Extracting Social Determinants of Health from Pediatric Patient Notes Using Large Language Models: Novel Corpus and Methods

Yujuan Fu[1*], Giridhar Kaushik Ramachandran[2*] , Nicholas J Dobbins[1]
Namu Park[1], Michael Leu[1], Abby R. Rosenberg[3,4,5], Kevin Lybarger[2], Fei Xia[1,6]
Özlem Uzuner[2], Meliha Yetisgen[1]
[1]Department of Biomedical Informatics & Medical Education, University of Washington
[2]Department of Information Sciences and Technology, George Mason University
[3]Dana-Farber Cancer Institute, [4]Boston Children's Hospital,
[5]Harvard Medical School, [6]Department of Linguistics, University of Washington
[1,4]Seattle, WA, USA, [2]Fairfax, VA, USA, [3,4,5]Boston, MA, USA
{velvinfu, ndobb, npark95, mgl27, fxia, melihay}@uw.edu
{gramacha, klybarge, ouzuner}@gmu.edu, abbyr_rosenberg@dfci.harvard.edu
[*]Authors contributed equally to this paper.

Abstract

Social determinants of health (SDoH) play a critical role in shaping health outcomes, particularly in pediatric populations where interventions can have long-term implications. SDoH are frequently studied in the Electronic Health Record (EHR), which provides a rich repository for diverse patient data. In this work, we present a novel annotated corpus, the Pediatric Social History Annotation Corpus (PedSHAC), and evaluate the automatic extraction of detailed SDoH representations using fine-tuned and in-context learning methods with Large Language Models (LLMs). PedSHAC comprises annotated social history sections from 1,260 clinical notes obtained from pediatric patients within the University of Washington (UW) hospital system. Employing an event-based annotation scheme, PedSHAC captures ten distinct health determinants to encompass living and economic stability, prior trauma, education access, substance use history, and mental health with an overall annotator agreement of 81.9 F1. Our proposed fine-tuning LLM-based extractors achieve high performance at 78.4 F1 for event arguments. In-context learning approaches with GPT-4 demonstrate promise for reliable SDoH extraction with limited annotated examples, with extraction performance at 82.3 F1 for event triggers.

Keywords: Social Determinants of Health, Pediatrics, Information Extraction, Large Language Models

1. Introduction

Health outcomes and quality of life are affected by the conditions in which people work and live and are referred to as Social Determinants of Health (SDoH) (Centers for Disease Control and Prevention, 2022). SDoH are particularly important in pediatric populations because health disparities have a long-term impact on future attainment of health, including educational and economic success (Thompson et al., 2016; Dickson et al., 2023). Clinicians have continuously adapted practices by systematically gathering pediatric patients' SDoH during clinical consultations (Garg et al., 2013; Ho et al., 2016; Kazak et al., 2015). Previous research has identified screening and intervention for SDoH risks in pediatric patients associated with better health outcomes and highlighted the necessity for a more comprehensive SDoH tool (Morone, 2017).

However, there are difficulties in documenting SDoH in Electronic Health Records (EHRs) in a tabular format, mainly due to the diversity of SDoH determinants, individual determinants' infrequent occurrence, and inconsistent reporting practice (Linfield et al., 2023). Many pediatric SDoH elements are primarily documented within the clinical narratives from EHRs. Such predominance of unstructured SDoH information in the EHRs impedes the systematic collection and utilization of SDoH information in clinical and research settings, limiting the potential for data-driven inventions to improve individual and public health.

To address these challenges, natural language processing (NLP) information extraction (IE) models are needed to extract semantic representations of SDoH, to enable large-scale and real-time use of this information. IE in the clinical domain and, more broadly, in the general domain has predominantly used fine-tuning-based techniques; recent advancements in instruction-tuned large language models (LLMs) (Thirunavukarasu et al., 2023), trained on large data repositories, are enabling in-context learning approaches.

Although there is a robust body of IE research exploring SDoH for adult populations, including the development of annotated data sets and data-driven IE models, there is comparatively little IE research investigating the SDoH of pediatric patients. In this work, we present the **Ped**iatric **S**ocial **H**istory **A**nnotation **C**orpus (**PedSHAC**), an annotated corpus of ten distinct SDoH determinants on clinical narratives from pediatric patients from the

LREC-COLING 2024, pages 7045–7056
20-25 May, 2024.

University of Washington (UW) hospital system. This corpus bridges the gaps in the literature by creating a human-annotated comprehensive and fine-grained corpus of SDoH phenomena for pediatric patients.

To the best of our knowledge, our novel pediatric SDoH corpus, PedSHAC, is the first annotated corpus of pediatric clinical narratives to utilize comprehensive and fine-grained SDoH annotations, including assigning SDoH labels such as *Status* and *Type* that could be incorporated into structured data fields within EHRs to represent patient information better. We believe that this corpus will be a valuable resource in support of understanding the role of SDoH in managing children's health and improving outcomes. Using PedSHAC, we explored various LLM-based IE strategies and demonstrated that detailed SDoH representations can be extracted with high accuracy. The de-identified PedSHAC corpus, annotation guideline, and code are made available through our GitHub[1].

2. Related Work

Our contributions include a novel corpus of pediatric clinical narratives with fine-grained annotations (PedSHAC) and comprehensive IE model development for benchmarking. To contextualize both contributions, we describe literature related to both SDoH corpora and IE methods.

