Evaluating and Enhancing Large Language Models (LLMs) in the Clinical Domain

Yujuan Fu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Meliha Yetişgen, Chair

Fei Xia

Wen-Wai Yim

Trevor A. Cohen

Su-In Lee

Program Authorized to Offer Degree:
Biomedical Informatics

University of Washington

Abstract

Evaluating and Enhancing Large Language Models (LLMs) in the Clinical Domain

Yujuan Fu

Chair of the Supervisory Committee:
Meliha Yetişgen
Biomedical and Health Informatics

Recent advancements in large language models (LLMs) have demonstrated human-level performance on many specialized medical tasks, even without annotated training data. However, three main challenges remain: (1) due to the sensitive and highly specialized nature of clinical narratives, as well as the high cost of human expert annotation, there is a lack of high-quality, well-structured, and clinically meaningful datasets for LLM training and evaluation; (2) current medical LLMs show limited generalization ability to interpret and extract complex clinical information on certain unseen natural language understanding (NLU) tasks; and (3) as LLMs are typically trained on vast amounts of data, there is a substantial risk of data contamination, where evaluation benchmarks unintentionally overlap with training data, leading to inflated test performance and potentially reduced performance on truly novel tasks.

In this work, we address these limitations through three core aims: (1) develop benchmark datasets for clinical information extraction (IE), a key NLU subtask, across two critical medical domains, and evaluate the performance of multiple state-of-the-art (SOTA) transformer-based language models (LMs), under both fine-tuning and in-context learning settings; (2) develop a more generalizable medical NLU model via instruction tuning, demonstrating enhanced performance on previously unseen clinical NLU datasets; and (3) systematically review existing detection approaches for data contamination and evaluate those approaches

on datasets used during pre-training and fine-tuning LLMs, with our own and three other widely used open-source LLMs.

In summary, our work contributes to the development of both clinical benchmarks and robust LLMs, as well as highlighting the ongoing challenges in benchmarking LLMs' generalizability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AGENT-STYLE LARGE LANGUAGE MODELS (LLM AGENTS): LLMs enhanced with advanced capabilities such as memory retrieval, strategic planning, collaboration with other models, and tool usage.

AUTO-REGRESSIVE TEXT GENERATION: A generation text approach in LMs where each token is produced sequentially by conditioning on all previously generated tokens, enabling the model to capture logical dependencies and maintain coherence in the generated text.

BENCHMARKS: Standardized datasets and scoring metrics used to measure how well a system performs on a specific task, providing a consistent basis for evaluation and comparison.

BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT): An encoder-only, transformer-based language model known for its strong performance on natural language understanding tasks.

CHAIN-OF-THOUGHT (COT) PROMPTING: A technique that guides language models to generate intermediate reasoning steps before producing a final answer, thereby improving performance on complex reasoning tasks.

CLINICAL DECISION SUPPORT: A set of tools or systems designed to assist healthcare professionals by providing relevant, evidence-based insights.

CLINICAL REPORT GENERATION: The task of producing structured or summarized clinical documentation from raw data, such as doctor-patient conversations or lengthy clinical notes.

DATA CONTAMINATION: The overlap between an LM's training and evaluation datasets.

ELECTRONIC HEALTH RECORD (EHR): A computer system for storing and managing important patient information documented during clinical care, including medical history, test results, diagnoses, treatments, and medications.

FINE-TUNING: A supervised training process that adapts a pre-trained language model to a specific task or dataset by updating its parameters using labeled data, enabling specialized predictions such as classification or sequence labeling.

IN-CONTEXT LEARNING (ICL): A technique where language models learn to perform tasks by conditioning on a few examples provided within the input prompt, without additional parameter updates or fine-tuning.

INFORMATION EXTRACTION (IE): An NLP task that transforms unstructured clinical narratives into structured labels, enabling efficient and large-scale data analysis.

INSTRUCTION FINE-TUNING: The process of fine-tuning a pre-trained LLM on a diverse set of tasks, datasets, and prompt formats to improve its ability to follow instructions and generalize to unseen tasks and domains.

INTER-ANNOTATOR AGREEMENT (IAA): A measure of consistency among different human annotators labeling the same data, indicating the reliability and quality of the annotation process.

LANGUAGE MODEL (LM): A computational model that assigns probabilities to sequences of words, enabling tasks such as text generation, prediction, and understanding.

LARGE LANGUAGE MODEL (LLM): A type of LM that is scaled up significantly in model parameter size and trained on vast datasets, enabling enhanced capabilities in generating, predicting, and understanding language with improved accuracy and generalization across diverse tasks.

MASKED LANGUAGE MODELING (MLM): A pre-training objective where a random subset of input tokens is masked, and the model learns to predict these masked tokens based on surrounding context. Also known as the *Cloze task*, it leverages the distributional hypothesis that words occurring in similar contexts have related meanings.

MULTIPLE-CHOICE QUESTION ANSWERING (MCQA): A task that involves selecting the correct answer from a set of candidate options given a question and supporting context.

NAMED ENTITY RECOGNITION (NER): An NLP task that labels meaningful text spans (named entities) with their corresponding types, such as persons, locations, or organizations.

NATURAL LANGUAGE GENERATION (NLG): The task of generating human-like text from structured or unstructured input representations.

NATURAL LANGUAGE INFERENCE (NLI): A task that involves determining the logical relationship between two sentences, typically whether one sentence entails, contradicts, or is neutral with respect to the other.

NATURAL LANGUAGE PROCESSING (NLP): A field at the intersection of linguistics, computer science, and artificial intelligence focused on enabling machines to understand, interpret, generate, and manipulate human language.

NATURAL LANGUAGE UNDERSTANDING (NLU): A subfield of NLP that focuses on interpreting and extracting meaningful data from unstructured text.

NEXT SENTENCE PREDICTION (NSP): A binary classification pre-training task in which the model is given pairs of sentences and trained to determine whether the second sentence logically follows the first in the original text. NSP helps the model understand inter-sentence relationships, benefiting downstream tasks like question answering and natural language inference.

PRE-ANNOTATION: The process of using a baseline system to generate initial ground-truth labels of varying quality, which are subsequently reviewed and corrected by human annotators.

PRE-TRAINING: The initial training phase, where an LM like BERT learns from large-scale unlabeled text corpora using self-supervised objectives to capture language patterns before fine-tuning on specific tasks.

REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF): A training technique that uses human feedback to guide the reinforcement learning process, aligning the model's outputs with human values and expectations to produce safer and more helpful responses.

RELATION EXTRACTION (RE): An NLP task that identifies and classifies semantic relationships between recognized entities in text.

RETRIEVAL-AUGMENTED GENERATION (RAG): A method that enhances language model generation by incorporating relevant information retrieved from an external knowledge base into the input context, improving performance especially in domains requiring up-to-date or specialized knowledge.

SCALING LAW: An empirical relationship showing that LLM performance improves predictably as model size, amount of training data, and diversity of fine-tuning tasks increase, enabling larger models to learn more complex semantic patterns.

SEMANTIC TEXT SIMILARITY: The task of assessing how closely two sentences align in meaning by assigning a semantic similarity score.

SENTENCE AND DOCUMENT CLASSIFICATION: The task of assigning a meaningful label to a text snippet or document.

TOKENIZATION: The process of splitting words into one or more tokens from the fixed vocabulary. This method enables models to effectively handle rare or out-of-vocabulary words by breaking them down into smaller, recognizable units.

TRANSFORMER: A deep neural network architecture introduced by Vaswani et al. [305]. It is built around a self-attention mechanism that captures contextual relationships among tokens in a sequence, regardless of their distance. The use of multi-head attention enables parallel computation and scalability, making transformers well-suited for large-scale natural language processing tasks.

VOCABULARY: A fixed set of tokens or subword units that an LM recognizes and uses to encode text. The vocabulary serves as the building blocks for representing input text during model training and inference.

ACKNOWLEDGMENTS

Pursuing a PhD has been a challenging yet incredibly rewarding journey for me. I could not have accomplished it without all the support of the wonderful people I've met along the way.

First and foremost, I would like to express my deepest gratitude to my amazing advisor, Prof. Meliha Yetişgen. I began my journey into NLP when I joined her lab, and she is the most knowledgeable, supportive, and kind advisor, who always encourages me to believe in myself. She guided me through every aspect of the academic journey: keeping up with the literature, conducting rigorous experiments, writing and presenting papers, navigating the review process as both an author and a reviewer, and organizing seminars. She is especially outcome-oriented and highly collaborative—she connected me with many brilliant co-advisors and collaborators, and taught me how to manage large-scale projects, meet deadlines, and lead successful research efforts. I feel incredibly fortunate to have been her student.

Besides my advisor, I am also grateful to the other professors who advised me throughout my PhD, in the order I met them: Prof. Özlem Uzuner, Prof. Kevin Lybarger, Dr. Wen-wai Yim, Prof. Fei Xia and Dr. Asma Ben Abacha. I remember how my journey evolved from designing annotation guidelines and managing annotation teams, to running baseline experiments, developing evaluation metrics, and eventually leading my own projects in literature review, system development, and result analysis. I still recall those weekly meetings where we carefully went through every detail of a project. I deeply appreciate the valuable feedback I received during those times - it's amazing to see how far our manuscripts evolved from the first draft to the final version.

Finally, I also want to thank Prof. Trevor Cohen and Prof. Su-In Lee for generously

making time to serve on my committee, for all the valuable feedback they provided during my general and final exams.

To my fellow students and lab mates - thank you for being not just collaborators but true friends. We've learned from each other, supported one another, and grown together. In the order I met you, I want to thank Wesley Surento, Xian Su, Sitong Zhou, Giridhar Kaushik Ramachandran, Nicholas Dobbins, Weipeng Zhou, Namu Park, Arjun Chakraborty, Changyuan Peter Qiu, Zhaoyi Sun, Zixuan Avery Yu and Sihang Zeng.

Finally, I want to thank my family—my father, mother, and grandparents. You were my first teachers and have given me unconditional support since I was a little girl. To my fiancé and best friend, Shirui Chen - thank you for encouraging me to pursue a career in NLP and for sharing both our happiest and most difficult moments.

Chapter 1

INTRODUCTION

1.1 Context and Motivation

Electronic Health Record (EHR) systems provide an effective means of storing and managing important patient information captured during the hospital course, documenting information such as medical history, test results, and medications [219]. According to a 2021 survey, 96% non-federal acute care hospitals have adopted a certified EHR system [218]. In addition to structured components, unstructured clinical narratives within the EHR system [77], are written by healthcare providers and capture nuanced patient information such as symptoms [365], diagnostic hypotheses, and treatment choices. Thus, understanding those unstructured clinical narratives is essential for treatment decision making [128], patient care management, quality insurance, and secondary clinical research [265].

Despite their benefits, the implementation of EHR systems has also been associated with notable drawbacks, most notably the extensive administrative demands and mental strain they impose on clinicians [337]. These pressures have contributed to severe burnout among care providers, increasing the risks of depression, substance abuse, and early career departure [337]. Recent advances in natural language processing (NLP) systems, particularly through large language models (LLMs), have shown the potential to mitigate these burdens by streamlining clinical workflows. For example, automatic clinical text summarization [304], diagnosis and treatment recommendation [331], and automatic error correction [4], are promising applications to reduce the clerical workload and cognitive fatigue associated with EHR usage. Beyond clinical workflows, these LLMs also show promise in secondary applications such as medical text simplification for the general public [91] and automated disease surveillance [221].

Given the growing importance of clinical LLMs, rigorous evaluation is critical to ensure their safety, reliability, and ethical deployment [267]. During this process, benchmark datasets provide a consistent and objective framework to assess LLM performance across standardized tasks. These benchmarks not only enable meaningful comparisons between LLMs, but also help identify their strengths and weaknesses, thereby guiding researchers in system improvement and development.

1.2 Objectives and Contributions

In this work, we focus on evaluating and enhancing LLMs for clinical applications, with particular focus on information extraction (IE), medical natural language understanding (NLU), and data contamination detection among evaluation benchmarks.

More specifically, our contributions can be summarized as the following three items:

1. Clinical notes often contain detailed, unstructured descriptions of patient conditions written by healthcare providers. To allow processing such rich information at scale and in real time, we apply IE methods that convert text into structured data. In this work, we focus on developing benchmark data sets and IE systems for two important domains: pediatric social determinants of health (SDoH) [87] and cancer [89]. In both domains, our IE systems reach performance close to human inter-annotator agreement (IAA). The IE systems and benchmark datasets described in this chapter are based on work presented in our prior publications [87, 89].

2. Medical NLU is an important NLP application that interprets and extracts meaningful information from unstructured texts [1]. We introduce a unified format for span-based and multiple-choice NLU tasks, including IE. Using this unified format, we then fine-tune a biomedical LLM on a new instruction dataset built from multiple publicly available medical sources. Our fine-tuned system, BioMistral-NLU, achieves strong generalization and outperforms other baseline models on unseen medical NLU benchmarks [90]. The corresponding chapter is reproduced from our prior publication [90].

3. Although LLMs have shown great performance on various benchmark tasks, there have been concerns about the reliability of those evaluation results, due to data contamination, where evaluation data overlap with training data. We review 50 studies, identify common assumptions behind contamination detection methods, and test three assumptions through case studies on 4 open-source LLMs, including our BioMistral-NLU. Our findings show that all tested assumptions fail under data distribution shifts, suggesting the challenge of detecting direct data contamination. Our results also indicate that, given the vast amount of training data, current LLMs may learn from data distributions rather than memorize exact examples [88]. The corresponding chapter is reproduced from our prior publication [88].

1.3 Guide for the Reader

The remainder of this thesis is organized as follows:

Chapter 2. Background and Motivations: This chapter reviews the key literature that underpins our research agenda, covering four focal areas from Chapter 3 to Chapter 6

Chapter 3 and 4. Clinical Information Extraction (IE): Two Key Tasks: The two chapters investigate information extraction (IE) as a method for transforming unstructured clinical narratives into structured data, enabling downstream applications such as EHR management and risk prediction. Focusing on the pediatric (Chapter 3) and cancer (Chapter 4) domains, each chapter begins by outlining the clinical importance of IE, and then details our contributions, including the development of annotation schemas, the construction of domain-specific datasets, and the design of evaluation protocols to assess both data quality and system performance. We proceed to describe the deployment of state-of-the-art (SOTA) transformer-based language models, trained using supervised fine-tuning and in-context learning, tailored to IE tasks in these two clinical settings. Finally, we present experimental results demonstrating that our models achieve performance levels comparable to human inter-annotator agreement (IAA) across both datasets.

Chapter 5. BioMistral-NLU: Generalized System for Medical Natural Lan-

guage Understanding (NLU): This chapter begins by defining the task of medical NLU, highlighting its importance, and also identifying a key challenge: generalized LLMs often have a great performance gap to in-domain fine-tuned systems on those specialized medical NLU tasks. Motivated by this gap, we introduce a unified NLU framework that includes task formulation, a standardized NLU prompting format, and the creation of a novel instruction-tuning dataset, MNLU-Instruct. Based on this unified NLU framework, we describe the development of our instruction-tuned BioMistral system, including the training pipeline, evaluation setup, and baseline comparison systems. Finally, we present experimental results demonstrating that our system generalizes effectively to unseen standardized medical benchmarks, outperforming even larger models.

Chapter 6 Detecting Data Contamination in Medical Benchmarks: This chapter begins by defining the critical issue of data contamination in LLM benchmarks and highlighting its potential to impact model performance, thereby motivating our study. It then describes the methodology for collecting relevant literature and introduces a taxonomy of contamination types, which is categorized by level (instance-level vs. dataset-level) and contamination type (direct vs. indirect). Afterwards, the chapter provides rigorous categorization and mathematical definitions for those assumptions underlying various detection methods, with a focus on direct contamination, the most commonly studied category. It also reviews approaches targeting indirect and task-level contamination. Finally, it presents a case study that evaluates three core assumptions across four open-source LLMs. The results reveal that these assumptions frequently fail, especially under data distribution shifts, highlighting the need for better detection and regulation methods for data contamination.

Chapter 7 Conclusion and Future Work: The final chapter summarizes the key contributions of this thesis, reflects on limitations, and outlines future directions.

Appendix A NLP Background: This chapter in the appendix provides some background knowledge for clinical NLP, covering key applications, commonly used systems, and standard evaluation methodologies.

1.4 References

The chapters of this dissertation are adapted from my peer-reviewed, first-author publications listed below:

- Chapter 3 : adapted from Fu et al. [87];
- Chapter 4 : adapted from Fu et al. [89];
- Chapter 5 : adapted from Fu et al. [90];
- Chapter 6 : adapted from Fu et al. [88].

The literature review in Chapter 2 synthesizes and contextualizes findings from all of the above works.

Chapter 2
BACKGROUND AND MOTIVATIONS

In this chapter, we survey the background literature that motivates each of our contributions: the two clinical information-extraction tasks presented in Chapters 3 and 4, the generalized medical NLU system detailed in Chapter 5, and the data-contamination detection framework explored in Chapter 6.

2.1 Extracting Social Determinants of Health from Pediatric Patient Notes

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) [84]. 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 [291, 67]. Clinicians have continuously adapted practices by systematically gathering pediatric patients' SDoH during clinical consultations [93, 116, 148]. 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 [212].

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 [174]. 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) [290], 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 Chapter 3, we will 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 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.

As our work of PedSHAC encompasses both dataset curation and system development, the subsequent sections situate our contributions within the broader landscape of SDoH corpora and SDoH IE methodologies.

2.1.1 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 a singular focus, such as smoking status [299, 259], homelessness [104, 29], and substance use [317, 342, 47, 13]. Previous research also addresses SDoH factors in specific contexts, such as sexual health [80] and hospital readmission rate [213]. Our prior SDoH work investigated adult SDoH factors using a fine-grained, event-based annotation scheme

encompassing detailed status and type labels for adults [189].

Pediatric SDoH factors such as adverse childhood experiences were researched in the adult patient population [29, 328, 327]. 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 [61] and food insecurity [23] are especially important to pediatric patients. The clinical notes of pediatric patients may describe employment associated with patient caregivers [156, 334]; at the same time, patient parents' mental health [277] become important as pediatricians continually evaluate whether children may be at risk for child abuse and neglect [79]. PedSHAC bridges this gap in the literature with comprehensive fine-grained annotation of SDoH determinants with a focus on pediatric patients.

In the following sections, we will describe the related work on the relevant corpora and clinical IE systems.

2.1.2 SDoH IE Methods

SDoH IE is an increasingly explored task, and the modeling approaches range from manually curated rules [235, 111], traditional/shallow machine learning models [55, 317], neural networks [29, 94], to transformer-based LLMs [235, 37].

Bidirectional Encoder Representations from Transformers (BERT) [66] is frequently used in SDoH extraction tasks for text classification [349, 350, 108] and entity and relation extraction [249, 190]. Sequence-to-sequence approaches that utilize generative LLMs, like Text-to-Text Transfer Transformer (T5) [242], have also achieved high performance [251]. The most recent generation of LLMs, such as GPT-4 [226], are pre-trained on large amounts of data and instruction-tuned [229], 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 [6, 338].

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.

2.2 Extracting Medical Problem and Drug Information from Oncology Notes

Clinical notes capture detailed descriptions of patient status and disease progression from the care provider's perspective through unstructured text [161]. In oncology, these narratives are comprehensive and cover diverse symptoms [65], multiple drug cycles, and side effects [261]. Such unstructured notes contain valuable information that complements structured data in electronic health records (EHRs) [77], such as symptoms [365], diagnostic hypotheses, and treatment decisions. Understanding these clinical narratives is crucial for treatment decisions [128], patient management, and quality assurance.

Natural language processing (NLP) methods for information extraction (IE) can convert unstructured narratives into structured data [318], enabling large-scale, real-time use of those rich information in clinical decision support applications and generation of real-world evidence in learning health systems. High-performing IE models require advanced techniques and annotated datasets for training and evaluation. Existing research has produced systems that extract information about cancer diagnoses and treatments [60]; however, a significant gap remains. Specifically, existing frameworks do not fully capture the relationships between cancer diagnoses, symptoms, and medications. This gap highlights the need for a unified approach for characterizing medical problems and drug information, including their interconnections. Chapter 4 details our work to address this gap.

In the following sections, we will describe the related work on the relevant corpora and clinical IE systems.

2.2.1 Medical Problem and Drug Data Sets

Clinical IE includes a variety of classification tasks, including text classification, relation extraction (RE), and event extraction (EE) [318]. We utilize EE to characterize medical problems and drugs, and RE to determine the relationships between them. Each event includes a trigger representing a clinical concept (problem or drug) and fine-grained attributes (e.g., assertion or anatomy) [60]. Previous research on event-based medical problem extraction either (1) focuses on a subset of problems, such as symptoms [365] or diseases and disorders [240], or (2) lacks granular annotation. For instance, the 2010 i2b2/VA challenge [301] only included assertion attributes for medical problems (present vs. absent), without detailing severity or anatomy. Cancer-focused EE studies [354, 56, 40] often overlook key factors, like symptoms (excluding pain) [113] and comorbidities [239], which are essential for understanding diagnosis and treatment. We present a comprehensive annotation schema that captures all medical problems [188, 296, 365].

