drug overdose, alcohol-related liver disease and suicide. *Int J Environ Res Public Health*. 2022;19:12395. <https://doi.org/10.3390/ijerph191912395>
50. Pan W, Kastin AJ. Can sleep apnea cause Alzheimer's disease? *Neurosci Biobehav Rev*. 2014;47:656-669. <https://doi.org/10.1016/j.neubiorev.2014.10.019>
51. Hsu E, Malagaris I, Kuo Y-F, et al. Deep learning-based NLP data pipeline for EHR-scanned document information extraction. *JAMIA Open*. 2022;5:ooac045. <https://doi.org/10.1093/jamiaopen/ooac045>
52. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
53. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery; 2016:785-794.
54. Nori H, King N, McKinney SM, et al. Capabilities of GPT-4 on medical challenge problems; 2023.
55. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20:37-46. <https://doi.org/10.1177/001316446002000104>
56. Bhate N, Mittal A, He Z, et al. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using GPT model; 2023.
57. Lituiev DS, Lacar B, Pak S, et al. Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients. *J Am Med Inform Assoc JAMIA*. 2023;30:1438-1447. <https://doi.org/10.1093/jamia/ocad054>
58. Yu Z, Yang X, Dang C, et al. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. *AMIA Annu Symp Proc AMIA Symp*. 2021;2021:1225-1233.
59. Wilbanks BA, Moss J. Evidence-based guidelines for interface design for data entry in electronic health records. *CIN Comput Inform Nurs*. 2018;36:35-44. <https://doi.org/10.1097/CIN.0000000000000387>
60. Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy. *Nature*. 2024;630:625-630. <https://doi.org/10.1038/s41586-024-07421-0>
61. Keloth VK, Selek S, Chen Q, et al. Large language models for social determinants of health information extraction from clinical notes—a generalizable approach across institutions. *medRxiv*. 2024;2024. <https://doi.org/10.1101/2024.05.21.24307726>
62. Boamah SA, Weldrick R, Lee T-SJ, et al. Social isolation among older adults in long-term care: a scoping review. *J Aging Health*. 2021;33:618-632. <https://doi.org/10.1177/08982643211004174>
63. Drageset J, Kirkevold M, Espehaug B. Loneliness and social support among nursing home residents without cognitive impairment:

- a questionnaire survey. *Int J Nurs Stud*. 2011;48:611-619. <https://doi.org/10.1016/j.ijnurstu.2010.09.008>
64. Cheng S-T, Lee CKL, Chow PK-Y. Social support and psychological well-being of nursing home residents in Hong Kong. *Int Psychogeriatr*. 2010;22:1185-1190. <https://doi.org/10.1017/S1041610210000220>
 65. Lybarger K, Ostendorf M, Yetisgen M. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *J Biomed Inform*. 2021;113:103631. <https://doi.org/10.1016/j.jbi.2020.103631>
 66. Feller DJ, Bear Don't Walk Iv OJ, Zucker J, et al. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Appl Clin Inform*. 2020;11:172-181. <https://doi.org/10.1055/s-0040-1702214>
 67. Han S, Zhang RF, Shi L, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform*. 2022;127:103984. <https://doi.org/10.1016/j.jbi.2021.103984>
 68. Lybarger K, Yetisgen M, Uzuner Ö. The 2022 n2c2/UW shared task on extracting social determinants of health. *J Am Med Inform Assoc JAMIA*. 2023;30:1367-1378. <https://doi.org/10.1093/jamia/ocaf012>