2.1. SDoH Corpora

The interplay of various social and economic factors on patient health has led to an increased interest in investigating SDoH. To facilitate SDoH exploration, multiple SDoH corpora have been developed. However, their annotation schema might have generally lacked granularity and comprehensiveness, or the patient population might have limited extension into the pediatric domain

For the adult population, many studies have focused on a limited number of SDoH factors with singular focus such as smoking status (Uzuner et al., 2008; Savova et al., 2008), homelessness (Gundlapalli et al., 2013; Bejan et al., 2018), and substance use (Wang et al., 2015; Yetisgen and Vanderwende, 2017; Carrell et al., 2015; Alzubi et al., 2022). Previous research also addresses SDoH factors in specific contexts, such as sexual health (Feller et al., 2018) and hospital readmission rate (Navathe et al., 2018). Our prior SDoH work investigated adult SDoH factors using a fine-grained, event-based annotation scheme encompassing detailed status and type labels for adults (Lybarger et al., 2021).

[1]https://github.com/uw-bionlp/PedSHAC

Pediatric SDoH factors such as adverse childhood experiences were researched in the adult patient population (Bejan et al., 2018; Wu et al., 2022b,a). The rest of prior SDoH work focused on adult populations doesn't necessarily extend to pediatric-patient-focused corpora, because pediatric populations have unique SDoH factors and there are many factors associated with caregivers that impact the SDoH and health of pediatric patients. For example, education access (DeJong et al., 2016) and food insecurity (Baer et al., 2015) are especially important to pediatric patients. The clinical notes of pediatric patients may describe employment associated with patient caregivers (Kuhlthau and Perrin, 2001; Xie et al., 2023); at the same time, patient parents' mental health (Stallard et al., 2004) become important as pediatricians continually evaluate whether children may be at risk for child abuse and neglect (Farrell et al., 2017). PedSHAC bridges this gap in the literature with comprehensive fine-grained annotation of SDoH determinants with a focus on pediatric patients.

2.2. Extraction methods

SDoH IE is an increasingly explored task, and the modeling approaches range from manually curated rules (Patra et al., 2021; Hatef et al., 2019), traditional/shallow machine learning models (Clark et al., 2008; Wang et al., 2015), neural networks (Bejan et al., 2018; Gehrmann et al., 2018), to transformer-based LLMs (Patra et al., 2021; Bompelli et al., 2021).

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is frequently used in SDoH extraction tasks for text classification (Yu et al., 2021, 2022; Han et al., 2022) and entity and relation extraction (Richie et al., 2023; Lybarger et al., 2023a). Sequence-to-sequence approaches that utilize generative LLMs, like Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020), have also achieved high performance (Romanowski et al., 2023). The most recent generation of LLMs, such as GPT-4 (OpenAI, 2023), are pre-trained on large amounts of data and instruction-tuned (Ouyang et al., 2022), enabling prompt-based learning methods with zero or few in-context examples. Recent work demonstrates the use of GPT-based models in few-shot clinical IE (Agrawal et al., 2022; Yang et al., 2023).

This work explores pediatric SDoH extraction using multiple transformer-based methods, including fine-tuning through BERT- and T5-based models, and in-context learning using GPT-4. Our experiments showed human-comparable performance through fine-tuning and relatively high performance through in-context learning. Our pipeline is versatile and can be readily adapted to

various IE tasks, as a reference for the broader research community.

3. Methods

3.1. Dataset

This work utilized the clinical notes of pediatric patients from the UW hospital system. The patient cohort consists of a random sample from the general pediatric population to improve generalizability across patient demographics. The clinical notes span a ten-year period (1/1/2012-12/31/2021) with 198k distinct notes from 36k distinct patients. Clinical notes are organized into topical sections that are delineated by specific heading formats. Patient SDoH can be described throughout the clinical narrative; however, SDoH are most frequently documented in the social history sections of the clinical notes. To focus the annotation on SDoH-dense portions of the clinical notes, we applied a rule-based approach to identify topical section headings and the social history sections, yielding 11k social history sections for 8k distinct patients. The social history section text for a patient can be very similar or identical across notes, so we randomly selected one social history section per patient, resulting in 8k patients, each with a single social history section. Finally, we randomly sampled 1,260 out of 8K social history sections for SDoH annotations.

3.1.1. Annotation scheme

We created detailed annotation guidelines for ten SDoH (referred to here as *event types*), as listed in Table 1. The three substance events, *alcohol*, *drug*, and *tobacco*, are annotated and evaluated separately, but their performance is reported together due to their relatively low frequency.

Each event is characterized by a trigger and multiple arguments that describe the event's status, type, and status. The *trigger* is a span with an event-type label. Each *argument* attaches to the corresponding trigger and is assigned a multi-class label, referred to here as a *subtype* label [2], representing *normalized* SDoH concepts (such as *Status - past*, *current*) that are more suitable for downstream clinical applications. Because the most important clinical information is usually stored in a structured format in EHRs, the normalized SDoH concepts as labels can be directly added to other structured information to create a more comprehensive patient representation. Arguments can be categorized into *required* and *optional*. The required arguments define the most important attributes of the event. A trigger can only be annotated if all required arguments can be resolved.

The annotation scheme and event type distribution are specified in Table 1. SDoH information was annotated using the BRAT rapid annotation tool (Stenetorp et al., 2012). Figure 1 is an example describing the patient's living arrangement and caregivers' employment.

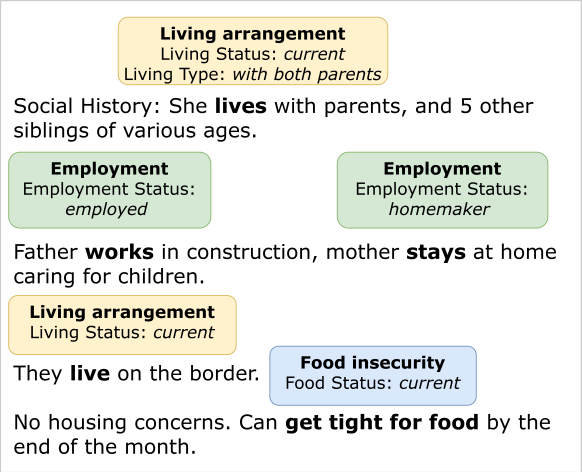

Figure 1: An Annotation example: the triggers are in boldface. The box above a trigger shows the event type, arguments and subtype labels.