The clinical relationships between drugs and medical problems inform treatment and diagnosis, but are often complex. Most existing literature narrowly focuses on a subset of possible relations, including adverse drug events [114, 126] and clinical temporal relationships [283, 308, 32] in clinical notes and gene-cancer interactions in biomedical literature [147, 43]. The 2010 i2b2/VA challenge annotated six detailed relations between medical problems and treatments in discharge summaries, which we use to characterize interactions between medical problems and drugs in clinical narratives of oncology notes.

2.2.2 Clinical IE Approaches

IE allows for the secondary use of clinical narratives in clinical and translational research [318], as well as near real-time EHR clinical decision support functionalities. It can enhance understanding of drug discontinuation, symptom monitoring, and adverse event management [11, 68, 173, 217]. Most clinical IE approaches employ separate models for event and relationship extraction in multi-step processes. These models have progressed from rule-

based systems [18, 260, 275] and feature-engineered models [58, 59, 255] to neural networks, culminating in transformer architectures [305, 160]. Bidirectional Encoder Representations from Transformers (BERT) [66], an encoder-only model, has shown superior performance in clinical IE [160]. Clinical variants of BERT [12] have achieved high performance in clinical IE tasks, like drug-problem RE [252]. Some BERT-based architectures, like Packed Levitated Marker (PL-Marker) [341], adopt a two-step approach for extracting spans (arguments) and relationships (argument roles), including IE for medical imaging reports [234]. BERT architectures with multiple output layers, like Span-based Event and Relation Transformer (SpERT) [76], can jointly extract spans (arguments) and relationships (argument roles), including the extraction of social determinants of health [190].

Recent progress in generative language models (GLMs) includes encoder-decoder models, like the Fine-tuned Language Net Text-To-Text Transfer Transformer (Flan-T5) [54], and decoder-only models, like the Large Language Model Meta AI (LLaMA) [293] and Generative Pre-trained Transformers 4 (GPT-4) [226]. GLMs have set new performance standards in various clinical benchmark tasks. Smaller models like T5 [241] have been fine-tuned for specialized tasks such as medical document classification, named entity recognition (NER) [181], and medical database query generation [70]. GLMs with billions or trillions of parameters, such as GPT-4, are called Large Language Models (LLMs) and excel at in-context learning (ICL) [362]. ICL allows these models to adapt to tasks based solely on prompts without changing their parameter weights and has been successfully applied to clinical IE tasks [243, 87, 118].

2.3 Generalized Medical NLU System

Medical Natural Language Understanding (NLU) encompasses a variety of tasks aimed at enabling machines to understand, interpret, and respond to clinical and biomedical language. Within this domain, research has focused on tasks like Information Extraction (IE) and Document Classification (DC) [330]. To create a more holistic understanding of medical NLU, two major benchmark datasets have been curated: the Biomedical Language

Understanding Evaluation (BLUE) [236] and the Biomedical Language Understanding and Reasoning Benchmark (BLURB) [102]. These datasets, which include a wide range of medical NLU tasks, are widely used to assess the capabilities of various LLMs in medical contexts [81, 320, 49].

The evaluation of medical foundation models focuses on complex medical tasks, such as medical exams (MedQA dataset) [253, 14], document summarization [253, 335]. Those tasks require medical knowledge and reasoning abilities, beyond simple medical NLU. However, when evaluated in simple NLU tasks, medical foundation models can fall short to smaller-scale fine-tuning-based approaches. For example, GPT-4 falls short compared to traditional NLP models in SDoH event extraction [243], and Me-LLaMA underperforms relative to traditional BERT-based methods in zero-shot named entity recognition (2010 i2b2) [301] and relation extraction (2013 DDI) [115]. Our experimentation on the Pediatric SDoH and CACER datasets from Section 3.2 4.2, also supports this observation. These shortcomings are often attributed to errors in adhering to the desired output format and misinterpreting human intentions within the highly specialized medical domain. This highlights the necessity for targeted Instruction tuning of LLMs in medical NLU.

In this section, we review the literature on generalized medical NLU systems, with particular emphasis on *instruction-tuning*, a widely adopted strategy for developing generalized LLMs. Building on the gaps identified in the prior literature, Chapter 5 presents our research on a generalized medical NLU framework.

2.3.1 Generalized System for Medical NLU

Recent studies have examined how generalized LLMs can perform medical NLU tasks without specific training. For instance, Agrawal et al. (2022) [7] demonstrated the potential of LLMs for clinical NLU tasks through few-shot ICL. Hu et al. (2023) [122] evaluated the performance of ChatGPT on clinical Named Entity Recognition (NER) tasks, using a subset of NLU datasets. Similarly, Wang et al. (2023) [320] proposed a new prompting strategy for multiple clinical NLU tasks using proprietary LLMs like ChatGPT [2] and GPT-4 [5]. However,

their evaluations were limited to small sample sizes from the BLUE benchmark. Chen et al. (2023) [49] and Feng et al. (2024) [81] also evaluated several LLMs using the BLURB benchmark. While ChatGPT and GPT-4 outperformed other LLMs, they significantly lagged behind fine-tuned, in-domain systems. This performance gap underscores the need for more generalized systems tailored to medical NLU.

2.3.2 Instruction Tuning for Medical NLU

Instruction tuning involves fine-tuning pre-trained LMs on a wide array of instruction-following tasks, enabling the model to understand and respond to natural language instructions. This process helps LLMs generalize to unseen tasks in zero-shot or few-shot settings [54, 229]. Instruction-tuning datasets typically cover a range of tasks such as reasoning, question-answering, dialogue, and summarization [359]. Instruction tuning has been explored in general domain NLU tasks, such as Information Extraction (IE) [316, 134, 258, 310, 183] and NER [366, 361]. These models have shown promise in generalizing to diverse tasks by leveraging a unified instruction format. However, to date, there has been limited adaptation of such generalized systems to the medical domain.

Several studies have made strides toward adapting instruction tuning for medical NLU tasks, primarily focusing on dialogue-based systems like ChatDoctor [352] and MedAlpaca [109]. Additionally, medical foundation LLMs such as MedGemini [253] and Taiyi [186] show potential across various NLU tasks, though comprehensive evaluations are still lacking. Many prior systems have focused on specific medical NLU tasks, such as Table Question Answering (QA) [187], NER [361], and Information Extraction [258], with little exploration of tasks like sentence similarity or natural language inference (NLI). As of now, there is no unified system capable of generalizing across all critical medical NLU tasks. In this work, we aim to fill this gap by evaluating our system in a zero-shot setting, using two well-established benchmarks that cover seven essential medical NLU tasks.

2.4 Data Contamination

Recent advances in LLMs have led to impressive results across a wide range of natural language processing tasks, positioning them as powerful tools for general-purpose applications [5, 20]. However, one major challenge that undermines fair and accurate evaluation of LLMs is the issue of **data contamination**, where overlap exists between an LLM's training corpus and its evaluation benchmarks [25]. Such overlap can lead to artificially high scores, creating the false impression that a model has effectively generalized to new data [25, 256, 168].

To address these concerns, researchers have proposed various techniques to detect contamination in language model training data. These techniques are not only important for evaluating model performance fairly, but also for identifying the presence of proprietary or sensitive content in training corpora [336, 205].

Despite the emergence of these methods, all existing contamination detection techniques rely on specific assumptions—about the model's access, the structure of the training and evaluation data, or the nature of language generation—that may not generalize across all contexts[1]. While prior surveys have discussed detection strategies or mitigation techniques, none have comprehensively examined or validated the foundational assumptions these methods depend on [336, 123, 121].

To address the gaps outlined above, Section 6 will present our work on a systematic categorization of the current research on data contamination and an empirical assessment of the effectiveness of these approaches.

[1]This review includes contamination detection approaches applicable to both large and small language models.

Chapter 3

PEDSHAC: SOCIAL DETERMINANTS OF HEALTH FROM PEDIATRIC PATIENT NOTES

In this chapter, we present our novel pediatric SDoH corpus, **Ped**iatric **S**ocial **H**istory **A**nnotation **C**orpus (**PedSHAC**), which 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].

3.1 Methods

In this section, we will describe the details of our dataset curation and system development.

3.1.1 Creating the PedSHAC Corpus

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

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

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. Clinical notes in PedSHAC are well-represented across different age groups, emphasizing early childhood and adolescent cohorts. PedSHAC's patient age distribution is visualized in Figure 3.1.

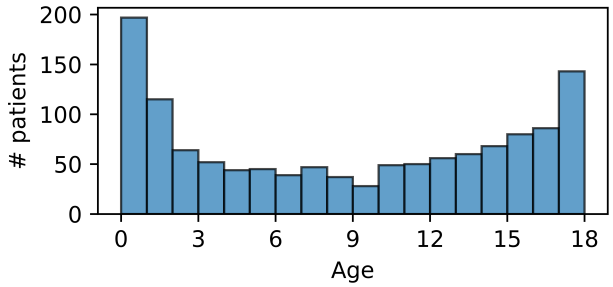

Figure 3.1: Patient age distribution in the PedSHAC corpus.

Annotation Scheme

We created detailed annotation guidelines for ten SDoH (referred to here as *event types*), as listed in Table 3.1. 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 defined by a trigger and a set of arguments that specify the event's SDoH category, current status, and subtype. 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 3.1. SDoH information was annotated using the BRAT rapid annotation tool [278]. Figure 3.2 is an example describing the patient's living arrangement and caregivers' employment.

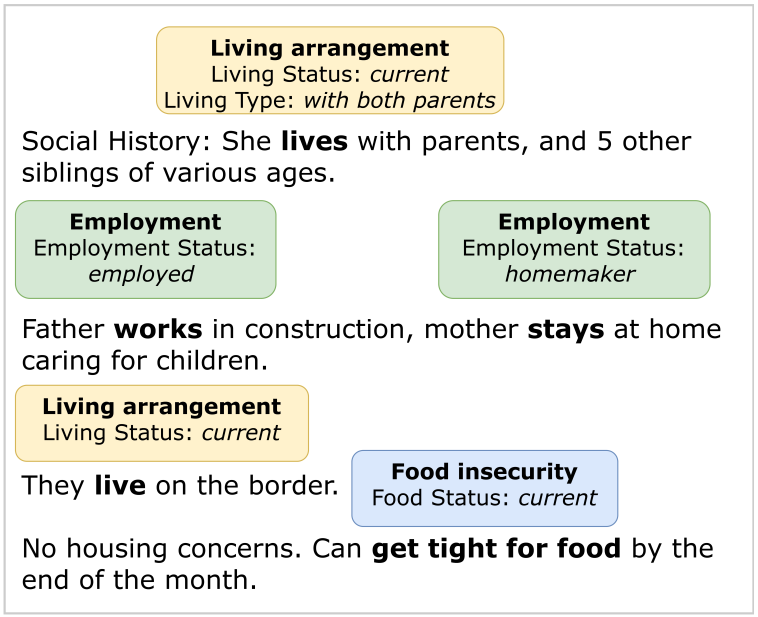

Figure 3.2: An annotation example: the triggers are in boldface. The box above a trigger shows the event type, arguments, and subtype labels.

[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 [191].

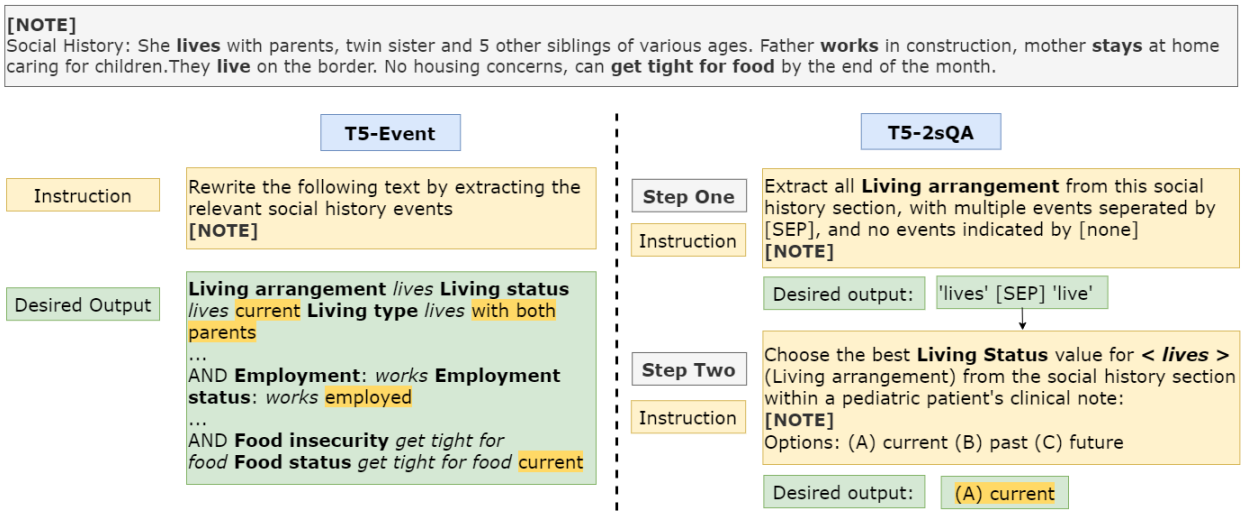

Figure 3.3: Our one-step (T5-Event) and two-step (T5-2sQA) extraction models. T5-Event extracts all SDoH events, including triggers and arguments, in one query. T5-2sQA extracts triggers and arguments in separate queries, where Step Two includes the predicted triggers from Step One.

IE Evaluation

We follow the previous N2C2 SDoH challenge [191] 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.

Event	Trigger & Arg.	Trigger examples & Argument subtypes	# labels			IAA
			Train	Validation	Test	F1
Adoption	Trigger	“adopted", ...	27	4	9	100.0
Education	Trigger	“5th grade" , “junior year", ...	227	35	74	80.0
Access	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 3.1: 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.

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 using the criteria in Section 4.1.3. 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 3.1 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.

3.1.2 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 [191], we use our high-performing, multi-label variation of the Span-based Entity and Relation Transformer model (mSpERT) [76, 190][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

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

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 [182, 193] demonstrates that entity and relation extraction tasks can be reformulated into text2event tasks using generative encoder-decoder models like T5 [242, 54] and decoder-only models like GPT-4 [226]. We map each event annotation to a structured text representation [251, 182][4]. Figure 3.3 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.

Fine-tuning T5 with Two-step QA Prompting (T5-2sQA): We utilize a two-step pipeline approach to first extract trigger spans [193] and then resolve subtype labels through multiple-choice questions [194]. Figure 3.3 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 [6]. 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.

---

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

The larger context window of recent LLMs, including GPT-4 [226], 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.

3.1.3 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 [12]. For T5 experimentation, we initialized from Flan-T5-Large (780M) [54], 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.

3.2 Results

3.2.1 Trigger and Argument Evaluation

Following the evaluation criteria described in Section 4.1.3, we report the extraction performance on the withheld PedSHAC test set in Table 3.2 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) [30] for all approaches, but only present a subset of significance testing results in Table 3.2 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.

[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 3.2: 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.

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 that 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.

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

3.2.2 Event-level Evaluation

We additionally assess performance using a more rigorous *event-level* evaluation criteria, which requires the equivalence (defined in Section 4.1.3) 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 3.3 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 3.2.1, yet exclude the results from Table 3.3 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 3.2 with Table 3.3, 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.

3.2.3 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 [129], both generative models (T5 and GPT-4) show a problem of hallucination [129], 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 [95], and constrained decoding [182].

3.3 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 [189] in the annotation phase could help save annotation costs, (2) GPT-4 prompt-tuning including the involvement of medical experts, more strategic selection of in-context examples (e.g. semantically similar instance from a trained text encoder [177]), automatic prompt generation [367], and self-verification [324], to further 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.

3.4 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.

Chapter 4

CACER: CLINICAL CONCEPT ANNOTATIONS FOR CANCER EVENTS AND RELATIONS

In this chapter, we focus on two cancer populations: prostate cancer and diffuse large B-Cell lymphoma (DLBCL). Prostate cancer is the second most common cancer among men and exhibits considerable heterogeneity [268, 245]. Some patients have a less aggressive, chronic form, whereas others face a highly aggressive disease linked to increased morbidity and mortality, necessitating more intensive treatments [245]. Similarly, DLBCL represents the most common aggressive lymphoma [285]. Together, these cancers exemplify the spectrum of chronic and aggressive cancer trajectories.

We introduce **C**linical Concept **A**nnotations for **C**ancer **E**vents and **R**elations (CACER), a novel corpus of cancer patient clinical oncology notes from Fred Hutch Cancer Center, with detailed annotations for medical problems, drugs, and their relationships. We benchmark this dataset with high-performing models and outline future directions. Key contributions include:

- We provide comprehensive, fine-grained annotations for 48k medical problem and drug events and 10k drug-problem and problem-problem relations. CACER will be made available to the research community pending institutional review board approval.

- We develop state-of-the-art extractors using BERT models and GLMs via supervised fine-tuning and ICL approaches. For GLMs, we explore various prompting strategies, including Question-Answering (QA). RE encompasses long-distance, inter-sentence relationships.

4.1 Methods

4.1.1 Creating the CACER Corpus

Data Set Collection

We used a clinical data set of outpatient records for two types of cancer, prostate cancer and DLBCL, from the Fred Hutch Cancer Center from 2015-2018. This data set includes 1,453 prostate cancer patients (with 10k notes) and 818 patients with DLBCL (with 11k notes). CACER consists of 575 clinical notes randomly sampled from this data set. Following de-identification, we randomly sampled notes containing over 30 lines and manually excluded clinical notes that were duplicates, EHR templates, or overlapped with other notes. We divided the data into training (400 notes), validation (60 notes), and test (115 notes) subsets, ensuring that no patient was included in multiple subsets. Appendix B.1 provides detailed dataset statistics and patient demographics.

Event Annotation Schema

The CACER annotation schema encompasses medical problem (*Problem*) and drug (*Drug*) events, along with the relations between them. It builds upon the schema used in our previous research on COVID-19 symptoms [188] and lung and ovarian cancer symptoms [296]. We expanded the symptom annotation guidelines to encompass all medical problems and drug events in the cancer domain.

Each *Problem* event is marked by a *trigger* and multiple attributes (i.e., *arguments*). The *Problem* trigger is a text span that most clearly and concisely expresses the medical problem. An argument has a name (i.e., 'argument type") such as *Duration* and a value that corresponds to a text span such as 'three months". For some argument types, such as *Assertion*, their values can be normalized into a pre-defined set of subtypes (e.g., 'present", 'absent"). For instance, the subtype 'present" might correspond to text spans such as 'is observed" or 'was found", or it is simply implied by the occurrence of a disease name. We call

the argument with a subtype label a '*labeled*" argument and the one without a '*span-only*" argument. Figure 4.1 provides examples for both types of arguments. *Drug* events only include a trigger, without any arguments. We will address the potential for incorporating more fine-grained drug annotations within oncology notes in the discussion section. Table 4.1 summarizes the event annotation schema. Our annotation guideline is available on the project's GitHub page.

Event	**Trigger & Arg.**	**Trigger examples & Argument subtypes**	**# labels**			**IAA F1**	
			Train	**Valid**	**Test**	**Without preAnn**	**With preAnn**
Drug	Trigger*	'ibuprofen', 'lupron', ...	11118	2104	3534	97.5	97.1
Problem	Trigger*	'cancer', 'vomitting'...	21453	3575	6555	89.6	96.5
	Assertion*	{present, hypothetical, absent, conditional, possible, not_patient}	21453	3575	6555	87.2	94.0
	Change	{worsening, no_change, improving, resolved}	1440	293	418	78.4	85.1
	Severity	{mild, moderate, severe}	775	168	254	84.6	89.8
	Anatomy	'prostate', 'back', ...	9880	1638	2877	82.9	91.9
	Characteristics	'recurrant', 'metastatic', ...	4749	830	1439	66.2	87.2
	Duration	'1 year', 'two weeks', ...	930	171	298	92.9	81.7
	Frequency	'every day', 'rarely', ...	245	35	79	92.3	69.6
Overall	-	-	72043	12389	22009	88.4	94.1

Table 4.1: Event schema, pre-annotation performance, and inter-annotator agreement (IAA) for the CACER dataset. The required trigger and arguments are labeled by *. A valid event must have all required arguments.