3.1.2. IE evaluation

We follow the previous N2C2 SDoH challenge (Lybarger et al., 2023b) evaluation criteria. We evaluate the trigger and argument extraction performance for each event. Two triggers are considered *equivalent* if they have the same event type and overlapping spans. The trigger extraction is framed as a named entity recognition task, and the precision, recall, and F1 are calculated. Two arguments are considered *equivalent* if they are attached to equivalent triggers and have the same argument type and subtype labels, and are evaluated using precision, recall, and F1.

3.1.3. Annotator agreement

Six medical students at UW annotated SDoH events in our dataset. We first performed two practice rounds to train the annotators and refine the annotation guidelines, with 5 and 10 notes, respectively. After the practice rounds, each note was annotated by two annotators (double annotation), with a third annotator adjudicating disagreements. The Inter-Annotator Agreement (IAA) is evaluated

[2] *arguments* and *subtype* labels can be considered as attribute names and attribute values. We chose this naming convention following the previous N2C2 SDoH challenge (Lybarger et al., 2023b).

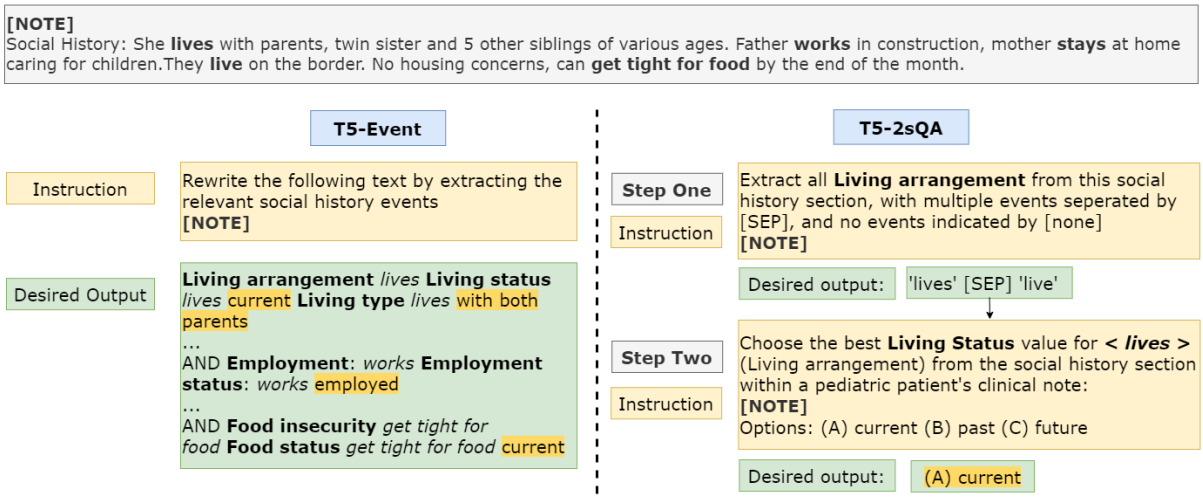

Figure 2: Our one-step (T5-Event) and two-step (T5-2sQA) extraction models. T5-Event extracts all SDoH events, including triggers and arguments, in one query. T5-2sQA extracts triggers and arguments in separate queries, where Step Two includes the predicted triggers from Step One.

Event	Trigger & Arg.	Trigger examples & Argument subtypes	# labels			IAA
			Train	Validation	Test	F1
Adoption	Trigger	“adopted”, ...	27	4	9	100.0
Education Access	Trigger	“5th grade” , “junior year”, ...	227	35	74	80.0
	Status	(yes,no)	227	35	74	80.0
Employment	Trigger	“Employment: ... ”, “works”, ...	390	45	117	81.1
	Status	(employed, unemployed, retired, on disability, student, homemaker)	390	45	117	77.8
Food Insecurity	Trigger	“food stamps”, “food insecurity”, ...	37	5	8	40.0
	Status	(current, past, none)	37	5	8	40.0
Living Arrangement	Trigger	“lives”, “foster care”, ...	676	101	195	90.4
	Status	(current, past, future)	676	101	195	88.5
	Type*	(with both parents, with single parent, with other relatives, with foster family, with strangers)	566	86	160	88.4
	Residence*	(home, institution, homeless)	136	22	38	38.1
Mental Health	Trigger	“depression”, “self-harm”, ...	45	11	15	66.7
	Status	(current, past, none)	45	11	15	53.3
	Experiencer	(patient, parent/caregiver)	45	11	15	66.7
Substance Use - Alcohol / Drug / Tobacco	Trigger	“meth”, ”alcohol”, “smokes”,...	265	38	78	86.4
	Status	(current, past, none)	265	38	78	85.7
	Experiencer	(patient, parent/caregiver)	265	38	78	73.2
Trauma	Trigger	“mentally abusive”, “bullying”, ...	132	23	33	88.9
	Status	(yes, no)	132	23	33	88.9
	Type	(divorce / separation, loss, psychological, physical, domestic violence, sexual)	132	23	33	84.6