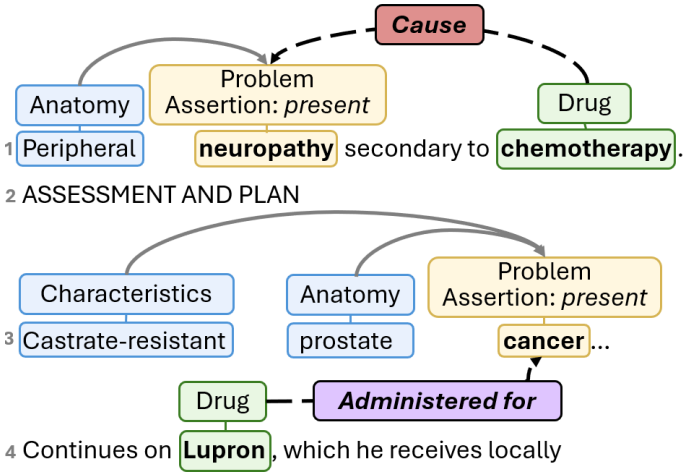

Figure 4.1: Annotation example from CACER. line 1 includes the intra-sentence relation, *Causes* and line 2-4 contains the inter-sentence relation, *administered for*. The inter-sentence relation is indicated by the note section subtitle, 'ASSESSMENT AND PLAN', linking the treatment to the main diagnosis.

Relation Annotation Schema

We use six *Drug-Problem* and *Problem-Problem* relation types from the 2010 i2b2/VA corpus [301], as shown in Table 4.2. Each relation includes a head, a tail, and a relation type. For *Drug-Problem* relations, the head is a *Drug* trigger, and the tail is a *Problem* trigger. For *Problem-Problem* relations, both the head and the tail are *Problem* triggers, where the tail *Problem* describes characteristics of the head *Problem* or identifies it as the cause or superclass of the tail.

Relation annotation requires either: (1) explicit linguistic cues indicating a relationship or (2) implicit relationships based on medical knowledge. Most relations are intra-sentence and are based on verb phrases like 'prescribed for.' Inter-sentence relations can be inferred by section headers, e.g., "Cancer Treatment ->... on <DATE>, Lupron is given." indicates an *AdminFor* relation between cancer and lupron. Otherwise, they are frequently implicit and

rely on general medical knowledge. A given *Problem* or *Drug* may be mentioned multiple times in a note, but only the closest pairs of triggers are annotated. We define the relation *context window* as the smallest continuous sequence of sentences that includes both head and tail triggers. In CACER, $< 1\%$ of the relations have a context window longer than five sentences.

Annotation is carried out using a web-based annotation tool, BRAT [279]. Example annotations are shown in Figure 4.1.

Relation	**Head entity**	**All relations**				**Intra-sentence relations**			
		# **labels**			**IAA**	# **labels**			**IAA**
		Train	**Valid**	**Test**	**F1**	**Train**	**Valid**	**Test**	**F1**
AdminFor	Drug	3715	762	1422	75.4	2456	373	783	86.9
NotAdminBecause		130	49	54	50.6	106	37	44	50.0
Causes		729	232	321	75.1	644	199	278	75.7
Improves		502	87	133	62.5	342	53	65	71.1
Worsens		257	68	121	43.0	199	52	70	47.2
ProblemInteracts WithProblem (PIP)	Medical problem	1257	259	350	54.1	1141	238	304	54.5
Overall	-	6590	1457	2402	69.6	4888	952	1544	74.6

Table 4.2: Relation schema, statistics and IAA for the CACER dataset. All relation tails are medical problems.

4.1.2 Building IE Systems

The CACER IE task comprises multiple subtasks. The extraction of *Drug* and *Problem* events requires the identification of triggers and argument spans, prediction of the relationships (argument roles) between triggers and arguments, and resolution of the subtype labels for labeled arguments. Additionally, the relations between *Drug* and *Problem* must be predicted. The subtasks can be performed through multi-step extraction approaches, as well as end-to-

end approaches that jointly extract all phenomena. Our experiments used BERT models and GLMs, incorporating both fine-tuning and ICL. Hyperparameters, such as batch size, gradient accumulation steps, and number of epochs, were optimized on the validation set, and the best-performing models were then evaluated on the test set. Model performance was assessed using a one-sided, bootstrap T-test with 10,000 iterations, where a sample is a note, and the p-value threshold is 0.05.

For BERT-based encoder models, we selected SpERT [76] and PL-Marker [341], both using Bio+ClinicalBERT [12], pre-trained on biomedical and clinical texts, as their encoder. For GLMs, we fine-tuned Flan-T5-large [54] and Llama3-8B-instruct [20] and used GPT-4 in an ICL setting [226]. All GLMs used a consistent prompt format, as shown in Figure 4.2 and Figure 4.3. For fine-tuned GLMs, we applied parameter-efficient fine-tuning (PEFT) with low-rank adaptation (LoRA) [119]. GPT-4 experiments were conducted in a Health Insurance Portability and Accountability Act-compliant Azure environment.

Event Extraction (EE)

Almost all event arguments co-occur in the same sentence as the trigger, so EE models operate on sentences. The GLM event extraction used a common format [251, 87], where the input prompt contains a task description and the target sentence, and the output is a structured text representation of the extracted events ordered by the location of the trigger. If no events are extracted, the output is 'None'. Each event is presented in a uniform template comprising a trigger and all arguments. Individual triggers or arguments start with a type label, followed by a subtype label or spans. Multiple spans of the same argument type are separated by the token <s>, as in '<Problem> pain <Assertion> present <Anatomy> back <s> neck <Duration> ...'. Events are separated by the [SEP] token. To be considered valid, all predicted text spans must occur exactly as they do in the original sentence. GPT-4 experimentation utilized ICL, where the prompt included a concise summary of the annotation guidelines and two randomly selected in-context sentence examples (one containing both *Problem* and *Drug* events and another without events). The complete EE prompts and

condensed annotation guideline can be found in the Appendix B.2 and Appendix B.3.

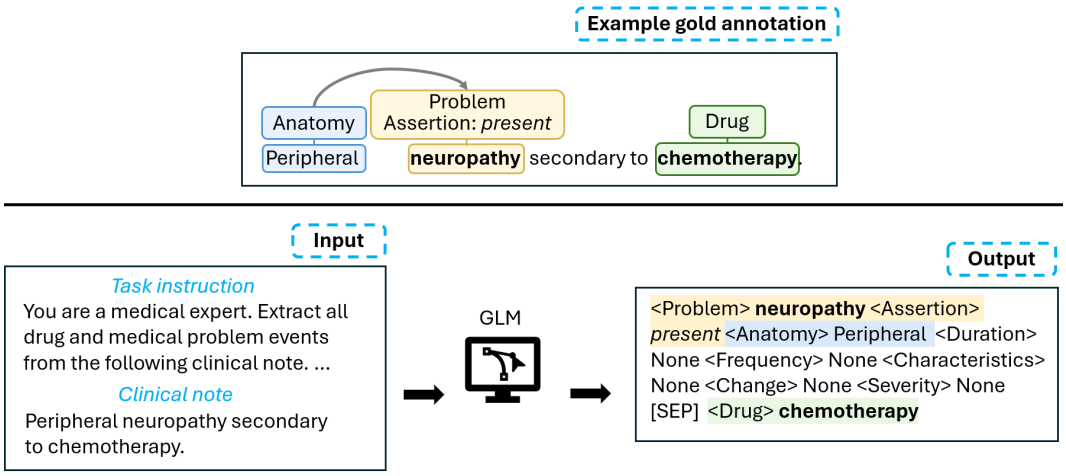

Figure 4.2: Input and output formats for GLMs in EE.

Relation Extraction (RE)

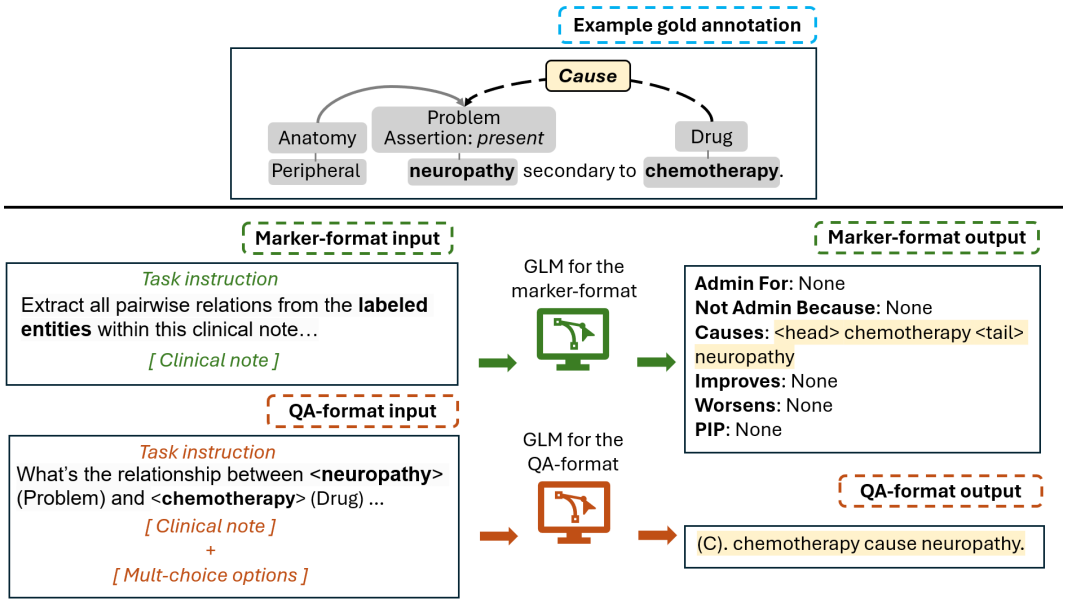

Figure 4.3: Input and output formats for GLMs in RE.

RE is unconstrained by sentence boundaries, and relations may span the entire note. We define the context window for a relation as the smallest continuous sequence of sentences that includes both head and tail events. We consider the RE input to be all context windows that are smaller than 400 BERT tokens and five sentences. This approach covers 98.7% of all relations. Details of the context window implementation are in the Appendix B.4.

Figure 4.3 presents the two RE formats for GLMs: *Marker* and *QA*. The Marker format aligns with PL-Marker, using special markers to denote *Drug* and *Problem* triggers in the input. The model outputs one relation type per line, returning a default *None* label if no relation is identified. If multiple instances of a relation type are found, they are separated by a [SEP] token. To avoid duplication, only relations extracted from their specific context windows are considered valid predictions, even when there is overlap between intra- and inter-sentence relations. The QA format identifies relations through multiple-choice questions, following previous work [270]. For each possible event pair in a context window, a unique input prompt is created that includes task instructions, the context window with labeled event pairs, and the potential relations based on condensed annotation guidelines. There are five possible *Drug-Problem* relations, one asymmetric *Problem-Problem* relation, and the option to indicate no relation by 'None of the above'. The complete RE prompts can be found in the Appendix B.2.

4.1.3 Inter-Annotator Agreement (IAA) and IE Evaluation

We calculated IAA and evaluated the performance of our IE systems using precision, recall, and F1. For events, we employed a relaxed evaluation similar to the N2C2 SDoH challenge [191]. Two triggers are *equivalent* if they have the same event type and overlapping spans. Two labeled arguments are equivalent if (1) their argument types match, (2) their subtype labels match, and (3) they are attached to equivalent triggers[1]. Span-only arguments are considered equivalent if (1) their argument types match, (2) their spans overlap, and (3) they

[1]For labeled arguments, the subtype label normalizes its corresponding span into a categorized value. Therefore, we focus on the subtype label and do not additionally evaluate the argument span.

are attached to equivalent triggers. For relations, equivalence is established if (1) the head and tail triggers are equivalent and (2) the relation types match.

4.2 Results

4.2.1 CACER Annotation

CACER was annotated by three University of Washington (UW) medical students. To familiarize themselves with the schema and refine the guidelines, annotators independently annotated five notes per training session, across a total of six such sessions. Across the annotation rounds, events were annotated first, followed by the annotation of relations between events. IAA was calculated per round to monitor quality, and disagreements were discussed. The training set was singly annotated, 78% of the validation set was doubly annotated, and the test set was fully doubly annotated.

In the first two rounds, annotators demonstrated consistent annotation of triggers, enabling the development of high-performing pre-annotation models. Previous research has shown that using pre-annotation can speed up the gold standard development for clinical NER, while maintaining similar IAA [175]. We trained a pre-annotation model on 29 notes from rounds 1-2 to assist annotators in focusing on more complex tasks and accelerate annotation. The pre-annotation model was a multi-label variation of SpERT (mSpERT) [76, 190], built with Bio+Clinical BERT [12]. We evaluated the pre-annotation model on four singly-annotated notes containing 293 *Problem* and 174 *Drug* triggers, achieving 85.2 and 83.3 F1, respectively. Starting in round 3, notes were pre-annotated with triggers and arguments, but without the trigger-argument linkages. Annotators reviewed and corrected these pre-annotations, and added the trigger-argument linkages in the process.

Table 4.1 presents the IAA for events with and without the use of pre-annotation. Pre-annotation improved the overall IAA and the IAA for *Problem* triggers but did not impact the IAA for *Drug* triggers or some *Problem* arguments. Annotators reported pre-annotation was helpful as it saved time, which is consistent with the findings from previous research

[175]. However, we observed a decline in IAA for the two less frequent *Problem* arguments, *Duration* and *Frequency*. This decline is likely due to the limited training data (29 notes) used to create the pre-annotation classifier and the linguistic variability in the expression of these terms, such as 'at night' for *Frequency*. The pre-annotation classifier has a higher rate of false negatives for these two arguments, resulting in some missed annotations and lower IAA. To address these issues and improve IAA, we reviewed disagreements during feedback discussions after each annotation round. Additionally, all test samples were doubly annotated to ensure annotation consistency.

Table 4.2 presents the IAA for relations, focusing on intra-sentence relations, where sentence boundaries are defined using SpaCy (en_core_web_sm) [2]. Intra-sentence relations account for 70.7% of relations. The micro-average relation IAA is 74.6 F1 for intra-sentence relations and 69.6 F1 for all relations. Inter-sentence relations pose significant challenges due to their implicit expression, making them more difficult to identify and define. To enhance the RE IAA, we conducted a quality check on singly annotated notes using an RE model to identify annotation inconsistencies. We trained the FLAN-T5 model using the GLM-QA format with gold standard events, as described in the following Section, Relation Extraction, on singly annotated notes from one annotator and then used to predict results for notes annotated by another. A third annotator reviewed and corrected any discrepancies between the model and human labels.

[2]https://spacy.io/

4.2.2 Event Extraction (EE) Results

Event	Trigger & Arg.	# Labels	BERT-based LM		GLM		
			SpERT	PL-Marker	Flan-T5	Llama 3 8B	GPT-4 (ICL)
Drug	Trigger	3,534	94.0	93.9	91.4	**95.0**	82.9
Problem	Trigger	6,555	93.1	93.1	90.9	**93.2**	68.2
Problem	Assertion	6,555	89.3	89.6	86.4	**89.8**	62.7
Problem	Change	418	72.0	71.5	**75.2**	73.0	35.0
Problem	Severity	254	**74.7**	73.3	69.7	70.5	35.7
Problem	Anatomy	2,877	81.7	**83.0**	67.4	79.9	57.0
Problem	Characteristics	1,439	71.9	**75.0**	57.5	68.8	17.0
Problem	Duration	298	**72.9**	72.7	67.5	69.6	22.4
Problem	Frequency	79	64.6	62.2	65.3	**65.6**	45.1
Overall	–	22,009	88.3	**88.8**	83.9	88.2	61.7

Table 4.3: Event extraction F1 performance. Except for GPT-4 (ICL), all approaches are finetuned on the CACER train set. **Bold** numbers represent the highest numerical scores. Underlined numbers denote the top-performing systems, indicating statistical significance over non-underlined systems. There is no significant difference between any of the underlined systems.

Table 4.3 presents the EE performance. BERT and Llama3 show no significant difference in overall F1 scores but significantly outperform Flan-T5 and GPT-4 (ICL). The top three models achieve performance close to IAA (without pre-annotation) for triggers, frequent arguments, and overall performance. However, for less frequent arguments (*Severity*, *Duration*, and *Frequency*), the best models underperform compared to IAA. For GLMs, EE performance decreases in sentences with multiple events due to the generation of longer texts, increasing the

likelihood of cascading errors from previously generated tokens. We noted that GPT-4 (ICL) significantly underperforms other models in almost every category of trigger and argument performance, except for the infrequent argument, *Frequency*. This underperformance is primarily due to GPT-4 (ICL)'s occasional non-compliance with in-context annotation guideline instructions. For instance, GPT-4 (ICL) tends to merge attributes such as *Characteristics* and *Anatomy* into *Problem* triggers, resulting in the omission of meaningful arguments.

Error Analysis

Error Types	**Error Subtypes**	**Language Models**				
		SpERT	**PL-Marker**	**Flan-T5**	**Llama 3 8B**	**GPT-4 (ICL)**
False Negatives	Abnormal measures	12	3	2	4	0
	Uppercases/Abbreviations	5	4	2	4	2
	Headers	18	24	16	14	6
	Long spans	6	6	4	0	0
	Dense events	0	0	42	14	0
	Others	24	28	42	32	78
False Positives	Overlapping spans	25	0	0	0	0
	Hallucinations	0	0	0	0	23
	Others	32	25	26	18	101
Mis-classification	Arguments	4	2	2	2	12
	Labels	6	4	4	2	10
	Switched events and arguments	2	6	4	5	12
Overall	-	134	102	144	95	244

Table 4.4: Types of EE errors in 5 sampled notes from the CACER test set. The 5 notes have 987 event triggers and arguments in total. The category, dense events, refers to the scenario where locations of multiple events are in close proximity. The category, *switched events and arguments*, refers to the scenario where an event trigger is classified as an argument, or vice versa.

We conducted an error analysis by randomly sampling 5 notes from the test set and manually characterizing the associated EE errors. Table 4.4 presents a distribution of error types by LM.

There are diverse sources of false negatives (FNs), including: (1) abnormal test results (FN *Problem* events) such as '*PSA started rising up to 9.2*'; (2) words in all caps, to express emphasis or represent abbreviations and acronyms such as '*ATIVAN*', '*VALIUM*', '*SVI*, '*ECE*' (casing issues may be mitigated with case-normalization in preprocessing); (3) *Anatomy* arguments in section headers such as '*Genitourinary*: no rash/erythema in groin area'; (4) long *Problem* triggers, such as '*activity is pretty much confined to walking to the bathroom from the bed*'. (5) locations of multiple events in close proximity (dense events), which are usually from a procedural checklist such as 'he denied any significant symptoms including *headache*, *loss of consciousness*, *vision change*, *chest pain*, ...'. There can be more than 15 such *Problem* events in a single sentence, and Flan-T5 and Llama 3 8B can fail to capture the last few events.

As a frequent source of false positives (FPs), SpERT tends to classify multiple overlapping spans. For example, in the phrase '*red raised rash*', SpERT classifies three *Characteristics* arguments, '*red*', '*raised*', and '*red raised*'. These overlap errors could be reduced through post-processing by merging the overlapping spans of the same type. GPT-4 tends to generate hallucinations for labeled arguments, which are either nonsensical or not accurately reflective of the source content provided [129]. For example, GPT-4 generates the *Severity* argument label, *severe*, for the event '*metastasis cancer*', even though no descriptions of its severity are involved. The other FPs are frequently associated with nuanced discrepancies to the annotation task definition, and we find many common FPs among multiple LMs. For example, the clinical measure, '*PSA nadir*' is misclassified as a *Problem* event; 'happened last night' is a single time point but misclassified as a *Duration* argument. Those types of FPs are especially frequent for GPT-4.

The mis-classifications involve (1) incorrect argument types, for example, 'pain *8/10*' is a *Severity* argument but is predicted as *Characteristics*; (2) incorrect subtype labels for labeled

arguments, for example, the *Assertion* in 'prescribed for future *pain*' should be *hypothetical* but is predicted as *possible*; (3) misclassification of event triggers as arguments or vice versa, for example, 'bony *metastasis* with concerning *lesions*' should have two *Problem* events ('*metastasis*' and '*lesions*') but '*lesions*' labeled as a *Characteristics* argument.

4.2.3 Relation Extraction (RE) Results

CACER includes both intra- and inter-sentence relations. Table 4.5 presents the RE performance: Table 4.5A includes all relations, and Table 4.5B includes only intra-sentence relations. Both tables detail overall performance, assuming gold standard events and end-to-end performance using predicted events. In the RE tasks, Llama with a QA format achieved the highest F1 but was not significantly better than SpERT or PL-Marker. In Table 4.5A, the QA prompting strategy surpassed the Marker approach, suggesting that the more constrained QA format is a better approach for GLMs. Additionally, breaking down complex RE tasks into individual entity-pair tasks achieves higher performance.