Table 1: Annotation scheme and event statistics for PedSHAC, where * indicates optional arguments. The train, validation, and test sets contain 894, 121, and 245 notes, respectively. The IAA micro-averaged F1 (%) is calculated on the last round of double annotation, consisting of 90 notes. The IAA F1 micro averages on triggers, arguments, and triggers plus arguments are 85.1, 80.0, and 81.9, respectively.

using the criteria in Section 3.1.2. We doubly annotated 360 notes through 4 rounds (90 notes per round) and then singly annotated the remaining 885 notes. PedSHAC has an IAA micro average of 85.1 F1 across all triggers and 80.0 F1 across all arguments in the last double-annotation round with 90 notes. Low IAAs are from infrequently occurring events such as *Food Insecurity* and *Mental Health*, and the annotation group carefully discussed every disagreement. PedSHAC is split into

training, validation, and test sets. Table 1 presents the distribution of SDoH for each split along with the IAA for all event types. The entirety of the test set and the majority of the validation set are doubly annotated.

3.2. SDoH Information Extraction

We experimented with various LLM types and learning strategies, including i) fine-tuning BERT, ii) fine-tuning T5, and iii) in-context learning with GPT-4. The generative model experimentation with T5 and GPT-4 explored multiple prompting strategies, including i) single-step text2event (Event), and ii) two-step question answering (2sQA). Both prompting approaches were explored with T5 through fine-tuning and GPT-4 through in-context learning.

Fine-tuning BERT (mSpERT): Following prior work in the N2C2 SDoH challenge (Lybarger et al., 2023b), we use our high-performing, multi-label variation of the Span-based Entity and Relation Transformer model (mSpERT) (Eberts and Ulges, 2020; Lybarger et al., 2023a)[3], as the BERT baseline. mSpERT is a span-based extractor that jointly extracts entities and relations. In the PedSHAC extraction task, mSpERT assigns multiple labels to a given span and assumes all predictions for a given span are associated with the same event. As all PedSHAC arguments share the same span as the trigger, mSpERT did not generate any relation predictions between spans.

Fine-tuning T5 with single-step text2event Prompting (T5-Event): Recent work (Lu et al., 2021; Ma et al., 2023a) demonstrates that entity and relation extraction tasks can be reformulated into text2event tasks using generative encoder-decoder models like T5 (Raffel et al., 2020; Chung et al., 2022) and decoder-only models like GPT-4 (OpenAI, 2023). We map each event annotation to a structured text representation (Romanowski et al., 2023; Lu et al., 2021)[4]. Figure 2 illustrates our T5-Event approach. Input sequences included the entire social history section and a model instruction. The target sequence was a sequence of SDoH events containing trigger type and text span, followed by the required and optional arguments. The trigger span was repeated with its argument to associate the arguments with the trigger span. Multiple events in the output were separated with 'AND' for parsing. T5-Event extracts all PedSHAC SDOH events for a social history section in one step.

[3] https://github.com/Lybarger/sdoh_extraction

[4] https://github.com/romanows/SDOH-n2c2/

Fine-tuning T5 with two-step QA Prompting (T5-2sQA): we utilize a two-step pipeline approach to first extract triggers spans (Ma et al., 2023a) and then resolve subtype labels through multiple-choice questions (Ma et al., 2023b). Figure 2 illustrates our two-step approach. In step 1, the model input is a prompt specifying the target event type and the social history section text, and the model's desired output is a list of trigger spans associated with the target event type. In step 2, we apply multi-choice QA to resolve the argument subtype labels for each identified trigger and each argument type relevant to the event type. The input prompt specifies the argument, the relevant trigger within the note, and all possible argument subtypes. An additional choice, "none," is added for optional arguments, indicating the argument may not be present for that event. The model output is the selected subtype.

GPT-4 with In-context Learning: Previous research demonstrates LLMs can achieve high performance through in-context learning (Agrawal et al., 2022). Additionally, some proprietary LLMs, including GPT-4, cannot currently be fine-tuned. Using prompt-based, in-context learning, information about the desired task is conveyed through instructions and few-shot examples. The larger context window of recent LLMs, including GPT-4 (OpenAI, 2023), which can accommodate up to 32k tokens, allows detailed text-based instructions and several response examples to be included in the prompt. We explored three in-context learning strategies: i) **Event** and **2sQA** – simple instructions without explanation of annotated phenomena. For **GPT-Event**, our instruction contained a list of all the event and argument types and an illustration of the T5-Event output format using a randomly chosen example note. **GPT-2sQA** uses the same prompts provided to T5-2sQA, ii) **GPT + guide** – *2sQA* prompt with a brief description of target trigger/argument based on a summary of the annotation guideline, iii) **GPT + 3-shot** – three few-shot examples, in addition to the *GPT + guide* prompts. For the *+3-shot* setting, we randomly selected three example social history sections from the train set per GPT query, with some restrictions: (1) for trigger extraction: the three example notes contained zero, one, and more than one triggers of specific event type respectively; (2) for required argument extraction, three randomly selected examples of events with that argument type (positive examples); and (3) for optional argument extraction such as *residence*, one random negative example as an event without that argument, and two random positive examples, are included from event associated with the argument type.

3.3. Experimental Paradigm

In fine-tuning, we trained extraction models on the train set, optimized the hyperparameters on the validation set, and applied the best-performing models to the withheld test set. In in-context learning, we utilized the annotation guideline and examples from the train set. We initialized the BERT-based mSpERT model from Bio+ClinicalBERT (Alsentzer et al., 2019). For T5 experimentation, we initialized from Flan-T5-Large (780M) (Chung et al., 2022), an instruction-tuned T5 variant. For GPT-4 experiments, we used OpenAI's GPT-4-32k (version: 2023-03-15-preview) with the chat completion API provided through our HIPAA-compliant Azure server instance and utilized the 'role' preset ('system', 'user', and 'assistant') arguments for providing our prompts. The system message includes the same instructions as the T5 experiments (except for the subtype options) and the distilled annotation guideline. The user message includes the note and subtype options for the argument extraction. We utilize multiple user-assistant input pairs to simulate the conversation history as in-context learning few-shot examples.

4. Results

4.1. Trigger and argument evaluation

Following the evaluation criteria described in Section 3.1.2, we report the extraction performance on the withheld PedSHAC test set in Table 2 under two settings: i) fine-tuning with mSpERT and T5 and ii) in-context learning with GPT-4. We validate the F1 scores and assess significance using a pairwise non-parametric test (bootstrap test, p-val < 0.05) (Berg-Kirkpatrick et al., 2012) for all approaches, but only present a subset of significance testing results in Table 2 due to lack of space. We consider the mSpERT model as a baseline for all approaches, with GPT-Event and GPT-2sQA base as a baseline for in-context learning approaches. The '*' indicates performance fine-tuning approaches with significance over mSpERT or vice versa and † marks in-context learning models with significantly higher performance than GPT-Event and GPT-2sQA base. The highest performance in each row is boldfaced.

Comparing performance against human IAA[5], GPT+3-shot shows comparable performance in trigger micro average (82.3 F1) to corresponding IAA (85.1 F1), and T5-2sQA shows argument micro average (78.4 F1) close to corresponding IAA (80.0 F1). For event types with lower IAA rates, such as *Mental Health* (trigger and all arguments) and *Living Arrangement* (*residence* argument), the extraction performance is also lower, indicating complexity in the SDoH descriptions.

For fine-tuning approaches, all models exhibit high trigger extraction performance with no significant difference. Comparing arguments micro average, T5-2sQA demonstrates significantly better performance than mSpERT, as well as all other in-context learning models. But on the level of individual argument types, T5-2sQA performance is similar to mSpERT and T5-Event, with the exception of the *Living Arrangement - type* argument. We observed no significant difference between T5-Event and T5-2sQA, indicating with sufficient fine-tuning data, the Flan-T5-large model can extract multiple events with complex, fine-grained event annotations appearing at the same time.

Comparing in-context-learning approaches with GPT-4, GPT-Event and GPT-2sQA base approaches demonstrate relatively lower performance when limited scheme information is incorporated into the prompt. Similar to the T5-Event and T5-2sQA models, the GPT-Event and GPT-2sQA base approaches have no significant difference in the trigger and argument extraction performances. Starting from GPT-2sQA base, adding the guidelines (+guide) provides the model with a detailed annotation scheme description, leading to significant improvement as 8.5 (from 71.3 to 79.8) among triggers and 9.8 (from 60.0 to 69.8) among arguments. Adding three in-context learning examples further improves the performance (GPT+3-shot) from the base 2sQA with 11.0 (from 71.3 to 82.3) among triggers and 11.6 (from 60.0 to 71.6) among arguments. Adding the guidelines to the GPT-2sQA model (+guide) shows comparable trigger performance with the fine-tuned models. The GPT+3-shot achieves the highest trigger extraction performance, albeit without statistically significant improvement from the GPT+guide. Specifically, the GPT+3-shot model shows a significant increase in performance for *Education access*, *Employment*, and *Substance Use* extraction over GPT-Event and GPT-2sQA base, while showing a significant increase even over mSpERT for *Employment* extraction. The GPT+3-shot model demonstrates similar performance to the fine-tuned models for extracting *Education Access*, *Employment*, *Living Arrangement*, and *Substance Use* event types.

[5]Note that the last round IAA is not directly comparable to LLM performance. Because (1) IAA is from the last double-annotation round, while the model performance is calculated on the whole test set, (2) the test set has resolved the annotator disagreement from the IAA. Therefore, the IAA is not an upper bound for LLM performance on the test set, but a reference to 'good' performance.

Event	Trigger & Arg.	# gold labels	Extraction performance (F1)						
			Fine-tuning			In-context learning			
			mSpERT	T5-Event	T5-2sQA	GPT-Event	GPT-2sQA		
							base	+guide	+guide +3-shot
Adoption	Trigger	9	**84.2**	82.4	**84.2**	58.1	66.7	66.7	54.5
Edu. Access	Trigger	74	78.0	79.1	84.1	71.6	75.9	84.9	**85.7**†
	Status	74	78.0	79.1	84.1	71.6	53.3	**85.5**†	84.5†
Employment	Trigger	117	75.1	78.9	81.1	69.1	73.4	85.5*†	**89.2***†
	Status	117	71.4	76.3	74.3	60.8	64.0	76.9†	**80.6***†
Food Insecurity	Trigger	8	**93.3**	87.5	**93.3**	53.3	0.0	70.0	87.5
	Status	8	**93.3**	87.5	**93.3**	53.3	0.0	70.0	87.5
Living Arrg.	Trigger	195	84.8	**86.5**	85.4	82.3	80.9	83.7	84.0
	Status	195	82.6	83.4	**84.4**	80.2	78.4	81.0	78.4
	Type	160	83.3	82.7	**88.7***	76.6	75.4	81.2	77.9
	Residence	38	63.5	**67.6**	62.2	27.7	27.2	28.0	28.6
Mental Health	Trigger	15	38.1	25.0	36.4	26.3	51.9	**53.3**	51.6
	Status	15	28.6	25.0	34.8	26.3	35.7	38.7	**43.8**
	Experiencer	15	9.5	8.3	17.4	21.1	35.7*	40.0*	**43.8***
Subst. Use	Trigger	78	**85.5***	81.6	81.9	54.1	64.2	73.5†	80.2†
	Status	78	81.4	78.1	**81.9**	50.8	63.2	69.0	76.8†
	Experiencer	78	74.5	80.3	**80.6**	49.2	63.2	72.1†	80.0†
Trauma	Trigger	33	62.1	54.5	53.3	58.6	5.7	55.3	**70.2**
	Status	33	51.7	54.5	54.2	58.6	5.7	55.3	**63.2**
	Type	33	55.2	51.5	54.2	55.2	5.7	55.3	**66.7**
Micro Avg	Trigger	529	79.6	79.5	80.9	69.9	71.3	79.8†	**82.3**†
	Arguments	844	75.3	76.0	**78.4***	62.0	60.0	69.8†	71.6†