In Table 4.5B, GPT-4 (ICL) exhibited significantly poorer performance than all other methods, with many false positives. A major source of false positives for *Problem-Problem* relations is the listing of multiple *Problem*s, such as symptoms, adverse events, and comorbidities, in close proximity within oncology notes, without contextual descriptors or medical knowledge to indicate actual relationships. GPT-4 inference incurs higher computational and financial costs, so we did not evaluate GPT-4 (ICL) further in other experiment settings.

For intra-sentence RE with gold standard events in Table 4.5B, except for GPT-4 (ICL), all approaches achieved higher overall F1 than human IAA. However, when longer-distance, inter-sentence relations are included (Table 4.5A), performance is similar to IAA. This underscores the challenges of long-distance RE, where co-reference and varied relation expressions are involved. Such relations are also often indicated by section headers without explicit descriptions.

Table 4.5A, all relations

Relation	# Labels	BERT-based LM		GLM - marker format			GLM - QA format		
		SpERT	PL-Marker	Flan-T5	Llama 3 8B	GPT-4 (ICL)	Flan-T5	Llama 3 8B	GPT-4 (ICL)
AdminFor	1,422	-	71.9	56.4	57.4	-	75.0	**76.4**	-
NotAdminBecause	54	-	52.9	42.4	35.0	-	**57.4**	**57.4**	-
Causes	321	-	75.7	51.9	57.3	-	78.4	**78.7**	-
Improves	133	-	56.4	43.2	43.8	-	**63.5**	54.3	-
Worsens	121	-	28.6	31.6	28.4	-	**42.6**	34.6	-
PIP	350	-	**58.0**	38.0	40.1	-	52.9	54.3	-
Overall	2,402	-	67.4	50.3	52.1	-	69.2	**70.3**	-
Overall (predicted events)	2,402	61.8	62.0	48.6	51.6	-	62.2	**65.3**	-

Table 4.5B, intra-sentence relations

Relation	# Labels	BERT-based LM		GLM - marker format			GLM - QA format		
		SpERT	PL-Marker	Flan-T5	Llama 3 8B	GPT-4 (ICL)	Flan-T5	Llama 3 8B	GPT-4 (ICL)
AdminFor	765	-	84.6	78.4	79.8	71.8	86.8	**88.8**	83.7
NotAdminBecause	44	-	55.8	57.1	50.7	35.4	61.9	**62.5**	36.4
Causes	274	-	80.8	71.0	75.5	63.1	83.5	**84.4**	74.4
Improves	65	-	**73.8**	65.6	62.1	45.9	72.5	68.4	56.7
Worsens	70	-	40.4	41.1	40.8	32.7	**52.8**	45.6	49.5
PIP	304	-	**61.5**	53.8	58.1	22.4	56.4	58.6	46.3
Overall	1,522	-	76.5	69.3	71.7	56.7	76.7	**78.9**	65.9
Overall (predicted events)	1,522	70.2	70.6	64.8	69.4	-	67.9	**73.1**	-

Table 4.5: Relation extraction performance. Except for GPT-4 (ICL), all approaches are finetuned on the CACER train set. Except for the two overall F1 with predicted events, all RE F1 are based on gold events. **Bold** numbers represent the highest numerical scores. Underlined numbers denote the top-performing systems, indicating statistical significance over non-underlined systems. There was no significant difference among any of the underlined systems.

Error Analysis

Error Types	Error Subtypes	All Relations (Predicted Events)				Intra-sent Relations (Gold Events)
		SpERT	PL-Marker	Flan-T5	llama 3 8B	GPT-4 (ICL)
False Negatives	FN events	6	4	23	11	-
	Relations implicated by symbols	6	2	0	0	0
	Dense events	10	5	4	9	8
	Other Intra-sentence relations	2	14	2	0	1
	Other Inter-sentence relations	15	12	9	5	-
False Positives	FP events	4	2	8	3	-
	Dense events	8	10	9	5	8
	Other Intra-sentence relations	1	1	9	3	7
	Other Inter-sentence relations	0	0	15	11	-
Mis-classification	-	4	7	8	5	7
Overall	-	56	57	87	52	31

Table 4.6: Types of RE errors in 5 sampled notes from the test set. The 5 notes have 128 relations in total, and 87 of the 128 relations are intra-sentence. All GLMs' predictions are generated under the QA prompting format.

We conducted an error analysis for RE using the same 5 notes as in Table 4.6 by manually characterizing the RE errors. We analyzed all RE errors for BERT-based LMs and GLMs with the QA format. For GPT-4 (ICL), we only analyzed errors associated with intra-sentence relations, as GPT-4 experimentation did not include inter-sentence relations. Table 4.6 presents a breakdown of RE error types. Cascading errors from EE, including FN and FP event trigger predictions, result in RE errors. Another major error source is from dense events with ambiguities and possibly hierarchical interactions. Consider the following example, 'Significant *hypotension* while hospitalized, consistent with *sepsis*. Persisted despite *antibiotics*.' PL-Marker captured the *Worsen* relation between '*antibiotics*' and '*sepsis*', but

missed the *Worsen* relation between '*antibiotics*' and '*hypotension*' and the *PIP* relation between '*sepsis*' and '*hypotension*.'

While BERT-based models generated more FNs, GLMs generated more FPs. The FNs for BERT-based models usually come from (1) relations implicated by punctuations, such as the sentence with a colon, '*AKI*: Cr peaked to ..., likely due to supratherapeutic *cyclosporine*', resulting in an FN associated with '*cyclosporine*' *Causes* '*AKI*'; (2) inter-sentence relations where the multiple sentences discuss the same conditions, such as the sentences about the metastasis, 'pelvis revealed extensive bony *metastasis* ... *Lupron* was given', resulting in the FN relation, '*Lupron*' *AdminFor* '*metastasis*'. FPs for GLMs are usually from (1) over-predicting relations when multiple events occur in close proximity, like a list of events; and (2) inter-sentence relations with the same trigger names, for example, '... mild rash and groin *pain* developing; monitor ... prescribed *Oxycodone* prn for *pain*', where the model must disambiguate which reference to 'pain' is associated with the 'Oxycodone' prescription.

The misclassifications stem from *Drug-Problem* relations, where models often confuse *Causes* with *NotAdminBecause*. While the former indicates that the *Problem* event is an adverse effect of the *Drug*, the latter does not. Sentences with ambiguities, such as '*Drug A* was discontinued because of *Problem B*', require medical knowledge to distinguish between these two relations.

4.3 Discussion

4.3.1 Drug Annotation Schema

Previous studies on fine-grained drug annotations in clinical narratives, such as those reported in i2b2 2009 [300], n2c2 2018 [114], and MADE 1.0 [126], have explored the detailed characterization of *Drug* events through attributes including dosage, route, and frequency. However, these attributes are not directly transferable to cancer treatment protocols, where drugs are typically administered as part of a regimen. For example, the R-CHOP regimen — a commonly used treatment for non-Hodgkin lymphoma (NHL) — combines multiple

drugs (Rituximab, Cyclophosphamide, Doxorubicin Hydrochloride, Vincristine Sulfate, and Prednisone) each with specified dosages, routes, and frequencies that are synergistically designed to maximize therapeutic efficacy. Such regimens reflect a complex matrix of attributes that differ significantly from the singular drug attributes documented in traditional drug annotation tasks, thereby rendering these methods less applicable to oncology. Our approach of annotating *Drug* event triggers as specific regimens can potentially be integrated with a hierarchical pharmacological knowledge base [199]. These knowledge bases are designed to characterize component medications and therapeutic contexts, among other elements, providing a structured framework for understanding and organizing complex oncology [199].

4.3.2 Comparison Across Models

In EE and RE, comparing BERT models (SpERT and PL-Marker) and the top-performing GLM (Llama3) reveals no significant differences in overall end-to-end performance. In RE with gold standard triggers, Llama3-QA achieves the highest numerical performance. EE involves identifying specific textual spans and their relationships, while relation prediction focuses on reasoning about pre-identified events, relying on comprehensive language understanding. Although Llama3 performs similarly to BERT models in trigger identification, it excels in understanding relationships between triggers. However, Llama3-QA requires a separate query for each potential relation, unlike BERT models that extract all relationships in a sentence in one step. This higher performance of Llama3 comes at an increased computational cost.

It is important to consider the extraction task requirements to balance performance benefits and computational costs. LLMs excel in tasks demanding medical knowledge and advanced reasoning and have achieved high performance in complex tasks, like medical exams [270], common-sense reasoning [159], and dialogue summarization [343]. However, for tasks that require lower levels of abstraction, like identifying *Drug* triggers, the enhanced capabilities of LLMs may be unnecessary, as smaller models like BERT and T5 can achieve comparable high performance. Future research could investigate the minimal scale of language models required to match human performance, especially for tasks that do not require high

levels of abstraction. This would enable more computationally efficient models to achieve significant performance gains without the exponential increase in computational demands.

4.3.3 Considerations for Clinical Deployment

Our experiments show that fine-grained *Problem* and *Drug* information can be extracted with similar performance to human annotators. Deploying such IE systems can enable large-scale, real-time information usage in EHR-embedded clinical decision support applications and generate real-world evidence for learning health systems. This may lead to more effective, safe, and efficient care and a broader evidence base for decision-making at various levels in healthcare systems.

Although the best-performing RE models achieve performance comparable to IAA, the overall IAA of 69.6 F1 indicates the inherent annotation challenges of this task. To address false positives, it may be possible to incorporate confidence scores when annotating and predicting relations. For example, BERT models and open-source GLMs can directly generate such scores as SoftMax probabilities; GLMs can also implement a second verification step to reduce false positives [95]. False negatives are difficult to detect and may lead to severe consequences if adverse drug events are missed. An ensemble approach using multiple models may help capture more relations, but the effectiveness of this strategy depends on model quality and does not guarantee improved performance [52].

To facilitate large-scale deployment, the extracted *Problem* and *Drug* event triggers could be normalized to standardized medical lexicons like the International Classification of Diseases (ICD-10) [228, 130, 274]. This normalization would enhance the integration of our IE systems into existing systems.

4.4 Conclusion

This study presents the Clinical Concept Annotations for Cancer Events and Relations (CACER), a novel corpus that provides detailed annotations of medical problems, drugs, and their relationships from clinical narratives of oncology notes. Our baseline experiments

with state-of-the-art transformer-based models achieved performance levels comparable to annotator agreement. Error analysis revealed several challenges facing current high-performing models: (1) enhancing the ability to generalize to unseen events for EE and (2) deciphering the complex context and medical knowledge required for RE. Future work will focus on (1) aggregating decisions from both fine-tuned and ICL LMs to balance precision and recall, (2) exploring parameter-efficient fine-tuning (PEFT) of LLMs using techniques such as Low-Rank Adaptation (LoRA) [120], and (3) integrating domain-specific knowledge into RE techniques and improving model capacity to accurately identify relationships across longer textual distances.

Chapter 5

BIOMISTRAL-NLU: GENERALIZED SYSTEM FOR MEDICAL NATURAL LANGUAGE UNDERSTANDING (NLU)

Driven by these motivations outlined in Section 2.2, in this chapter, we will (1) propose a unified NLU framework for NLU tasks in GLMs, (2) instruction-tune BioMistral-7B for medical NLU tasks using the proposed framework, and (3) evaluate our instruction-tuned system and demonstrate its superior overall zero-shot NLU performance on two standardized medical NLU benchmarks, outperforming larger, SOTA foundation models.

5.1 Method

5.1.1 A Unified NLU Framework

Task Formulation

We approach the NLU problem by reframing it as a text generation task. Based on NLU task definitions from the BLURB benchmark [102], we categorize the most common NLU tasks into two main output formats: span extraction and QA format. **Free-text Format** is deployed when some model outputs involve free-text spans extracted from the original input. This format can include label types, text spans, and subtype labels. It is a superset of the event extraction formats used in the two datasets from Chapter 3 and 4. **QA Format** is deployed when all model outputs are constrained to specific label types. This format is widely adopted [194], which has also demonstrated good performance in cancer RE, as shown in Section 4.2 with the CACER dataset.

Using the main output formats, we formulate seven important clinical NLU tasks under a unified NLU format. This format makes it easier to evaluate different NLU tasks and could help transfer knowledge when the system is fine-tuned for a broader set of tasks. Six

of these tasks are based on the BLUE and BLURB benchmarks and include named entity recognition (NER), document classification (DC), relation extraction (RE), multiple-choice question answering (QA), natural language inference (NLI), and semantic text similarity (STS). Additionally, we've added event extraction (EE), which is widely studied in the medical field [85]. In EE, an event consists of a trigger and several arguments that describe the event. The extraction of event triggers (ETE) and arguments (EAE) is similar to NER, while classifying event arguments (EAC) into subtypes is like sequence classification.

Unified NLU Format

These seven medical NLU tasks can be grouped into three main categories: (1) token classification, (2) sequence classification, and (3) sequence regression. Figure 5.1 demonstrates example input-output text pairs for each NLU task.

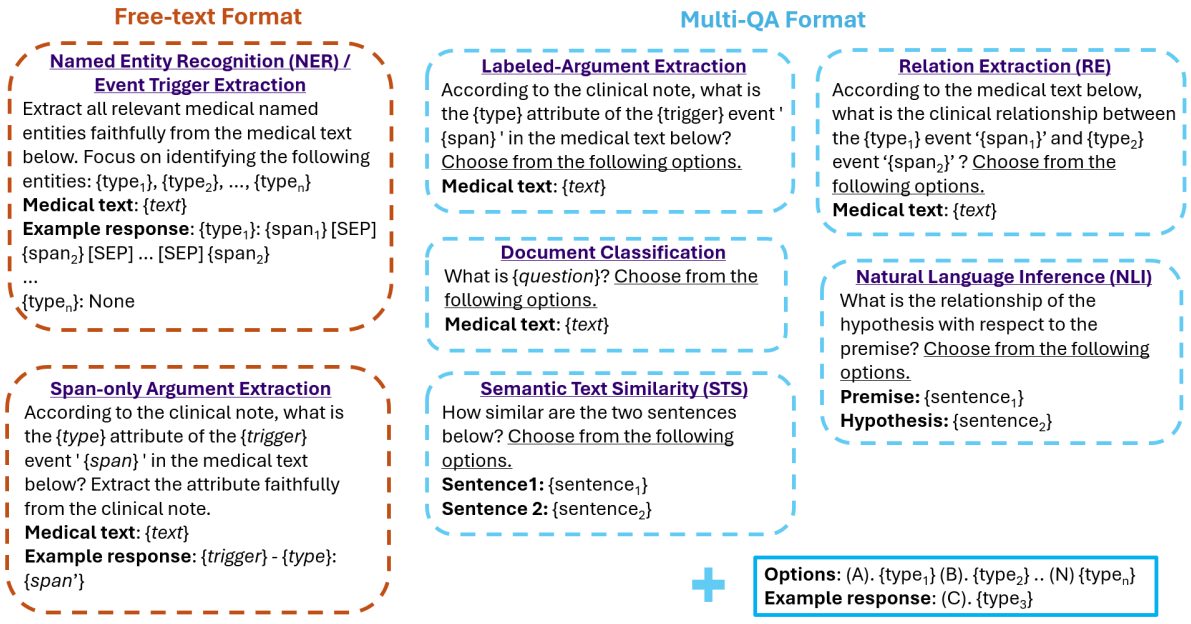

Figure 5.1: Examples of input-output formats for each medical NLU task within our proposed unified framework.

NER, ETE, and EAE are **token classification tasks**, where each token in the input sequence is assigned a specific class label[1]. In token classification, the input consists of a user instruction along with pre-defined token labels, and the target text. The output is organized such that each line contains the class label and the corresponding tokens identified as belonging to that class, listed in the order they appear in the target tokens. Continuous tokens belonging to the same class are grouped into text spans (i.e., entities), separated by ellipses ("..."). For instance, an entity might be labeled as "Disease: fever...headache". If no tokens are identified as entities, the output will be "None".

EAC, DC, RE, QA, and NLI are **sequence classification tasks**, where a class label is assigned to the entire input token sequence (see Table 5.3). In sequence classification, the user instruction provides a set of predefined class labels, typically formatted as multiple-choice options, which is common in instruction tuning [54]. For example, "(B) fevers *happen with* headache". The model's output is one or more of these multiple-choice options. In DC, the input text is the medical document. In RE, it is the text snippet with labeled named entities. For NLI, the input consists of a premise-hypothesis pair. In QA, the user instruction presents a task-related question, and the input is the relevant medical text.

STS is a **sequence regression task**, where a numeric score is assigned to the entire input, representing the semantic text similarity between two input sentences. Our user instructions for STS include multi-choice options that reflect the scoring criteria from the original publication. For example, "(A) The two sentences are on different topics (score 0)."

MNLU-Instruct Dataset

To build the instruction tuning dataset MNLU-Instruct, we focus on seven key medical NLU tasks and utilize clinical and biomedical NLU datasets from publicly available resources. The MNLU-Instruct does not overlap with any datasets from our evaluation benchmarks, BLUE and BLURB. In order to evaluate the generalizability of our proposed system, we deliberately

[1]Tasks like NER are often referred to as sequence labeling tasks in NLP literature [112]. However, in this work, we adopt the term "token classification tasks" for consistency with the BLURB framework [102].

exclude any QA datasets from MNLU-Instruct, reserving medical QA as novel evaluation tasks. Beyond NLU tasks, we additionally include three medical summarization tasks, to enhance the system's text summarization and comprehension abilities. Due to the scarcity of public medical datasets for NLI and STS, we also incorporate general-domain datasets, including SNLI [39], Multi-NLI [325], and SIS-B [309]. The final MNLU-Instruct dataset is derived from 33 publicly accessible datasets, as detailed in Table 5.1.

Task	**Datasets used for instruction-tuning**
NER	i2b2 2006DeID [298], i2b2 2011Coreference [302], i2b2 2012Temporal [283], i2b2 2014 DeID [281], GENIA [346], linnaeus [154], tmVar [321], DrugProt [208], BioRed [185], GNorm [211], NLM-Gene [124], ClinicalIE [7], BC4CHEMD [155], PubMed PICO [135], PICO-Data [216]
EE	i2b2 2009Medication [300], i2b2 2018ADE [114], n2c2 2022SDoH [191]
DC	i2b2 2006Smoking [299], i2b2 2008Obesity [297], TrialStop [247], n2c2 2018 [282], 2024 SemEval Task 2 [140], MTSamples [3]
RE	i2b2 2011Coreference [302], i2b2 2012Temporal [283], BioRed [185], EUADR [303], DrugProt [208]
NLI	BioNLI [28], SNLI [39], Multi-NLI [325]
STS	SIS-B [309]
Summ	PubMedSum [57], CDSR [105], AciDemo [343]

Table 5.1: The MNLU-Instruct dataset, which is used for fine-tuning: NLU and summarization datasets and tasks curated from existing open-source medical corpora.

The input-output pairs for NLU tasks in MNLU-Instruct are constructed using a task-agnostic prompting approach, as illustrated in Figure 5.1. This method directly adapts predefined labels from the original datasets, expanding shortened labels (e.g., from 'GENERIF' to 'Gene reference into a function'). To enhance the diversity of the dataset, we randomize the order of task labels for each input-output pair. This includes shuffling token labels in token classification tasks and multi-choice options in sequence classification and regression tasks. When training splits are missing or when datasets contain only a few examples, we use the entire dataset for training. A complete list of dataset labels, prompts, and statistics

is available in our project GitHub repository[2].

5.1.2 System Development and Experiment Setup

In this section, we will introduce our instruction-tuning setup, evaluation datasets, evaluation metrics, and comparative systems.

System Development: Instruction-tuning

We hypothesize that instruction-tuning on a diverse yet relevant set of tasks can improve the generalizability of LLMs for medical NLU tasks. To test this hypothesis, we fine-tune a high-performing medical LLM on MNLU-Instruct and evaluate its performance in a zero-shot setting. Figure 5.2 demonstrates our instruction-tuning and zero-shot evaluation pipeline.

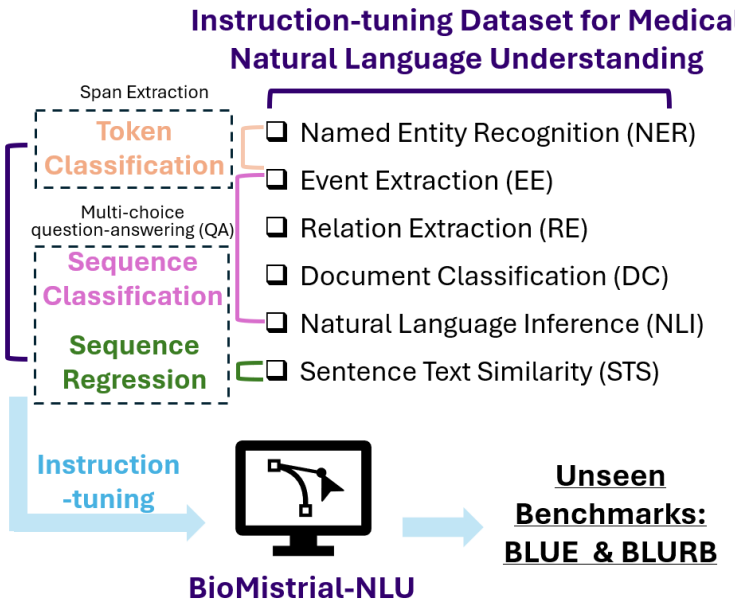

Figure 5.2: Proposed system development, evaluation, and deployment pipeline for our foundation NLU model.

[2]https://github.com/uw-bionlp/BioMistral-NLU

We select BioMistral-7B-DARE as our baseline system, which was the state-of-the-art open-source LLM for multiple medical QA tasks at the time of our experiment[3] [159]. For simplicity, we refer to BioMistral-7B-DARE as BioMistral throughout this work. BioMistral-7B is adapted from Mistral-7B-Instruct-v0.1 [131] by pre-training on PubMed articles. BioMistral-7B-DARE is created by merging the model weights of BioMistral-7B with the original Mistral-7B-Instruct-v0.1 using the Drop And REscale (DARE) technique [348]. DARE enables domain adaptation while preserving the model's instruction-tuned capabilities.

We fine-tune BioMistral using all model parameters on MNLU-Instruct, resulting in BioMistral-NLU-FT. However, fine-tuning LLMs on specialized domains can risk degrading their generalization ability for broader tasks [9]. To mitigate this, and to preserve the versatility of the original BioMistral, we apply the technique of Drop And REscale (DARE) [347] as recommended by Labrak et al. [159]. DARE merges LMs fine-tuned from the same base model by randomly dropping most delta parameters (i.e., differences from the base model) and rescaling the rest, enabling the creation of a combined model without additional training. Using this approach, we construct our final system, **BioMistral-NLU**, by merging the parameters of BioMistral-NLU-FT with those of BioMistral.

The experiment is conducted using the alignment-handbook[4] package. Based on the engineering guidelines from the alignment-handbook GitHub discussion, we set the number of epochs to 3, the batch size to 16, and the learning rate to 2e-04 with a warm-up ratio of 0.1, using 4 A100 GPUs. The other hyperparameters follow the default configuration specified by the alignment-handbook. For inference, we use the vllm package[5] and set the temperature to 0. The entire fine-tuning process for BioMistral-NLU takes less than one day.

---

[3] https://huggingface.co/BioMistral/BioMistral-7B-DARE

[4] https://github.com/huggingface/alignment-handbook

[5] https://github.com/vllm-project/vllm

Evaluation Datasets and Metrics

We evaluate BioMistral-NLU in a zero-shot setting using the BLURB and BLUE benchmarks. Due to the sensitivity of clinical note-based corpora, we exclude two inaccessible datasets from BLUE, namely ShARe/CLEF [284] and MedSTS [319]. Some datasets appear in both benchmarks, leading to a total of 7 tasks and 15 unique datasets being evaluated.

The evaluation datasets are developed using the unified prompt format outlined in Figure 5.1. The entity types and multi-choice options for these datasets are provided in Tables 5.2 and 5.3.

Dataset	**Named entities**
BC2GM	Gene
BC5-chemical	Chemical
BC5-disease	Disease
NCBI-disease	Disease
JNLPBA	Protein, Cell type, RNA, Cell line, DNA
EBM PICO	Interventions, Participants, Outcomes

Table 5.2: NER datasets used in the evaluation.

For consistency with prior studies, we adopt the same evaluation criteria used in BLUE [236] and BLURB [102]. Token classification tasks are assessed using F1 scores at the entity level, except for the PICO dataset, which is evaluated at the token level. For sequence classification tasks, accuracy is used when class labels are balanced (e.g., in NLI and QA tasks). In cases where class labels are imbalanced (e.g., in RE), F1 scores are used. For the sequence regression task (STS), system outputs are converted into numerical integer scores and evaluated using Pearson correlation.

Task	Dataset	Multi-choice options
DC	HoC	10 cancer hallmarks
QA	PubMedQA	yes / maybe / no
	BioASQ	yes / no
RE	GAD	2 gene-disease relations
	DDI	4 drug-drug interactions
	ChemProt	5 chemical-protein relations
	i2b2-2010	8 medical problem relations
NLI	MedNLI	entails / neutral / contradicts
STS	BioSSES	5 similarity score definitions

Table 5.3: Sequence classification and regression datasets used in the evaluation.

Comparative Systems

We compare our proposed system, BioMistral-NLU, with our baseline, the original BioMistral-7B-DARE, as well as other high-performing systems.

Proprietary LLMs: We compare against **ChatGPT** [2] and **GPT-4** [5]. We reference prior research that evaluates these proprietary LLMs on the BLURB benchmark [49, 81][6]. ChatGPT's performance is reported under one-shot in-context learning (ICL), while GPT-4's performance is based on randomly selected 3-shot examples for Named Entity Recognition (NER) tasks and zero-shot for other tasks. Furthermore, the prompts for these models are strategically optimized for each dataset, making these systems highly competitive. Given that Feng et al. (2024) [81] demonstrated GPT-4's superiority over FLAN-T5-XXL [54], PMC-LLaMA-13B [326], and Zephyr-7B-Beta [295], we exclude these systems from further evaluation.

Open-source LLMs: We also compare with the original BioMistral-7B-DARE and

[6]GPT-4 version: gpt-4-0613. ChatGPT version: GPT-3.5, though the exact version was not specified in the original publication.

LLaMA-3.1-8B-Instruct[7] [20]. In our controlled experiments, we evaluate these open-source LLMs using our proposed unified prompting format, as explained in Section 5.1.1.

The evaluation of those LLMs is conducted in a zero-shot setting, except for NER datasets. Since the instruction-tuning phase for these foundation LLMs includes QA tasks, our baseline systems successfully generate outputs containing the original multi-choice options from sequence classification and regression tasks in a zero-shot setting. However, our desired token classification output format is less common during the instruction-tuning phase of these open-source models. To address this, we provide other baseline LLMs with an explanation of the output format along with two in-context examples. These 2-shot examples are randomly selected from the training split of each dataset, and we ensure that the outputs from these examples are distinct to prevent any bias toward a specific answer. More details about the prompts and few-shot sample selection can be found in the example prompts can be found in our project GitHub[8].

Task- and Dataset-Specific Fine-Tuned LM: **BERT-FT**. To better understand the performance gap between generalized foundation LLMs and domain-specific fine-tuned systems, we refer to the reported results of BERT-based models from the BLUE [236] and BLURB [102] benchmarks. For each dataset, the BERT-FT model is fine-tuned specifically on its respective training split.

5.2 Result

Following the practice in BLURB [102], we average system performance across datasets for an overview. As shown in Table 5.4, BioMistral-NLU outperforms the baseline BioMistral with an increase in the macro average score of 19.7 for BLURB and 16.7 for BLUE. Meanwhile, BioMistral-NLU outperforms the proprietary models, achieving an increase in the macro average score of 9.0 over ChatGPT, and 2.7 over GPT-4 for BLURB.

Our results demonstrate that instruction-tuning on diverse medical NLU tasks using our

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[8] https://github.com/uw-bionlp/BioMistral-NLU

unified format effectively improves the LLMs' generalizability to unseen NLU datasets. In this section, we will analyze the results and characterize the gaps between the systems.

5.2.1 Comparison Across Systems

When comparing BioMistral-NLU with the baseline BioMistral, we observe an average performance improvement of 33.7 in F1 scores for NER tasks and 8.2 for other tasks. This difference can likely be attributed to the instruction-tuning phase of BioMistral. While the NER task may have been less frequent during BioMistral's instruction-tuning, tasks such as QA, which use a similar prompting strategy, may overlap with some of BioMistral's instruction-tuning tasks. Therefore, instruction-tuning on a broader spectrum of NLU tasks could enhance the generalizability of LLMs across domains.

For comparison with proprietary LLMs in the BLURB benchmark, BioMistral-NLU outperforms GPT-4 by an average F1 score of 9.7 on NER tasks. However, for other BLURB tasks, BioMistral-NLU's performance is 2.0 higher than ChatGPT and 5.4 lower than GPT-4. This performance discrepancy is expected due to GPT-4's significantly larger parameter size and its exposure to a more diverse instruction-tuning corpus, which enhances its ability to generalize for complex tasks involving advanced reasoning, in line with empirical findings on model scaling [145, 54].

When comparing BioMistral-NLU to dataset-specific fine-tuned systems such as BERT-FT, we find a performance gap of 20.3 points in BLURB and 26.3 points in BLUE. This difference likely stems from the inherent ambiguity in medical NLU tasks, where even human annotators can disagree on interpretations, leading to discrepancies in task outcomes [333, 222]. BERT-FT systems leverage extensive domain-specific training data, which reduces such ambiguity. In contrast, BioMistral-NLU, which relies on generalized task definitions from input prompts, faces challenges in addressing these ambiguities effectively, making it difficult for it to match the performance of fine-tuned systems.

Task	Evaluation Metric	Dataset	# test instances	In-domain	Generalized LLMs with zero- or few-shot ICL				
				BERT-FT [236] [102]	Chat-GPT [49]	GPT-4 [81]	Llama-3-8B	BioMistral	
								Baseline	Ours
NER	Entity-level F1	BC2GM†	6,322	84.5	37.5	54.6	12.6	34.1	**61.5**
		BC5-chemical†*	5,385	93.3	60.3	78.2	52.5	45.0	**89.9**
		BC5-disease†*	4,424	85.6	51.8	63.9	38.7	33.7	**67.0**
		NCBI-disease†	955	89.1	50.5	66.0	33.5	39.9	**61.8**
		JNLPBA†	8,657	79.1	41.3	45.4	33.3	25.6	**64.4**
	Token-level F1	EBM PICO†	24,474	73.4	55.6	33.5	20.2	19.6	**55.3**
DC	F1	HoC†*	315	81.5	51.2	62.5	23.1	47.3	**63.8**
QA	Acc	PubMedQA†	500	60.2	76.5	70.6	71.0	**72.0**	70.2
		BioASQ†	263	94.8	88.6	85.7	78.7	74.9	**86.7**
RE	F1	GAD†	534	84.0	52.4	51.5	55.6	55.0	**58.5**
		DDI†*	5,761	82.4	51.6	37.7	**13.2**	10.0	13.0
		ChemProt†*	14,744	77.2	34.2	37.6	35.2	28.6	**38.1**
		i2b2-2010*	6,292	76.4	-	-	38.9	30.9	**41.8**
NLI	Acc	MedNLI*	1,422	73.5	-	-	49.1	49.3	**57.5**
STS	Pearson Corr	BioSSES†*	20	92.3	42.8	89.3	67.9	69.1	**80.8**
Overall	Macro average	BLURB†	-	82.9	53.4	59.7	41.2	42.7	**62.4**
		BLUE*	-	82.8	-	-	39.8	39.2	**56.5**

Table 5.4: Our proposed system, BioMistral-NLU's zero-shot performance on 15 unseen medical NLU datasets from 2 benchmarks: BLURB (labeled by †) and BLUE (labeled by *). **Bold** indicates superior performance over the BioMistra-7B and Llama-3-8B, which utilize the same, dataset-agnostic prompts as BioMistral-NLU. Underline indicates better performance over the ChatGPT and GPT-4 ICL, which utilize dataset-specific prompts.

5.2.2 Error Analysis

For the NER tasks, a notable challenge for BioMistral-NLU is accurately identifying the boundaries of named entities. For example, in the BC2GM gene NER dataset, the model predicted "Id - 1" as a named entity, while the gold-standard annotation was "mouse Id - 1".

To assess the impact of such discrepancies, we used a relaxed evaluation criterion, where named entities are considered equivalent if their spans overlap. Under this relaxed criterion, BioMistral-NLU shows an average F1 improvement of 15.5 across the five NER datasets.

Despite its instruction-tuning benefits, BioMistral-NLU experienced a drop in performance on the PubMedQA dataset compared to its baseline, BioMistral. This decline highlights the potential drawbacks of fine-tuning on specialized datasets, which can lead to the model's degradation in broader tasks. While the MNLU-Instruct dataset includes document summarization tasks, it does not directly transfer to the question-answering tasks, leading to a performance dip.

In RE tasks, BioMistral-NLU shows recall rates significantly higher than its precision (by 10 to 70 points), suggesting a tendency to identify a higher number of false positives. This issue arises primarily when the model predicts relations between entities that do not fit any of the predefined relation categories. Instead of recognizing "no relation," the model erroneously assigns a wrong relation label.

In the sequence regression tasks, such as BioSSES, BioMistral-NLU tends to predict intermediate similarity scores (e.g., 2 or 3) more frequently than extreme values (0, 1, 4, or 5). This pattern indicates a tendency to under-represent extreme values, which could be a consequence of the model's generalization on less specific input prompts.

5.3 Discussion

We have demonstrated that instruction-tuning on diverse medical NLU tasks can enhance LLMs' downstream generalization to unseen medical NLU datasets in a zero-shot setting. In this section, we will evaluate the impact of instruction dataset composition, focusing on two components: instruction-tuning tasks and domains.

5.3.1 Impact of Instruction-tuning Tasks

We aim to assess the impact of instruction-tuning task selection from two perspectives: (1) its relevance to downstream tasks and (2) its task diversity. Focusing on these two perspectives,

we fine-tune the baseline system, BioMistral, with different subsets of tasks used to build BioMistral-NLU. We evaluate the fine-tuned system on the 4 RE datasets from Table 5.4 in a zero-shot setting, and compare the macro-average F1 scores across the 4 RE datasets.

To study the impact of task relevance, we first construct two instruction-tuning setups: (1) with the RE task (**w/ RE**) and (2) with the DC task (**w/o RE**). We chose the DC task because DC employs a similar QA prompting format to RE and it contains 6 diverse datasets from Table 5.1. To study the impact of task diversity, besides DC and RE, we additionally include 2 and 4 more randomly selected tasks from Table 5.1. More specifically, our experiment settings are:

1. **w/ RE**:
   (a) 1 task: RE
   (b) 3 tasks: RE, NLI, NER
   (c) 5 tasks: RE, NLI, NER, EE, STS
2. **w/o RE**:
   (a) 1 task: DC
   (b) 3 tasks: DC, NLI, NER
   (c) 5 tasks: DC, NLI, NER, EE, STS

All fine-tuning experiments are controlled by using a fixed number of 50,000 data instances and running for three epochs. We maintain an equal number of instances for each task (i.e., 50,000/k instances per task when fine-tuning with k tasks), and randomly sample fine-tuning instances from all datasets within the same task.

After BioMistral is fine-tuned with the same number of instances, we observe the following from Figure 5.3: (1) Overall, setting 1 (with RE) consistently outperforms setting 2 (without RE), due to its relevance to the RE datasets used in downstream evaluation; (2) In both settings, system performance increases with the number of fine-tuning tasks, demonstrating the benefits of fine-tuning with multiple tasks; (3) When fine-tuning on a single task, whether fine-tuning improves system performance on downstream tasks depends on the similarity between fine-tuning task and the downstream task.

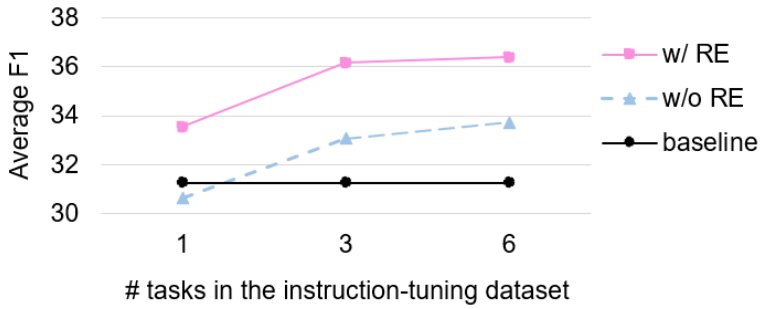

Figure 5.3: Average zero-shot performance on the 4 RE datasets, after instruction-tuning on 50k instances.

5.3.2 Impact of Instruction-tuning Domain

After demonstrating the benefits of diverse instruction-tuning tasks, we now examine individual tasks. Note that the BLUE benchmark includes both biomedical and clinical datasets: biomedical data comes from scientific publications, while clinical data consists of semi-structured clinical notes from patients [329]. In this section, we assess how domain selection affects downstream generalizability.

We follow a similar experimental setup as described in Section 5.3.1, fine-tuning BioMistral for three epochs over 25,000 data instances. The fine-tuned system is evaluated on six biomedical NER datasets from Table 5.2 in a zero-shot setting, using macro average F1 scores. The instruction-tuning NER datasets from MNLU-Instruct [9] are divided into biomedical and clinical splits. Our experiments include fine-tuning on a single split (**BioMed** / **Clinical**) and both splits (**Both**). We additionally combine single splits or include additional instances, creating a similar experiment setting with 50k instances. We use the 2-shot BioMistral described in Section 5.1.2 as the baseline system.

From Figure 5.4, we observe the following: (1) Instruction-tuning on the BioMed domain alone consistently outperforms tuning on the Clinical domain alone when using the same

[9]We also include event triggers as named entities.

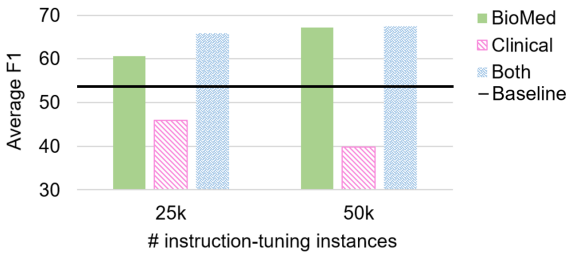

Figure 5.4: Average zero-shot performance on 6 biomedical NER datasets, when finetuned on different domains.

number of instances. (2) Compared to the baseline, instruction-tuning on the Clinical domain negatively impacts downstream performance on the BioMed domain. (3) Combining instances from both domains improves downstream generalizability to the BioMed domain, even with the same total number of instances. (4) Increasing the number of instances from the BioMed or Both domains improves performance, whereas more instances from the Clinical domain alone decrease performance.

5.4 Conclusion

In this chapter, we introduce a unified prompting format for 7 important medical NLU tasks, and develop an instruction-tuning dataset based on publicly available clinical and biomedical corpora. Our experiment demonstrates that fine-tuning across diverse medical NLU datasets improves the system's generalizability in a zero-shot setting with dataset-agnostic prompt tuning. Our ablation study underscores the necessity for instruction tuning across diverse medical NLU tasks, including domain-specific lexicon and common biomedical tasks.

Our future work will focus on further improving the generalized LLM's zero-shot performance on medical NLU tasks and narrowing its gap to in-domain fine-tuned systems. Because LLMs often struggle to adhere to in-context annotation guidelines [358], our future work will focus on integrating nuanced task descriptions from annotation guidelines into both the fine-tuning and inference stages [258]. Future work could also involve a self-verification step

[95] or using a knowledge base as augmentation [166] to reduce false positives in the sequence classification tasks.

Chapter 6

DETECTING DATA CONTAMINATION IN MEDICAL BENCHMARKS

Motivated by the gaps highlighted in Section 2.4, we make the following contributions related to data contamination detection:

1. Conduct a systematic review of 50 papers related to data contamination detection in LMs.
2. Present formal, mathematical definitions for different types and levels of data contamination.
3. Categorize and analyze the key assumptions and requirements underlying existing detection approaches, evaluating whether these assumptions are validated in current literature.
4. Use real-world case studies to show that several unverified assumptions fail under practical scenarios.

6.1 Literature Evaluation

6.1.1 Paper Collection

To systematically investigate approaches for data contamination detection, we implement a three-step literature review process with 3 sources of papers: (1) four key survey papers [336, 123, 121, 63] and papers from the 1st Workshop on Data Contamination at ACL 2024[1]; (2) relevant papers cited by the papers from Source (1); and (3) relevant papers cited by the papers from Source (2).

From the above three sources, we collect 50 relevant papers on data contamination.

[1] https://conda-workshop.GitHub.io/.

We additionally filter these papers according to the inclusion criterion: the paper must propose and/or evaluate detection approaches for data contamination in text datasets and LMs. Furthermore, eight studies solely discussing risks and mitigation strategies for data contamination do not meet the inclusion criteria and are excluded.

Consequently, our review includes a total of 50 papers. Among them, we systematically summarize those detection approaches and present formal mathematical representations for their underlying requirements and assumptions. We then evaluate whether these underlying assumptions are true under different scenarios, as described below.