Table 2: Model performance F1 (%) on event triggers and arguments from the PedSHAC withheld test set. The asterisk * indicates that performance was significantly better ($p<0.05$) than mSpERT or vice versa. The symbol † marks in-context learning models with significantly higher performance than GPT-Event and GPT-2sQA. The highest performance in each row is in boldface.

Event	# gold labels	Event extraction performance (F1)				
		Fine-tuning			In-context learning	
		mSpERT	T5-Event	T5-2sQA	GPT-Event	GPT+ 3-shot
Adoption	9	**84.2**	82.4	**84.2**	58.1	54.5
Edu. Acc.	74	78.0	79.1	84.1	71.6	**84.5**
Employment	117	71.4	73.5	74.3	60.8	**79.7**
Food. Insec.	8	**93.3**	87.5	**93.3**	53.3	73.7
Living Arrg.	195	72.8	69.7	**74.9**	19.8	12.6
Mental Health	15	9.5	8.3	17.4	21.1	**37.5**
Subst. Use	78	75.9	75.0	**80.6**	45.9	78.0
Trauma	33	51.7	51.5	53.3	55.2	**59.6**
Micro Avg	529	71.6	70.4	**74.7**	42.6	54.0

Table 3: Model performance F1 (%) with the *event-level* evaluation on the PedSHAC withheld test set.

4.2. Event-level evaluation

We additionally assess performance using a more rigorous *event-level* evaluation criteria, which requires the equivalence (defined in Section 3.1.2) of all arguments in an event type. A predicted event is considered correct if and only if its trigger overlaps with a trigger in the gold standard and all arguments in the event are correctly identified with the correct subtype labels. Table 3 presents the event-level performance for the best GPT-2sQA approach and the rest of the approaches. We conduct the same pairwise significance testing across all models as Section 4.1, yet exclude the results from Table 3 to improve readability.

The T5-2sQA model achieves the highest micro-average performance, as well as significantly better performance than the in-context learning approaches in *Living Arrangement*, *Substance Use*, and micro average. Both mSpERT and T5-Event have similar performance to T5-2sQA. There is no significant difference among all fine-tuning models in any event.

Note that the trigger extraction performance bounds event-level performance. Comparing Table 2 with Table 3, three fine-tuning approaches have a relatively small performance drop on the micro average from trigger to event, as 6.2 (from 80.9 to 74.7) for T5-2sQA, 8.0 (from 79.6 to 71.6) for mSpERT and, 9.1 (from 79.5 to 70.4) for T5-Event. This is because trigger extraction is a more challenging task, and the fine-tuning-based LLMs can correctly predict the argument if they are able to correctly identify the trigger. This demonstrates great promise for fine-tuning-based LLMs' downstream clinical use at the event extraction level.

On the other hand, the GPT+3-shot shows a performance drop of 28.3 (from 82.3 to 54.0). This is mainly because the GPT+3-shot model shows poor performance on some arguments (i.e. *Living Arrange - residence*) and the difficulty of predicting multiple arguments correctly at the same time for the same event.

4.3. Error Analyses

Comparing errors across different learning strategies, we observed that the fine-tuning models tend to have relatively lower recall than precision, while the in-context learning models tend to have lower precision than recall. While fine-tuning models perform well in extracting SDoH for event types well-represented in the training set, they demonstrate relatively poorer generalizability. This could be because fine-tuning models contain much fewer parameters than GPT-4 and have less prior knowledge about some SDoH factors. For example, if a *Mental Health* trigger phrase is uncommon and not previously seen in the train set, the fine-tuning models can fail to extract it. On the other hand, the in-context learning approaches tend to interpret SDoH extraction in a broader context and extract events outside the annotation scheme. For example, 'Dad </name>, Mom </name> and Sister </name>' is a list of the family members' names, which does not explicitly state the patient's living arrangement. However, the GPT+3-shot approach considers this span implying a *Living Arrangement* event and annotates it as a trigger.

Without fine-tuning, GPT+3-shot is very sensitive to the instructions provided in the form of the guideline. For example, our guideline did not state that the *residence* subtype needs to be explicitly mentioned, and GPT-4 predicted descriptions such as 'lives with parents' having the optional argument *residence* with the subtype *home*'. Such false positives resulted in a precision of 17.2 and 28.6 F1 for the *residence* argument. GPT+3-shot also sometimes extracts meaningful SDoH information but fails to overlap with the gold annotation, especially in the *Food Insecurity* events. For example, clinicians tend to follow a template format: 'Food insecurity: NO'. while GPT+3-shot tends to extract the phrase following the prefix and predicts 'No' as the trigger, the annotators annotate the prefix, 'Food insecurity', as the gold trigger. On the other hand, because T5-based approaches learn from abundant annotated data, they were able to learn from the actual implementation of the guide and implicitly understand edge cases that are not explicitly defined in the guide. Future GPT-based models could use better-designed prompts to incorporate more detailed instructions or better sample selection approaches for in-context learning.