Data contamination can occur at the instance or dataset level, and the detection approaches for them can be different. To facilitate the discussion, we would like to first provide formal mathematical definitions of data contamination at these levels, based on the descriptive definitions from previous research [336, 25, 123].

6.1.2 Instance-Level Contamination

In this study, we focus on text datasets and define a language instance x as a sequence of word tokens. Originally, **direct** instance-level contamination is defined as the presence of an instance x within an LM M's training set, D_M, i.e. $x \in D_M$ [336]. However, LMs often do not publish their exact training corpus, but instead refer to multiple datasets, as subsets of D_M [362]. These datasets typically undergo various pre-processing steps, such as de-duplication, filtering, masking, and removing noise. Consequently, LMs are trained on slightly different versions of the same dataset [230]. Meanwhile, there is also **indirect** instance-level contamination from greater variations of the dataset, such as machine paraphrasing [339].

To account for such minor differences and indirect contamination, in our Definition D1 below, we introduce a **Binary Indicator Function for Instance-Level Contamination**, $b(x, x')$, which returns *True* (1) if two instances are considered to be the same and *False* (0) otherwise. Researchers can determine what instances are considered to be the same by defining $b(x, x')$ accordingly.

Detection Approach	Require-ments (ID)	Assump-tions (ID)	Detection Rearch	Critiques on Those Approaches
Instance Similarity (9 papers)	Disclose (R1) & Release (R2)	None	[71], [78], [171], [250], [64], [339], [238], [165], [200]	
Prob. Analysis (16 papers)	None	Absolute Prob. (A1)	[273], [266], [205], [198], [276], [323], [170]	[198], [62], [74], [42], [206]
	Perturbed Instance (R3)	Ref. Prob. (A2)	[201], [198], [227]	[74], [198], [206]
	Ref. LM (R4)	Ref. Prob. (A3)	[45], [198], [209], [353] [204]	[198, 62, 74, 42], [206]
	Other	Other	[127], [357]	
Instance Gen. & Instance Select. (20 papers)	None	Verbatim Mem. (A4)	[46], [142], [195], [75]*, [292], [263], [100]*	
	Key Info. (R5)	Key Info. Gen. (A5)	[64], [244], [48], [231], [45], [44], [176], [100]	
	None	Gen. Variation (A6)	[73]	
	Metadata (R6)	Metadata Mem. (A7)	[257]* [146]*	[62]
Answer Mem. (5 papers)	Instance Perturb. (R7)	Answer Change (A8)	[176]*, [207]*, [345]*, [369]*, [246]*	

Table 6.1: Existing detection approaches for direct data contamination, their requirements and assumptions, and critiques they received. Some papers cover multiple detection approaches with different assumptions. Most detection methods apply to both instance- and dataset-level contamination, while * denotes those limited to dataset-level contamination. In this study, we show that the underlined assumptions may not be often satisfied.

Definition D 1. Instance-Level contamination: Let D_M be the training data of an LM M. The binary function $f(M, x)$ is defined as follows:

$$f(M,x) = \begin{cases} 1 & \text{if } \exists x' \in D_M, b(x,x') = 1 \\ 0 & \text{if } \forall x' \in D_M, b(x,x') = 0 \end{cases} \tag{6.1}$$

We define an instance x to be **seen** by M, or M is **contaminated** by x, iff $f(M, x) = 1$. Conversely, we define an instance x as **clean** or **unseen** by M iff $f(M, x) = 0$.

The detection of instance-level contamination is commonly referred to as **membership inference attack (MIA)**. The goal of MIA is to determine the probability of an instance being used to train an LM, namely, $\widehat{f}(M, x)$ [121].

6.1.3 Dataset-Level Contamination

Prior research implicitly refers to dataset-level contamination at two degrees: partial dataset contamination and full dataset contamination.

Definition D 2. Full Dataset Contamination: A dataset D is fully contaminated (**fully seen**) by an LM, if every instance within this dataset is contaminated. Namely, $\forall x \in D, f(M, x) = 1$.

When creating benchmarks for detecting data contamination, previous work typically generates the fully contaminated split. For example, [198] created contaminated and clean datasets, respectively from the training and validation splits of the LM's pretraining corpus. Shi et al. [266] focused on LMs which disclosed that they used Wikipedia event data for training, and created the contaminated dataset from the Wikipedia event data which were published before the LMs' release.

Definition D 3. Partial Dataset Contamination: A dataset D is partially contaminated (**paritially seen**) by an LM M, if at least one instance within D is seen. Namely, $\exists x \in D, f(M, x) = 1$.

In practice, especially when reporting contamination from benchmark datasets [73] or detecting copyrighted content [146, 48], people focus more on evaluating partial dataset contamination.

Definition D 4. Unseen/Clean Dataset: A dataset is clean (**unseen**) by an LM, if none of its instances is contaminated. Namely, $\forall x \in D, f(M, x) = 0$.

6.2 Detection of Direct Data Contamination

Direct data contamination is the most common and well-researched type of data contamination. In this section, we categorize the existing detection approaches, their requirements, assumptions, and the critiques they received (see Table 6.1). The requirements are defined as the preliminary conditions necessary for conducting certain detection approaches. The assumptions are what the authors of detection approaches assume to be true; the assumptions either are explicitly stated by the authors or can be inferred from the detection approaches.

Most detection methodologies for direct contamination are primarily developed to address instance-level contamination; however, they can be adapted to account for dataset-level contamination. Consequently, unless specified otherwise, this section will concentrate on instance-level contamination.

The performance of a detection method depends on how well its requirements are met and the reliability of its assumptions. Therefore, we group the detection approaches based on their assumptions and requirements.

6.2.1 Instance Similarity

When D_M is known, detection approaches based on **instance similarity** directly deploy Equation 6.1, by proposing a similarity function to measure $b(\cdot, \cdot)$ and comparing a new instance with every $x \in D_M$.

Previous research focuses on developing a better or more efficient similarity function. Examples of similarity calculation can be conducted through exact match [71], fuzzy match

[238, 162], automatic NLG evaluation metrics [78, 64], and another LM [339]. Tools have also been developed to allow efficient search within a large D_M, such as Data Portraits [200] and ROOTS Search Tool [238].

Although this approach does not rely on underlying assumptions, it has two requirements:

Requirement R 1. D_M needs to be disclosed.

Requirement R 2. D_M must be accessible, which is often hindered by legal, privacy constraints, and expired website links.

Case Study: To examine how often these two requirements are met, we analyzed the top 10 LMs on the Vellum LLM leaderboard[2]. We found that *none* of the LMs fulfilled R1, the most basic requirement, let alone R2 (see Appendix D.3 for details).

6.2.2 Probability Analysis

When the training dataset D_M is unavailable, but the LM M's output token probabilities are known, probability analysis has been used to detect potential instance-level contamination. We group those detection approaches by their assumptions, and unless specified otherwise, they have no requirements.

Absolute Probability

Given an instance x, probability analysis measures instance-level contamination through $P_M(x)$, the probability of the instance x based on an LM M.

Assumption A 1. Seen instances will have higher probabilities than unseen ones, and there exists a threshold, ξ_p, that separates seen instances from unseen ones:

$$P_M(x) \begin{cases} \geq \xi_p & \text{if } f(M,x) = 1 \\ < \xi_p & \text{if } f(M,x) = 0 \end{cases} \tag{6.2}$$

[2] `https://www.vellum.ai/llm-leaderboard#model-comparison`. Accessed on Oct 6, 2024.

Previous research measures $P_M(x)$ through perplexity [45, 170] or approximates it through LM loss [323], which can be impacted by the instance domain and simplicity. To improve upon this assumption, Shi et al. [266] evaluates only the average token probability of top $p\%$ least likely tokens in an instance (**Min $p\%$ Token**), assuming that unseen instances contain more low-probability outliers in Wikipedia events and books. Similarly, Song and Shmatikov [273] assesses probabilities of the k most frequent tokens.

Likewise, Srivastava et al. [276], Wei et al. [323], and Meeus et al. [205] proposed inserting special strings as watermarks into the training data, using the probability of these watermarks to detect data contamination.

However, Maini et al. [198] and Duan et al. [74] have demonstrated that the perplexity and min top p probabilities are close to random in detecting direct instance-level data contamination across different splits of the Pile dataset. Maini et al. [198] suggests that shifts in perplexity and infrequent word probabilities may be attributed to temporal events on platforms like Wikipedia, rather than contamination. Similarly, Cao et al. [42] highlighted that perplexity and token probability approaches are ineffective for code generation tasks.

Reference Probability by an Instance

Instead of assuming the probabilities of all the seen instances are higher than the probabilities of all the unseen instances, this approach compares the probabilities of similar instances.

Requirement R 3. There exists an algorithm which, given an instance x and an LM M, can automatically generate a similar unseen instance, x'.

Assumption A 2. If x and x' are similar and M has seen x but not x' , the probability of x should be much higher than that of x' based on M:

$$P_M(x) \begin{cases} \gg P_M(x') & \text{if } f(M,x) = 1 \\ \not\gg P_M(x') & \text{if } f(M,x) = 0 \end{cases} \tag{6.3}$$

Utilizing this assumption, Mattern et al. [201] construct a similar, reference instance x' by replacing individual words in x with their synonyms. However, in practice, the observation

that $P_M(x) \geq P_M(x')$ might result from replacement with rare words. This assumption has been proven false by Maini et al. [198] on different splits of the Pile dataset [92]. In addition to this synonym-based perturbation, Maini et al. [198] demonstrate the ineffectiveness of other perturbation approaches, including white space, characters, random deletion, and case changes.

Another study, Oren et al. [227], constructs the reference instance by randomly shuffling (exchanging) the order of sentences in the original instance. They make another assumption of the *exchangeability*, positing that all orderings of all the instances in an exchangeable benchmark should be equally likely if uncontaminated. This assumption might not be valid for coding and reasoning tasks.

Reference Probability by Another LM

This type of approach compares the probability of an instance based on two LMs.

Requirement R 4. Given an instance x, we can find another LM M' such that x is unseen by M'.

Assumption A 3. If x is seen by M but not M', then $P_M(x)$ should be much higher than $P_{M'}(x)$:

$$P_M(x) \begin{cases} \gg P_{M'}(x) & \text{if } f(M,x) = 1 \\ \not\gg P_{M'}(x) & \text{if } f(M,x) = 0 \end{cases} \tag{6.4}$$

Previous research has utilized term frequency [204], the zlib entropy [45], and another LM [45, 209] as the reference model. However, Maini et al. [198] and Duan et al. [74] have demonstrated that those reference models perform close to random guessing across various domains. Cao et al. [42] also show the zlib entropy does not work for code generation tasks. Instead of using reference probabilities at the sentence level, Zanella-Béguelin et al. [353] deploy a reference model for both individual token probability and its probability rank within the vocabulary.

6.2.3 Instance Generation and Instance Selection

In this section, we investigate underlying requirements and assumptions for detection approaches based on instance generation and instance selection.

Instance generation detects contamination by treating x as a prefix-suffix pair, $x = (x_p, x_s)$. These approaches evaluate the LM M's generated output, $M(x_p)$, conditioned on x_p. If $M(x_p)$ is similar or identical to x_s, x will be predicted as seen. Based on this core intuition, instance generation approaches vary in their assumptions regarding input-output pairs and language generation approaches. Unless specified otherwise, those approaches below focus on instance generation.

For instance selection, instead of directly generating answers, the LM is tasked with selecting the most likely x_s from a set of candidate options in a multi-choice format. However, detection approaches relying on instance selection face a fundamental limitation: even if x is unseen, the LM might still choose the correct x_s by accident. Consequently, these approaches are generally not employed to detect instance-level contamination but rather to assess the probability of full dataset contamination.

Verbatim Memorization

This type of approach assumes LMs can memorize their training data, to certain extent.

Assumption A 4. Given a prefix-suffix pair $x = (x_p, x_s)$, if x has been seen by an LM M, x_s can be generated (memorized) by M through greedy decoding, when given the input x_p.

$$M_g(x_p) \begin{cases} = x_s & \text{if} f(M, x) = 1 \\ \neq x_s & \text{if} f(M, x) = 0 \end{cases} \tag{6.5}$$

Duarte et al. [75] and Golchin and Surdeanu [100] define x_s as sentences or passages, and create similar instances to x_s by paraphrasing x_s using another LM. They use instance selection, and assume that the contaminated LM will be more likely to select the verbatim option.

However, instance-level contamination does not always lead to verbatim memorization. Utilizing instance generation, Kandpal et al. [142], Carlini et al. [44], Carlini et al. [46], and Tirumala et al. [292] demonstrate that verbatim memorization requires repeated exposures to this instance x during training, and a larger LM and longer input length x_p can result in better memorization. Schwarzschild et al. [263] used the minimum length of x_p needed to generate the desired output x_s to define the degree of memorization.

Similarly, Kandpal et al. [142] and Carlini et al. [45] study a relaxed version of this assumption, where the LM can generate x_s through different sampling strategies in decoding, such as top-k or top-p (Nucleus) sampling [117]. They reach a similar conclusion that data contamination does not necessarily lead to memorization.

Key Information Generation

This type of approach assumes that, if an LM has seen an instance, it can generate x's key information based on its context.

Requirement R 5. An instance x can be paraphrased into a slot-filling, context-key pair, $x = (x_c, x_k)$. The key x_k is usually a representative sub-span of x, such as dates and names. The rest tokens in x compose the context, x_c.

Assumption A 5. If x is seen, M will be able to produce similar output to x_k when given x_c.

$$S(M(x_c), x_k) \begin{cases} \geq \tau_s & \text{if } f(M, x) = 1 \\ < \tau_s & \text{if } f(M, x) = 0 \end{cases} \tag{6.6}$$

Here, $M(x_c)$ denotes the output of the LM M through a certain decoding method. $S(\cdot, \cdot)$ is a text similarity function, and τ_s is the contamination threshold. One can use the similarity functions described in Section 6.2.1.

Leveraging this assumption, prior studies have masked key information within specific datasets, including input questions in NLP benchmarks [64, 100, 176], column names in SQL code generation questions [244], character names in books [48], and labels in NLI and SST tasks [195].

Generation Variation

This type of approach explores how an LM's outputs vary if it has seen an instance during training.

Assumption A6. Suppose an instance x can be represented as a prefix-suffix pair, $x = (x_p, x_s)$. If an LM M has seen x, then given x_p, M will generate something identical or similar to x_s under different sampling strategies:

$$\text{Var}(\{M_\cdot(x_p)\}) \begin{cases} < \xi_v & \text{if } f(M, x_p) = 1 \\ \geq \xi_v & \text{if } f(M, x_p) = 0 \end{cases} \tag{6.7}$$

where $Var(\{M_\cdot(x_p)\})$ measures the variations of outputs from M under diverse, different sampling strategies when given x_p; ξ_v is a threshold, based on the type of input x_p and sampling strategies.

Dong et al. [73] defines the metric $Var(\cdot)$ as 'Contamination Detection via output Distribution' (CDD), and utilizes this assumption to detect memorization in coding and reasoning benchmarks. However, this assumption can lead to false positives for other tasks, such as multiple choices, where the output is more constrained and has less variation.

Metadata-based Memorization

This type of approach determines whether an LM has seen a dataset D by using D's metadata.

Requirement R6. Given a dataset D, we can construct an input prompt x_m including D's metadata m, such as dataset name, split, and format.

Assumption A7. If an LM M has seen a dataset D, when given D's metadata, M is able to generate instances that are very similar to some $x \in D$.

$$\begin{cases} \exists x \in D, S(M(x_m), x) \geq \tau_m & \text{if } D \text{ is seen} \\ \forall x \in D, S(M(x_m), x) < \tau_m & \text{if } D \text{ is unseen} \end{cases} \tag{6.8}$$

Here, $M(x_m)$ is the set of instances that M generates when given D's metadata m; $S(M(x_m), x)$ represents the highest similarity between x and any instance x' as a subsequence of $M(x_m)$; τ_m is the contamination threshold for $S(\cdot, \cdot)$.

Sainz et al. [257] and Golchin and Surdeanu [100] utilized this assumption to demonstrate that OpenAI systems memorized many instances from widely used benchmarks. However, this approach can have false negatives if the LM's training phase does not preserve the linkage between D's metadata and instances [62].

6.2.4 Answer Memorization

Answer memorization is usually conducted at the dataset level. It introduces perturbations to the original dataset, measures the LM's performance change, and aims to detect whether the LM's high performance is due to memorizing its answer.

Requirement R7. Given an LM M and an evaluation dataset D, one can generate a similar dataset D' that is unseen by M, by modifying every $x \in D$.

Assumption A8. Suppose datasets D and D' are similar and an LM M has seen D but not D', M's performance on D ($\text{Eval}(M, D)$) will be much higher than its performance on D' ($\text{Eval}(M, D')$).

$$\text{Eval}(M, D) \begin{cases} \gg \text{Eval}(M, D') & \text{if } D \text{ is seen} \\ \not\gg \text{Eval}(M, D') & \text{if } D \text{ is unseen} \end{cases} \tag{6.9}$$

Previous research evaluates answer memorization in multiple-choice (MC) tasks by introducing variations such as altering numbers in mathematical tasks [207], changing the order of MC options, etc. [344, 369]. Razeghi et al. [246] show that multiple LMs perform better on numerical reasoning tasks involving frequently occurring numerical variables in their pretraining data. Similar to this assumption, Liu et al. [176] detects if the LM can still predict the correct answer after removing all MC options.

6.3 Other Types of Contamination

Besides direct data contamination, previous research also investigates indirect data contamination (6 papers) and task contamination (5 papers).

6.3.1 Indirect Data Contamination

Indirect data contamination occurs when an instance x is not seen by an LM M, but something (x') derived from x is. For instance, x' can be a paraphrase or a summary of x [339].

Indirect data contamination is often hard to track and trace [25]. For example, OpenAI uses online user conversations for training, which could include variations of benchmark datasets[3]. Another example involves knowledge distillation, where an LM utilizes instances x_k generated by another LM M' during training, and these instances x_k may resemble instances from the training set $D_{M'}$ of M' [307].

Detection Approaches for Indirect Data Contamination

Compared to direct contamination, indirect data contamination is much more challenging to detect. Dekoninck et al. [62] and Cao et al. [42] show that many probability-based detection approaches are ineffective for indirect data contamination.

However, prior research showed that three approaches may still be applicable: (1) the instance similarity measured by another LM [339], (2) the CDD metric [73], which leverages Assumption A6 by measuring output variations rather than directly comparing with original instances, and (3) directly tracking the disclosed usage of datasets. For example, Balloccu et al. [25] reviewed the datasets evaluated using OpenAI APIs.

---

[3] `https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance`. Accessed on Oct 6, 2024.

6.3.2 Task Contamination

Task contamination occurs when any instance of the same task is seen by an LM [168]. Detecting task contamination is crucial for assessing an LM's generalizability to unseen tasks [54]. Tasks can include applications such as machine translation, summarization, and mathematical calculation. Task contamination is a broader concept than data contamination: if some labeled instances from a dataset are seen by an LM, the associated task is contaminated, but task contamination doesn't always imply the dataset has been seen.

Task contamination generally evaluates an LM's performance on a particular task at the dataset level. The idea is that if an LM has previously seen the task, its performance will be much higher compared to unseen tasks of similar difficulty.

For example, as noted by Aiyappa et al. [10], LLMs show improved performance on the same benchmark after model updates, which may be influenced by data contamination during LLMs' continuous training. Ranaldi et al. [244] and Li and Flanigan [168] also find that OpenAI models perform significantly better on benchmarks released before the model's release than on those released later, when task difficulty is controlled or performance is normalized using a baseline model. To ensure fair comparisons across tasks, Li and Flanigan [168] control task difficulty using a baseline system. However, Cao et al. [42] also note that LMs do not necessarily perform worse on more recent code generation benchmarks.

6.4 Case Study

Besides the case study in Section 6.2.1, we aim to evaluate whether the assumptions outlined in Table 6.1 are universally applicable across different domains, for direct and instance-level MIA.

6.4.1 Assumptions to Evaluate

As shown in Table 6.1, prior research has verified that 4 out of 8 assumptions can fail under certain conditions. Meanwhile, some assumptions have specific requirements, and their

applicability depends on how well these requirements are met. Therefore, we focus on two unverified assumptions that have no such requirements for evaluation, limiting confounding factors and deferring the testing of other assumptions to future studies. We also validate one verified assumption (Assumption A1) to confirm the consistency of our findings with prior research.

Assumption A1: Absolute Probability. In the assumption that seen instances will have a lower perplexity and fewer low-likely (outlier) tokens, we measure perplexity by an instance's first k tokens (**PPL_**k) [45]; **Min** p**% Token** by the average token probabilities among p% least likely tokens [266].

Assumption A4: Verbatim Memorization. We expand this assumption from the instance level to the token level, assuming an LM will memorize some tokens in seen instances. We measure the percentage of tokens in an instance ranked as the k most likely in casual language modeling (**Mem** k). The k value of 1 represents greedy decoding, and larger than 1 simulates the decoding with top k token sampling.

Assumption A6: Generation Variation. We evaluate the assumption that, given a seen prefix sequence, the LM exhibits less variation (i.e., higher certainty) in predicting the next token under different token sampling strategies. Since lower entropy indicates greater certainty, we measure entropy over the top k most likely tokens (**Entropy** k) (see Appendix D.4 for details).

6.4.2 Experiment Design

To enhance the generalizability of our results, we evaluate these assumptions using four different types of LLMs, with datasets used in different training phases, shown in Table 6.2. We also investigate the impact of model size on MIA performance, using seven Pythia models with parameter sizes ranging from 70M to 13B.

Except for the UltraChat dataset, each dataset consists of multiple smaller subsets from different domains. For our experiments, we randomly sample 9 subsets from each, and consider each subset as an individual dataset. To minimize distribution shifts between seen

and unseen datasets, we randomly select 1,000 instances from the training split (seen) and 1,000 instances from the test split (unseen) within each dataset. If a test split is unavailable, we sample from the validation split. If there are fewer than 1,000 unseen instances, we use the entire test split, ensuring that each split contains at least 100 instances.

LM	# **Params**	**Training Phase**	# **Epochs**	**Trainset Size**	**Batch Size**	**Seen & Unseen Datasets Used in Our Case Study**
Pythia [34]	70M - 12B	Pretraining	≈ 1.5	825 GiB	2M	Pile [92]
OLMo-2 [220]	7B		≈ 2	22.4 TB	$\approx$ 4M	AlgebraicStack [22]
Zephyr-7B-β [295]	7B	Supervised	1-3	9.3 GB	512	UltraChat [69]
BioMistral-NLU [90]	7B	Fine-tuning	2	3.6 GB	64	Medical-NLU [90]

Table 6.2: LMs and datasets used in the case study. Except for the UltraChat dataset, each dataset contains multiple subsets from different domains. The trainset refers to the whole trainset used in each LM's corresponding training phase, as described in their original paper, which is a superset of seen & unseen datasets used in our case study.

Following prior work [266], we evaluate MIA performance using the area under the ROC curve (AUC) at the instance level, representing the probability that a seen instance has a better score (higher or lower) than an unseen instance [101].

6.4.3 Results

Within-Domain MIA

Table 6.3 shows the AUC for each MIA method across representative subsets (domains) of each dataset. Complete results for all datasets and domains are available in Appendix D.5.1.

On pretraining datasets, all metrics perform close to random guessing, with AUC close to 50. We also observed the same pattern with different sizes of Pythia LMs. Our results for Assumption A1 are consistent with critiques they received (see Section 6.2.2). We suspect that during pretraining, LMs are more likely to learn underlying data distributions, instead

of memorizing specific instances.

On fine-tuning datasets, we observed a great variation in MIA AUC across domains. The best-performing metric, PPL_200 on the RE-2012temp dataset, can have an AUC as high as 99.4. This suggests that data contamination from memorizing training instances remains a risk during the fine-tuning phase. Overall, the performance of the perplexity-based metric improves as the number of tokens increases. This trend is linked to the fine-tuning process, where tokens at the beginning of the training instance serve as input prompts but are not explicitly learned during training.

Training Phase		**Pretaining**				**Supervised Fine-tuning**		
Model		**Pythia-6.9B**		**OLMO-2-7B**		**Zephyr-7B-β**	**BioMistral-NLU-7B**	
Assumptions & Metric		**Youtube-Subtitles**	**ArXiv**	**Github-Coq**	**Github-Isabelle**	**Ultra Chat**	**DC-MTSample**	**RE-2012temp**
A1	PPL_50	50.7	50.7	47.1	50.9	55.3	51.9	**62.5**
	PPL_100	50.4	51.1	48.2	53.1	59.1	60.0	**95.3**
	PPL_200	49.6	50.9	48.5	51.6	**60.1**	58.9	**99.4**
	Min 5% token	48.5	51.3	49.4	54.0	**63.6**	47.4	**92.9**
	Min 15% token	48.6	51.3	49.3	53.6	**63.0**	51.7	**93.3**
	Min 25% token	48.5	51.4	49.3	53.1	**61.5**	53.3	**93.4**
A4	Mem 5	49.1	52.7	52.1	48.7	52.2	47.9	41.4
	Mem 15	48.6	51.9	50.6	58.6	53.1	49.1	45.2
	Mem 25	48.2	51.6	49.4	59.2	53.1	50.3	48.9
A6	Entropy 5	49.3	52.0	47.2	52.3	54.3	55.4	**93.2**
	Entropy 15	49.0	52.0	47.8	52.1	54.1	54.8	**93.1**
	Entropy 25	48.9	51.9	48.6	52.5	54.0	54.1	**93.1**
Average AUC		49.0	51.8	49.0	53.3	55.9	52.0	**74.1**
PPL	Seen	13.2±16.0	7.9±3.7	10.5±8.1	8.4±5.2	5.3±4.6	1.7±0.2	1.4±0.1
	Unseen	12.7±10.6	8.0±3.6	9.9±7.2	9.2±6.5	6.2±4.1	1.8±0.2	3.0±1.6

Table 6.3: Average MIA AUC for different LMs. For LMs evaluated on multiple subsets (domains) of the same dataset, we present the results from the subsets with the lowest and highest average AUC. The last two rows, marked as 'PPL_200', represent the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Cross-Domains MIA with Data Distribution Shifts

Within the same domain, the similar average PPL between seen and unseen instances indicates that they have similar underlying distributions, but also a high variation (STD). However, PPL differs a lot across domains. We therefore examine the impact of distribution shifts from different domains on the MIA performance, with the scenario where seen and unseen instances are from different domains.

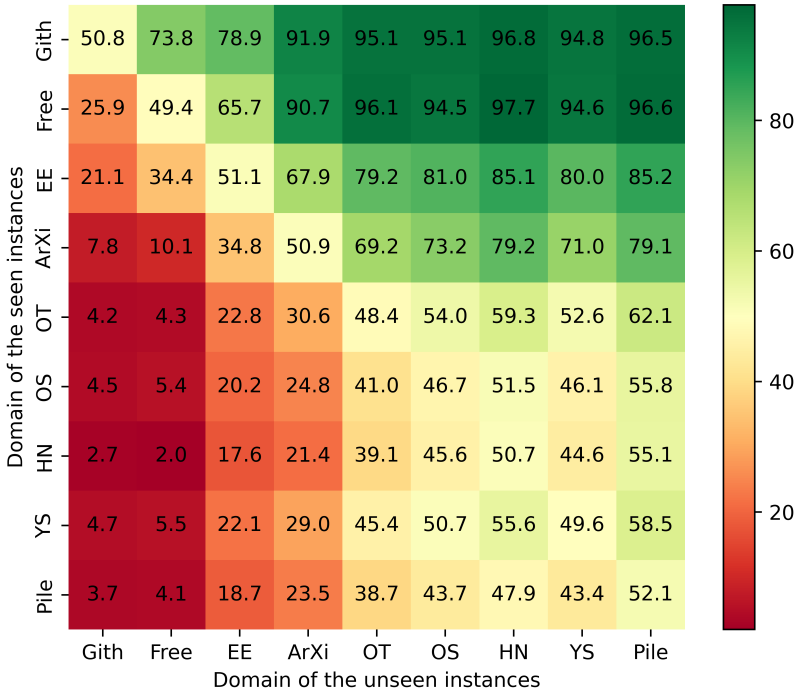

Figure 6.1: Average MIA AUC for the Pythia-6.9b model with PPL_200, when the seen and unseen instances are from different domains. The abbreviations represent the domains in Table D.9 in the Appendix D.5.1.

We present the AUC with PPL_200 in Figure 6.1. The AUC is high in the top-right corner when seen instances are from a domain with lower average PPLs and unseen instances are from a domain with higher average PPLs. Conversely, the bottom-left corner has low AUCs. This indicates that the accuracy of PPL_200-based MIA highly depends on the

domain difference, instead of the seen vs. unseen distinction. More information about the PPL_200 distribution within and across domains is in Appendix D.5.2. A similar trend is observed with other metrics (see Appendix D.5.2 & D.5.3).

6.5 Discussion

In our case study, we observed that the evaluated MIA approaches perform well only on certain domains of datasets used during the fine-tuning phase, but not during the pretraining phase. This discrepancy may be attributed to the significantly larger dataset and batch sizes employed in the pretraining phase.

Together, our case study and prior research show that 6 out of the 8 assumptions listed in Table 6.1 can often be invalid under certain conditions. The other two unverified assumptions, key information memorization (A5) and answer change due to memorization (A8), depend on very specific requirements, complicating their evaluation. Overall, our findings suggest that MIA remains a challenging task.

The limited effectiveness of MIA in pretraining phases suggests detecting data contamination in benchmarks remains an important challenge for LLM evaluation. While poor MIA performance may indicate a lack of direct instance memorization, models could still learn underlying distributional patterns of benchmark data, enabling artificially high performance on the benchmark datasets [62]. On the other hand, privacy and copyright risks persist, as LLMs might learn from sensitive/proprietary data (e.g., patented concepts) without triggering MIA alarms.

6.6 Conclusion

In this study, we present a comprehensive survey of 50 studies focused on data contamination detection and their underlying assumptions. Our theoretical analysis reveals that these assumptions may not apply consistently across different contexts.

Our case studies showed that 3 out of the 8 assumptions are not universally applicable across all training phases and dataset domains, especially for datasets used in the pretraining

stage. Our cross-domain MIA experiments additionally show that many assumptions measure an LM's goodness of fit, which is not necessarily the result of instance memorization due to data contamination. Thus, detecting data contamination remains challenging.

Links to all relevant papers and the code for the case study are available on our project GitHub[4].

6.7 Ethical Considerations

In this work, we employed multiple LMs and their training sets, which may contain sensitive and/or identifiable information. For example, the Pile [92] includes content crawled from the Internet. The BioMistral-NLU-7B and its training set contain datasets derived from de-identified clinical notes [90]. Therefore, we only downloaded the necessary instances and published the numerical results from our experiments. In our project GitHub, we only release our code for data sampling and MIA approaches to ensure reproducibility; we do not publish any actual data instances or LMs. We recommend the community to check the corresponding regulations before deploying the datasets and LMs for other purposes.

[4] https://github.com/velvinnn/LLM_MIA

Chapter 7
CONCLUSION AND FUTURE WORK

This chapter concludes the thesis by summarizing its key contributions, discussing limitations, and proposing directions for future work. Finally, I end this chapter with reflections and closing remarks.

7.1 Key Contributions

This thesis makes three primary contributions toward the reliable evaluation and enhancement of LLMs for clinical applications:

1. **Creation of high-quality clinical IE resources.** We developed two human-annotated benchmark datasets on two important clinical domains: *Pediatric SDoH* and *Cancer Symptoms & Treatments*. Each dataset is constructed with medical annotators, double annotation on the test set, and publicly available guidelines. These resources fill critical gaps in open clinical corpora and provide rigorous test beds for downstream model evaluation.

2. **Systematic benchmarking of state-of-the-art models.** By comparing (1) fine-tuned BERT variants, (2) instruction-following LLMs through ICL, and (3) fine-tuned instruction-following LLMs, we showed that task-specific fine-tuning still outperforms few-shot ICL by a non-trivial margin. Our best fine-tuned models achieved F_1 scores that approach inter-annotator agreement, demonstrating practical viability for clinical IE.

3. **Extending evaluation beyond IE to NLU.** After observing the limited performance of generalized LLMs on clinical IE tasks, we decided to improve their performance on

the broader medical NLU tasks, which encompass IE as a subset. We instruction-tuned the open-source LLM, *BioMistral*, on a curated suite of biomedical NLU tasks. Despite its moderate parameter count, BioMistral delivered SOTA *zero-shot* accuracy on the widely used BLURB and BLUE benchmarks, surpassing much larger closed-source LLMs.

4. **A principled study of data contamination in LLM evaluation.** We study another critical issue in evaluating LLM, data contamination, featuring the overlap between the training and evaluation sets. Data contamination can potentially inflates LLM performance on benchmark evaluations and leads to performance degradation on similar but unseen datasets. We survey existing detection methods for data contamination and mathematically categorize their underlying assumptions. Additionally, we conduct a case study examining three types of assumptions with four different open-source LLMs. Our results show that all tested assumptions fail in detecting the use of pretraining corpus and data domain distribution shifts. This suggests that, in such scenarios, LLMs may have been learning general data distributions rather than memorizing specific sentences.

7.2 Limitations

- **Dataset Availability:** Due to the sensitive nature of clinical narratives, we plan to release the two IE datasets following Institutional Review Board (IRB) approval and careful de-identification of the source clinical notes. Access will be granted to researchers upon signing an appropriate data use agreement.

- **Evaluation of BioMistral-NLU:** Currently, the evaluation of BioMistral-NLU is limited to six natural NLU tasks included in the BLURB and BLUE benchmarks. Further assessment across a broader range of NLU tasks and diverse clinical domains is needed for its generalizability and robustness.

- **Limitations from the Data Contamination Study:** The literature review included 50 studies, but relevant papers may have been missed due to search limitations. The focus was primarily on English-language models; detection methods for non-English LMs or non-text modalities were not addressed. Approaches from other machine learning domains may be transferable, but were not explored. LLM memorization varies based on multiple factors (e.g., dataset domain, size, batch size). Observations from the case study may not generalize to all LMs or training settings.

7.3 Future Work

More Comprehensive Evaluation of BioMistral-NLU To establish real-world reliability, BioMistral-NLU must be stress-tested on datasets that differ along multiple axes—not only in clinical specialty (e.g., radiology, emergency medicine, mental health) but also in note type (discharge summaries vs. progress notes), writing style, patient demographics, and institutional conventions. Systematically sampling these factors will reveal whether performance drops stem from vocabulary drift, discourse structure, or genuinely novel clinical concepts. Such analyses can in turn guide targeted instruction-tuning or lightweight adapter training for domains where zero-shot generalization remains weak.

Multi-Agent Clinical NLP Pipelines Emerging research frames LLMs as cooperative agents that can delegate subtasks to specialized models or retrieval services. Exploring how *BioMistral-NLU* can act as a specialized NLU agent, and coordinate with other agents, is also a promising direction.

Better Detection and Mitigation of Data Contamination Our case study showed that multiple assumptions under existing contamination detection algorithms can fail, with large pre-training corpora and data domain shifts. Future work should either develop more advanced algorithms for data contamination detection, or craft more standardized governance protocols on the use of existing benchmarks, as well as proprietary or sensitive data, for

model training.

Pathways to Clinical Adoption Bridging the gap between research and clinical deployment requires comprehensive, real-world studies that measure not only model performance but also clinician workload, decision latency, and patient outcomes. Key milestones include: integration with EHR systems through standards like FHIR [21]; grounded text generation that links LLM-generated responses with source evidence for more informed decision making; and rigorous bias audits across demographic and institutional strata.

7.4 Final Remarks

The advent of ChatGPT has ushered in a transformative era in NLP. During the latter half of my PhD journey, I witnessed the rapid evolution of the clinical NLP landscape, marked by innovations such as ICL [6], RAG [166], CoT [322] and the emergence of autonomous agents [315]. This period has been both exhilarating and challenging, offering unprecedented opportunities for exploration and growth.

Shunyu Yao's essay, *The Second Half*, articulates a pivotal shift in the AI research paradigm. He posits that while the initial phase of AI focused on developing new training methods and models, the forthcoming phase will emphasize rigorous evaluation, particularly in high-stakes domains like healthcare [340]. In this section, I want to bring up two important directions in LLM evaluation: generalization and clinical alignment. For researchers in clinical NLP, it is especially important to work closely with clinical practitioners to evaluate clinical NLP systems - not only in research settings, but also in how they influence physicians' workflows and decision-making in real-world clinical practice.

Recent studies present a nuanced picture of large language models (LLMs) like ChatGPT in clinical settings. For instance, one study by Goh et al. [98] found that GPT-4 assistance significantly improved physicians' clinician decision accuracy without increasing demographic bias, boosting accuracy from 47% to 65% for White male patients and 63% to 80% for Black female patients; however, another trial by Goh et al. [99] showed no significant improvement

in physician diagnostic reasoning when LLMs were used alongside traditional tools, though LLMs alone outperformed both physician groups, highlighting their diagnostic potential but also underscoring the need for better integration into clinical workflows. These contrasting outcomes may stem from differences in study design and context: the first study employed standardized video vignettes with controlled variables, allowing for focused assessment of AI assistance on specific clinical decisions, while the second study involved a broader range of diagnostic scenarios, reflecting greater complexity and variability in physician performance. Additionally, the manner, in which physicians receive GPT-4-generated recommendations after initial assessments versus having access to an LLM alongside conventional resources, could have influenced the effectiveness of AI assistance in enhancing clinical decision-making.

Given these motivations, more comprehensive evaluation of clinical NLP systems in realistic settings is essential. This includes both simulated and real-world evaluations. In simulated settings, the focus can lie in developing realistic benchmarks, designing more human-aligned automatic evaluation metrics, and assessing the consistency of LLM outputs with established clinical literature and guidelines. Real-world evaluation, on the other hand, should focus on human-computer interaction (HCI) research, particularly how to design and integrate these systems in ways that effectively support clinical practitioners within their actual workflows.

BIBLIOGRAPHY

[1] Natural-Language Understanding –an overview. r://www.sciencedirect.com/topics/computer-science/natural-language-understanding. Accessed: 2025-05-01.

[2] OpenAi: Introducing ChatGpt. r://openai.com/blog/chatgpt, 2022. Accessed: 2024-04-12.

[3] Welcome to Mtsamples. r://mtsamples.com/, 2023. Accessed: 2024-6-8.

[4] Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. Medec: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint arXiv:2412.19260*, 2024.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 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, December 2022. ACL. doi: 10.18653/v1/2022.emnlp-main.130.

[7] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large Language Models are Few-Shot Clinical Information Extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[8] Sangzin Ahn. The impending impacts of large language models on medical education. *Korean Journal of Medical Education*, 35(1):103, 2023. doi: 10.3946/kjme.2023.292.

[9] Samuel K Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo permutation symmetries. *arXiv preprint arXiv:2209.04836*, 2022.

[10] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-yeol Ahn. Can we trust the evaluation on ChatGpt? In Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta, editors, *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 47–54, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.trustnlp-1.5. URL https://aclanthology.org/2023.trustnlp-1.5.

[11] Matthew S Alkaitis, Monica N Agrawal, Gregory J Riely, Pedram Razavi, and David Sontag. Automated Nlp extraction of clinical rationale for treatment discontinuation in breast cancer. *JCO Clinical Cancer Informatics*, 5:550–560, 2021.

[12] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BerT embeddings. In Anna Rumshisky, Kirk Roberts, Steven Bethard, and Tristan Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL https://aclanthology.org/W19-1909.

[13] Raid Alzubi, Hadeel Alzoubi, Stamos Katsigiannis, Daune West, and Naeem Ramzan. Automated Detection of Substance-Use Status and Related Information from Clinical Text. *Sensors*, 22(24):9609, 2022. doi: 10.3390/s22249609.

[14] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[15] Anthropic. Introducing Claude 3.5 Sonnet, jun 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Accessed on Oct 6, 2024.

[16] Anthropic. Claude 3 Haiku: our fastest model yet, mar 2024. URL https://www.anthropic.com/news/claude-3-haiku. Accessed on Oct 6, 2024.

[17] Anthropic. Introducing the next generation of Claude, mar 2024. URL https://www.anthropic.com/news/claude-3-family. Accessed on Oct 6, 2024.

[18] Alan R Aronson. Effective mapping of biomedical text to the UmlS metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[19] Ron Artstein. Inter-annotator Agreement. In *Handbook of Linguistic Annotation*, pages 297–313. Springer, 2017.

[20] AI at Meta. Introducing Meta Llama 3: The most capable openly available Llm to date. r://ai.meta.com/blog/meta-llama-3/, 2024. Accessed: 2024-04-18.

[21] Muhammad Ayaz, Muhammad F Pasha, Mohammed Y Alzahrani, Rahmat Budiarto, and Deris Stiawan. The Fast Health Interoperability Resources (FhiR) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR medical informatics*, 9(7):e21929, 2021.

[22] Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An Open Language Model For Mathematics, 2023.

[23] Tamara E. Baer, Emily A. Scherer, Eric W. Fleegler, and Areej Hassan. Food Insecurity and the Burden of Health-Related Social Problems in an Urban Youth Population. *J Adolesc Health*, 57(6):601–607, 2015. ISSN 1054-139X. doi: https://doi.org/10.1016/j.jadohealth.2015.08.013. URL https://www.sciencedirect.com/science/article/pii/S1054139X15003365.