Consistent with errors identified by prior work (Ji et al., 2023), both generative models (T5 and GPT-4) show a problem of hallucination (Ji et al., 2023), outputting with improper formats, which range from minor modifications to spacing, punctuation, and casing. Another type of hallucinated response is spans that do not correspond to the original text, such as synonyms to the original SDoH determinants. We consider the generated output invalid if the predictions do not comply with the predefined output format or the predictions contain predicted spans that do not exactly match the original text. We observed a 3-5% invalid rate for trigger prediction and less than 1% for argument prediction in the QA approaches. Future work could apply approaches to better constrain the prediction within the note and annotation scheme, including rule-based post-editing such as minimum edit distance, self-verification (Gero et al., 2023) and constrained decoding (Lu et al., 2021).

4.4. Limitations

Our annotation of the SDoH events in PedSHAC is limited to a single hospital system and its pediatric population. The distribution of the SDoH events may not be representative of other pediatric populations. The relatively lower frequencies of some of the event types may result from the patient population at our institution. The current annotation scheme does not allow multiple events of the same event type to have the same trigger span. For example, in the sentence, 'He lives with grandma first, and then with his parents', both *past* and *current Living Arrangement* events should have the same trigger 'lives' but is not allowed. In future work, we plan to modify the annotation scheme to allow multiple events of the same type associated with the same trigger. Some downstream clinical research may need even more fine-grained annotation.

5. Conclusion

In this work, we present a novel corpus, PedSHAC, annotated for SDoH. Our corpus has 1,260 social history sections of pediatric patients annotated across 10 SDoH event types. We envision such fine-grained annotation on multiple critical SDoH types can help the research community study the impact of SDOH on other child health outcomes. We explored LLM-based IE across multiple dimensions, including pre-trained architectures – mSpERT, Flan-T5, and GPT-4; learning strategies – fine-tuning and in-context methods; and prompting approaches – one-step text-to-event and two-step QA. Our results demonstrate that detailed SDoH representations can be extracted from pediatric narratives with performance

comparable to human annotators, providing an automatic approach for incorporating valuable SDoH information in clinical and research applications.

Future work for the corpus development could include addressing the current limitations, through actual user studies to pinpoint the needs and possibly expanding the current SDoH annotation to encompass more hospital systems and pediatric subpopulations. We also plan to explore other IE approaches such as (1) using effective data selection strategies such as active learning (Lybarger et al., 2021) in the annotation phase could help save annotation costs, (2) GPT-4 prompt-tuning including the involvement of medical experts, automatic prompt generation (Zhou et al., 2022), and self-verification (Weng et al., 2022) to improve the response quality.

Our proposed automatic IE approaches allow extracted SDoH information to be directly incorporated in EHRs in a tabular form, we envision our work to help downstream clinical applications through better quantifying the presence of various SDoHs in pediatric populations.

6. Acknowledgements

This work was supported by the National Institutes of Health (NIH) - National Cancer Institute (Grant Nr. 1R01CA248422-01A1 and 1R21CA258242-01) and National Library of Medicine (Grant Nr. 2R15LM013209-02A1). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

7. Ethics

We obtained the necessary approvals from our institution's Institutional Review Board (IRB), with a waiver of patient content on using their clinical notes. SDoH sections in clinical notes may contain Protected Health Information (PHI) including potentially identifiable information, like names, occupations, contact information, and other identifiers. All researchers and annotators received the necessary human subjects training to interact with patient data, including PHI. We used secured servers to ensure data security. Our GPT-4 experiments were conducted on a Health Insurance Portability and Accountability Act (HIPAA)-compliant Azure environment and ensured that no queries would be recorded by OpenAI. The patient populations in our corpus may not the representative of populations at other institutions or the broader population, which may inadvertently bias our extraction models and impact the generalizability. We believe our work will benefit the practice of automatic SDoH IE from pediatric narratives, as well as the general domain IE through LLMs.

Bibliography

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on EMNLP*, pages 1998–2022, Abu Dhabi, United Arab Emirates. ACL.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. ACL.

Raid Alzubi, Hadeel Alzoubi, Stamos Katsigiannis, Daune West, and Naeem Ramzan. 2022. Automated detection of substance-use status and related information from clinical text. *Sensors*, 22(24):9609.

Tamara E. Baer, Emily A. Scherer, Eric W. Fleegler, and Areej Hassan. 2015. Food insecurity and the burden of health-related social problems in an urban youth population. *J Adolesc Health*, 57(6):601–607.

Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, et al. 2018. 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*, 25(1):61–71.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on EMNLP and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. ACL.

Anusha Bompelli, Yanshan Wang, Ruyuan Wan, et al. 2021. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. *Health Data Sci*, 2021.

David S Carrell, David Cronkite, Roy E Palmer, Kathleen Saunders, David E Gross, Elizabeth T Masters, Timothy R Hylan, and Michael Von Korff. 2015. Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inform.*, 84(12):1057–1064.

Centers for Disease Control and Prevention. 2022. Social determinants of health at CDC.

Hyung Won Chung, Le Hou, Shayne Longpre, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. 2008. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, 15(1):36–39.

Neal A DeJong, Charles T Wood, Madlyn C Morreale, Cameron Ellis, Darragh Davis, Jorge Fernandez, and Michael J Steiner. 2016. Identifying social determinants of health and legal needs for children with special health care needs. *Clin Pediatr (Phila)*, 55(3):272–277.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

Cheryl A Dickson, Berrin Ergun-Longmire, Donald E Greydanus, Ransome Eke, Bethany Giedeman, Nikoli M Nickson, Linh-Nhu Hoang, Uzochukwu Adabanya, Daniela V Pinto Payares, Summer Chahin, et al. 2023. Health equity in pediatrics: Current concepts for the care of children in the 21st century (dis mon). *Disease-a-Month*, page 101631.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *24th ECAI*.

Caitlin A Farrell, Eric W Fleegler, Michael C Monuteaux, Celeste R Wilson, Cindy W Christian, and Lois K Lee. 2017. Community poverty and child abuse fatalities in the united states. *Pediatrics*, 139(5).

Daniel J Feller, Jason Zucker, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. 2018. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. In *AMIA Annu Symp Proc*, volume 2018, page 422. AMIA.

Arvin Garg, Brian Jack, and Barry Zuckerman. 2013. Addressing the social determinants of health within the patient-centered medical home: lessons from pediatrics. *JAMA*, 309(19):2001–2002.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS One*, 13(2).

Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.

Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. 2013. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among us veterans. In *AMIA Annu Symp Proc*, volume 2013, page 537. AMIA.

Sifei Han, Robert F Zhang, Lingyun Shi, et al. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *J Biomed Inform*, 127:103984.

Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, et al. 2019. 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*, 7(3):e13802.

Karen Ho, Randi Zlotnik Shaul, Lee Ann Chapman, and Elizabeth Lee Ford-Jones. 2016. Standard of care in pediatrics: Integrating family-centred care and social determinants of health. *Healthc Q*, 19(1):55–60.

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

Anne E Kazak, Annah N Abrams, Jaime Banks, Jennifer Christofferson, Stephen DiDonato, Martha A Grootenhuis, Marianne Kabour, Avi Madan-Swain, Sunita K Patel, Sima Zadeh, et al. 2015. Psychosocial assessment as a standard of care in pediatric cancer. *Pediatr Blood Cancer*, 62(S5):S426–S459.

Karen A Kuhlthau and James M Perrin. 2001. Child health status and parental employment. *Arch Pediatr Adolesc Med*, 155(12):1346–1350.

Gaia H Linfield, Shyam Patel, Hee Joo Ko, Benjamin Lacar, Laura M Gottlieb, Julia Adler-Milstein, Nina V Singh, Matthew S Pantell, and Emilia H De Marchis. 2023. Evaluating the comparability of patient-level social risk data extracted from electronic health records: A systematic scoping review. *J. Health Inform.*, 29(3):14604582231200300.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the ACL and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. ACL.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023a. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 30(8):1389–1397.

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

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023b. The 2022 n2c2/UW shared task on extracting social determinants of health. *J Am Med Inform Assoc*, 30(8):1367–1378.

Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 2023a. DICE: Data-efficient clinical event extraction with generative models. In *Proceedings of the 61st Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 15898–15917, Toronto, Canada. ACL.

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*.

Jennifer Morone. 2017. An integrative review of social determinants of health assessment and screening tools used in pediatrics. *J Pediatr Nurs*, 37:22–28. Special Issue: Social Determinants of Health.

Amol S Navathe, Feiran Zhong, Victor J Lei, Frank Y Chang, Margarita Sordo, Maxim Topaz, Shamkant B Navathe, Roberto A Rocha, and Li Zhou. 2018. Hospital readmission and social risk factors identified from physician notes. *Health Serv. Res.*, 53(2):1110–1136.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Braja G Patra, Mohit M Sharma, Veer Vekaria, et al. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *J Am Med Inform Assoc*, 28(12):2716–2727.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*, 21(1):5485–5551.

Russell Richie, Victor M Ruiz, Sifei Han, Lingyun Shi, and Fuchiang Tsui. 2023. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. *J Am Med Inform Assoc*, page ocad046.

Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *J Am Med Inform Assoc*, page ocad071.

Guergana K Savova, Philip V Ogren, Patrick H Duffy, James D Buntrock, and Christopher G Chute. 2008. Mayo clinic nlp system for patient smoking status identification. *J Am Med Inform Assoc*, 15(1):25–28.

Paul Stallard, Philip Norman, Sarah Huline-Dickens, Emma Salter, and Jan Cribb. 2004. The effects of parental mental illness upon children: A descriptive study of the views of parents and children. *Clin Child Psychol Psychiatry*, 9(1):39–52.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the ACL*, pages 102–107, Avignon, France. ACL.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Ross Thompson, Paul H Dworkin, Georgina Peacock, Mary Ann McCabe, John K Iskander, Phoebe Thorpe, and Susan Laird. 2016. Addressing health disparities in early childhood.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, 15(1):14–24.

Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. 2015. Automated extraction of substance use information from clinical texts. In *AMIA Annu Symp Proc*, volume 2015, page 2121. AMIA.

Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. 2022. Large language models are reasoners with self-verification. *arXiv preprint arXiv:2212.09561*.

Jinge Wu, Rowena Smith, and Honghan Wu. 2022a. Adverse childhood experiences identification from clinical notes with ontologies and nlp.

Jinge Wu, Rowena Smith, and Honghan Wu. 2022b. Ontology-driven self-supervision for adverse childhood experiences identification using social media datasets.

Qian-Wen Xie, Xiangyan Luo, Roujia Chen, and Xudong Zhou. 2023. Associations between parental employment and children's screen time: A longitudinal study of china health and nutrition survey. *Int. J. Public Health*, 67:1605372.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 171–181. Springer.

Zehao Yu, Xi Yang, Chong Dang, et al. 2021. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. In *AMIA Annu Symp Proc*, volume 2021, page 1225.

Zehao Yu, Xi Yang, Yi Guo, et al. 2022. Assessing the documentation of social determinants of health for lung cancer patients in clinical narratives. *Front Public Health*, 10.

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