[24] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, 2022. URL https://arxiv.org/abs/2204.05862.

[25] Simone Balloccu, Patrıcia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source Llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, 2024.

[26] Dibyanayan Bandyopadhyay, Soham Bhattacharjee, and Asif Ekbal. Thinking machines: A survey of llm based reasoning strategies. *arXiv preprint arXiv:2503.10814*, 2025.

[27] Satanjeev Banerjee and Alon Lavie. MEteOr: An Automatic Metric for Mt Evaluation with Improved Correlation with Human Judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

[28] Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. BioNli: Generating a Biomedical Nli Dataset Using Lexico-semantic Constraints for Adversarial Examples. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104, 2022.

[29] Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, 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*, 25(1):61–71, 2018. doi: 10.1093/jamia/ocx059.

[30] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 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, July 2012. ACL. URL https://aclanthology.org/D12-1091.

[31] Markus Berndt and Martin R Fischer. The role of electronic health records in clinical reasoning. *Annals of the New York Academy of Sciences*, 1434(1):109–114, 2018.

[32] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 1052–1062, 2016.

[33] Prabin Bhandari and Hannah Brennan. Trustworthiness of Children Stories Generated by Large Language Models. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 352–361, Prague, Czechia, September 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.inlg-main.24. URL https://aclanthology.org/2023.inlg-main.24/.

[34] Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O' Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430, 2023.

[35] Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical

professionals. *Journal of Biomedical Informatics*, 137, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104274. URL https://www.sciencedirect.com/science/article/pii/S1532046422002799.

[36] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[37] Anusha Bompelli, Yanshan Wang, Ruyuan Wan, et al. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review. *textitHealth Data Sci*, 2021, 2021. doi: 10.34133/2021/9759016.

[38] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling Laws for Data Poisoning in Llms, 2024. URL https://arxiv.org/abs/2408.02946.

[39] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[40] Claudia Breischneider, Sonja Zillner, Matthias Hammon, Paul Gass, and Daniel Sonntag. Automatic extraction of breast cancer information from clinical reports. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 213–218. IEEE, 2017.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. 33:1877–1901, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[42] Jialun Cao, Wuqi Zhang, and Shing-Chi Cheung. Concerned with Data Contamination? Assessing Countermeasures in Code Language Model. *arXiv preprint arXiv:2403.16898*, 2024.

[43] Jiarun Cao, Elke M van Veen, Niels Peek, Andrew G Renehan, and Sophia Ananiadou. A novel Automated Approach to Mutation-Cancer Relation Extraction by Incorporating Heterogeneous Knowledge. *IEEE Journal of Biomedical and Health Informatics*, 27(2): 1096–1105, 2022.

[44] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL `https://www.usenix.org/conference/usenixsecurity19/presentation/carlini`.

[45] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[46] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. In *The Eleventh International Conference on Learning Representations*, 2022.

[47] David S Carrell, David Cronkite, Roy E Palmer, Kathleen Saunders, David E Gross, Elizabeth T Masters, Timothy R Hylan, and Michael Von Korff. Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inform.*, 84 (12):1057–1064, 2015. doi: 10.1016/j.ijmedinf.2015.09.002.

[48] Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. Speak, Memory: An Archaeology of Books Known to ChatGpt/Gpt-4. In *Proceedings of the 2023*

Conference on Empirical Methods in Natural Language Processing, pages 7312–7327, 2023.

[49] Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. An extensive benchmark study on biomedical text generation and mining with ChatGpt. *Bioinformatics*, 39(9):btad557, 2023.

[50] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEdiTroN-70B: Scaling Medical Pretraining for Large Language Models, 2023. URL `https://arxiv.org/abs/2311.16079`.

[51] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

[52] Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46, 2020.

[53] Zhixuan Chu, Mengxuan Hu, Qing Cui, Longfei Li, and Sheng Li. Task-driven causal feature distillation: Towards trustworthy risk prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11642–11650, 2024.

[54] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53, 2024.

[55] Cheryl Clark, Kathleen Good, Lesley Jezierny, Melissa Macpherson, Brian Wilson, and Urszula Chajewska. Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, 15(1):36–39, 2008. doi: 10.1197/jamia.M2442.

[56] Anni Coden, Guergana Savova, Igor Sominsky, Michael Tanenblatt, James Masanz, Karin Schuler, James Cooper, Wei Guan, and Piet C De Groen. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. *Journal of biomedical informatics*, 42(5):937–949, 2009.

[57] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097.

[58] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04)*, pages 423–429, 2004.

[59] Aron Culotta, Andrew McCallum, and Jonathan Betz. Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text. In Robert C. Moore, Jeff Bilmes, Jennifer Chu-Carroll, and Mark Sanderson, editors, *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages

296–303, New York City, USA, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/N06-1038.

[60] Surabhi Datta, Elmer V Bernstam, and Kirk Roberts. A frame semantic overview of Nlp-based information extraction for cancer-related Ehr notes. *Journal of biomedical informatics*, 100:103301, 2019.

[61] Neal A DeJong, Charles T Wood, Madlyn C Morreale, Cameron Ellis, Darragh Davis, Jorge Fernandez, and Michael J Steiner. Identifying social determinants of health and legal needs for children with special health care needs. *Clin Pediatr (Phila)*, 55(3): 272–277, 2016. doi: 10.1177/0009922815591959.

[62] Jasper Dekoninck, Mark Niklas Müller, Maximilian Baader, Marc Fischer, and Martin Vechev. Evading Data Contamination Detection for Language Models is (too) Easy. 2024.

[63] Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. Unveiling the Spectrum of Data Contamination in Language Model: A survey from Detection to Remediation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 16078–16092, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.951.

[64] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating Data Contamination in Modern Benchmarks for Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8698–8711, 2024.

[65] Teresa L Deshields, Patricia Potter, Sarah Olsen, and Jingxia Liu. The persistence of

symptom burden: symptom experience and quality of life of cancer patients across one year. *Supportive Care in Cancer*, 22:1089–1096, 2014.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BErt: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[67] 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. Health equity in pediatrics: Current concepts for the care of children in the 21st century (Dis Mon). *Disease-a-Month*, page 101631, 2023. doi: 10.1016/j.disamonth.2023.101631.

[68] Lisa DiMartino, Thomas Miano, Kathryn Wessell, Buck Bohac, and Laura C Hanson. Identification of uncontrolled symptoms in cancer patients using natural language processing. *Journal of pain and symptom management*, 63(4):610–617, 2022.

[69] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations, 2023.

[70] Nicholas J Dobbins, Bin Han, Weipeng Zhou, Kristine F Lan, H Nina Kim, Robert Harrington, Özlem Uzuner, and Meliha Yetisgen. LeafAi: query generator for clinical cohort discovery rivaling a human programmer. *Journal of the American Medical Informatics Association*, 30(12):1954–1964, 2023.

[71] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk

Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

[72] Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38, 2022.

[73] Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, and Ge Li. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. *arXiv preprint arXiv:2402.15938*, 2024.

[74] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do Membership Inference Attacks Work on Large Language Models? In *Conference on Language Modeling (COLM)*, 2024.

[75] André Vicente Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. DE-cOp: Detecting Copyrighted Content in Language Models Training Data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LO4xhXmFal.

[76] Markus Eberts and Adrian Ulges. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In *24th ECAI*, 2020. URL https://ebooks.iospress.nl/volumearticle/55116.

[77] Vera Ehrenstein, Hadi Kharrazi, Harold Lehmann, and Casey Overby Taylor. Obtaining data from electronic health records. In *Tools and technologies for registry interoperability, registries for evaluating patient outcomes: A user's guide, 3rd edition, Addendum 2 [Internet]*. Agency for Healthcare Research and Quality (US), 2019.

[78] Aparna Elangovan, Jiayuan He, and Karin Verspoor. Memorization vs. Generalization: Quantifying Data Leakage in Nlp Performance Evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, 2021.

[79] Caitlin A Farrell, Eric W Fleegler, Michael C Monuteaux, Celeste R Wilson, Cindy W Christian, and Lois K Lee. Community poverty and child abuse fatalities in the United States. *Pediatrics*, 139(5), 2017. doi: 10.1542/peds.2016-1616.

[80] Daniel J Feller, Jason Zucker, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T Yin, Peter Gordon, Noémie Elhadad, et al. 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, 2018. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371339/`.

[81] Hui Feng, Francesco Ronzano, Jude LaFleur, Matthew Garber, Rodrigo de Oliveira, Kathryn Rough, Katharine Roth, Jay Nanavati, Khaldoun Zine El Abidine, and Christina Mack. Evaluation of large language model performance on the Biomedical Language Understanding and Reasoning Benchmark. *medRxiv*, pages 2024–05, 2024.

[82] John R. Firth. A synopsis of linguistic theory, 1930–1955. In *Studies in Linguistic Analysis*, pages 10–32. Blackwell, 1957.

[83] Rudolph Flesch. A new Readability Yardstick. *Journal of Applied Psychology*, 32(3): 221–233, 1948. doi: 10.1037/h0057532.

[84] Centers for Disease Control and Prevention. Social Determinants of Health at Cdc, 2022. URL `https://www.cdc.gov/about/sdoh/index.html`.

[85] Giacomo Frisoni, Gianluca Moro, and Antonella Carbonaro. A survey on event extraction for natural language understanding: Riding the biomedical literature wave. *IEEE Access*, 9:160721–160757, 2021.

[86] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPtscore: Evaluate as You Desire. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365. URL https://aclanthology.org/2024.naacl-long.365/.

[87] Yujuan Fu, Giridhar Kaushik Ramachandran, Nicholas J. Dobbins, Namu Park, Michael Leu, Abby R. Rosenberg, Kevin Lybarger, Fei Xia, Özlem Uzuner, and Meliha Yetisgen. Extracting social determinants of health from pediatric patient notes using large language models: Novel corpus and methods. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7045–7056, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.618/.

[88] Yujuan Fu, Ozlem Uzuner, Meliha Yetisgen, and Fei Xia. Does Data Contamination Detection Work (Well) for Llms? A survey and Evaluation on Detection Assumptions. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5235–5256, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.291/.

[89] Yujuan Velvin Fu, Giridhar Kaushik Ramachandran, Ahmad Halwani, Bridget T McInnes, Fei Xia, Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. CACER: Clinical concept Annotations for Cancer Events and Relations. *Journal of the American Medical Informatics Association*, 31(11):2583–2594, 09 2024. ISSN 1527-974X. doi: 10.1093/jamia/ocae231. URL https://doi.org/10.1093/jamia/ocae231.

[90] Yujuan Velvin Fu, Giridhar Kaushik Ramachandran, Namu Park, Kevin Lybarger, Fei Xia, Ozlem Uzuner, and Meliha Yetisgen. BioMistral-Nlu: Towards More Generalizable Medical Language Understanding through Instruction Tuning. *AMIA 2025 Informatics Summit*, 2025.

[91] Haritha Gangavarapu, Giridhar Kaushik Ramachandran, Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. Adapting Biomedical Abstracts into Plain language using Large Language Models. *arXiv preprint arXiv:2501.15700*, 2025.

[92] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[93] Arvin Garg, Brian Jack, and Barry Zuckerman. Addressing the social determinants of health within the patient-centered medical home: lessons from pediatrics. *JAMA*, 309 (19):2001–2002, 2013. doi: 10.1001/jama.2013.1471.

[94] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *textitPloS One*, 13(2), 2018. doi: 10.1371/journal.pone.0192360.

[95] Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. Self-Verification Improves Few-Shot Clinical Information Extraction. *arXiv preprint arXiv:2306.00024*, 2023. doi: 10.48550/arXiv.2306.00024.

[96] Emily Getzen, Lyle Ungar, Danielle Mowery, Xiaoqian Jiang, and Qi Long. Mining for equitable health: Assessing the impact of missing data in electronic health records. *Journal of Biomedical Informatics*, 139:104269, 2023. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2022.104269. URL https://www.sciencedirect.com/science/article/pii/S153204642200274X.

[97] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, 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 medical education*, 9(1):e45312, 2023.

[98] Ethan Goh, Bryan Bunning, Elaine Khoong, Robert Gallo, Arnold Milstein, Damon Centola, and Jonathan H Chen. Chatgpt influence on medical decision-making, bias, and equity: a randomized study of clinicians evaluating clinical vignettes. *Medrxiv*, 2023.

[99] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

[100] Shahriar Golchin and Mihai Surdeanu. Data Contamination Quiz: A tool to Detect and Estimate Contamination in Large Language Models. *CoRR*, abs/2311.06233, 2023. doi: 10.48550/ARXIV.2311.06233. URL `https://doi.org/10.48550/arXiv.2311.06233`.

[101] Machine Learning Education Google. AUC (area under the roc curve). URL `https://developers.google.com/machine-learning/glossary#AUC`. Accessed: 2025-02-10.

[102] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[103] Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, and Hoifung

Poon. Distilling Large Language Models for Biomedical Knowledge Extraction: A case Study on Adverse Drug Events, 2023. URL https://arxiv.org/abs/2307.06439.

[104] Adi V Gundlapalli, Marjorie E Carter, Miland Palmer, Thomas Ginter, Andrew Redd, Steven Pickard, Shuying Shen, Brett South, Guy Divita, Scott Duvall, et al. 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, 2013. URL https://pubmed.ncbi.nlm.nih.gov/24551356/.

[105] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. Automated Lay Language Summarization of Biomedical Scientific Reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168, 2021.

[106] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892, 2012.

[107] Jacob Haimes, Cenny Wenner, Kunvar Thaman, Vassil Tashev, Clement Neo, Esben Kran, and Jason Schreiber. Benchmark inflation: Revealing llm performance gaps using retro-holdouts. *arXiv preprint arXiv:2410.09247*, 2024.

[108] Sifei Han, Robert F Zhang, Lingyun Shi, et al. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *\textitJ Biomed Inform*, 127:103984, 2022. doi: 10.1016/j.jbi.2021.103984.

[109] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. MedAlpaca–an open-source collection of medical conversational Ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[110] Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[111] Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, 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. *textitJMIR Med Inform*, 7(3):e13802, 2019. doi: 10.2196/13802.

[112] Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. A survey on recent advances in sequence labeling from deep learning models. *arXiv preprint arXiv:2011.06727*, 2020.

[113] Norris H Heintzelman, Robert J Taylor, Lone Simonsen, Roger Lustig, Doug Anderko, Jennifer A Haythornthwaite, Lois C Childs, and George Steven Bova. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *Journal of the American Medical Informatics Association*, 20(5):898–905, 2013.

[114] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[115] Marıa Herrero-Zazo, Isabel Segura-Bedmar, Paloma Mart\ ńez, and Thierry Declerck. The Ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[116] Karen Ho, Randi Zlotnik Shaul, Lee Ann Chapman, and Elizabeth Lee Ford-Jones. Standard of Care in Pediatrics: Integrating Family-Centred Care and Social Determinants of Health. *Healthc Q*, 19(1):55–60, 2016. doi: 10.12927/hcq.2016.24608.

[117] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[118] Danqing Hu, Bing Liu, Xiaofeng Zhu, Xudong Lu, and Nan Wu. Zero-shot information extraction from radiological reports using ChatGpt. *International Journal of Medical Informatics*, 183:105321, 2024.

[119] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LorA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[120] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[121] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[122] Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*, 2023.

[123] Shotaro Ishihara. Training Data Extraction From Pre-trained Language Models: A survey. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 260–275, 2023.

[124] Rezarta Islamaj, Chih-Hsuan Wei, David Cissel, Nicholas Miliaras, Olga Printseva, Oleg Rodionov, Keiko Sekiya, Janice Ward, and Zhiyong Lu. NLm-Gene, a richly annotated gold standard dataset for gene entities that addresses ambiguity and multi-species gene recognition. *Journal of biomedical informatics*, 118:103779, 2021.

[125] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308.

[126] Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MadE 1.0). *Drug safety*, 42:99–111, 2019.

[127] Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. Membership inference attack susceptibility of clinical language models. *arXiv preprint arXiv:2104.08305*, 2021.

[128] Kasper Jensen, Cristina Soguero-Ruiz, Karl Oyvind Mikalsen, Rolv-Ole Lindsetmo, Irene Kouskoumvekaki, Mark Girolami, Stein Olav Skrovseth, and Knut Magne Augestad. Analysis of free text in electronic health records for identification of cancer patient trajectories. *Scientific reports*, 7(1):46226, 2017.

[129] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput Surv*, 55(12):1–38, 2023. doi: 10.1145/3571730.

[130] Zongcheng Ji, Qiang Wei, and Hua Xu. Bert-based ranking for biomedical entity normalization. *AMIA Summits on Translational Science Proceedings*, 2020:269, 2020.

[131] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[132] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[133] Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A realistic Virtual Ehr Environment to Benchmark Medical Llm Agents, 2025. URL `https://arxiv.org/abs/2501.14654`.

[134] Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. Instruct and extract: Instruction tuning for on-demand information extraction. *arXiv preprint arXiv:2310.16040*, 2023.

[135] Di Jin and Peter Szolovits. Pico element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75, 2018.

[136] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[137] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQa: A dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.

[138] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MImiC-iIi, a freely accessible critical care database. *Sci. Data*, 3(1): 1–9, 2016. doi: 10.1038/sdata.2016.35.

[139] Martin Joos. Description of language design. *The Journal of the Acoustical Society of America*, 22(6):701–707, 1950. doi: 10.1121/1.1906660.

[140] Maël Jullien, Marco Valentino, and André Freitas. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. *arXiv preprint arXiv:2404.04963*, 2024.

[141] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. URL https://web.stanford.edu/~jurafsky/slp3/. Online manuscript released January 12, 2025.

[142] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pages 10697–10707. PMLR, 2022.

[143] Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. NUbiA: Neural Based Interchangeability Assessor for Text Generation. In Shubham Agarwal, Ondřej Dušek, Sebastian Gehrmann, Dimitra Gkatzia, Ioannis Konstas, Emiel Van Miltenburg, and Sashank Santhanam, editors, *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.evalnlgeval-1.4/.

[144] Tian Kang, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. UMls-based Data Augmentation for Natural Language Processing of Clinical Research Literature. *Journal of the American Medical Informatics Association*, 28(4):812–823, 2021. ISSN 1067-5027. doi: 10.1093/jamia/ocab003. URL https://doi.org/10.1093/jamia/ocab003.

[145] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess,

Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[146] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright Violations and Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458.

[147] Koya Kawashima, Wenjun Bai, and Changqin Quan. Text mining and pattern clustering for relation extraction of breast cancer and related genes. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 59–63. IEEE, 2017.

[148] 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. Psychosocial assessment as a standard of care in pediatric cancer. *Pediatr Blood Cancer*, 62(S5):S426–S459, 2015. doi: 10.1002/pbc.25730.

[149] Keylabs. Annotating Medical Text: Nlp Guidelines for Clinical Data, April 11 2025. URL https://keylabs.ai/blog/annotating-medical-text-nlp-guidelines-for-clinical-data/. Accessed: 2025-05-02.

[150] Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine. *IEEE Reviews in Biomedical Engineering*, 2025.

[151] Ritu Khare, Benjamin M. Good, Robert Leaman, Andrew I. Su, and Zhiyong Lu. Crowdsourcing in Biomedicine: Challenges and Opportunities. *Briefings in Bioin-*

formatics, 17(1):23–32, 2016. ISSN 1467-5463. doi: 10.1093/bib/bbv021. URL https://doi.org/10.1093/bib/bbv021.

[152] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in Nlp. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

[153] Katrin Klug, Katharina Beckh, Dario Antweiler, Nilesh Chakraborty, Giulia Baldini, Katharina Laue, René Hosch, Felix Nensa, Martin Schuler, and Sven Giesselbach. From admission to discharge: a systematic review of clinical natural language processing along the patient journey. *BMC Medical Informatics and Decision Making*, 24(1):238, 2024.

[154] Veysel Kocaman and David Talby. Biomedical named entity recognition at scale. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I*, pages 635–646. Springer, 2021.

[155] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The CheMdnEr corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7:1–17, 2015.

[156] Karen A Kuhlthau and James M Perrin. Child health status and parental employment. *Arch Pediatr Adolesc Med*, 155(12):1346–1350, 2001. doi: 10.1001/archpedi.155.12.1346.

[157] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[158] Wojciech Kusa, Harrisen Scells, Moritz Staudinger, and Allan Hanbury. Leveraging Cochrane Systematic Literature Reviews for Prospective Evaluation of Large Language Models. *The 1st Workshop on Data Contamination (CONDA)*, 2024.

[159] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of Open-Source Pretrained Large Language Models for Medical Domains. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.348. URL `https://aclanthology.org/2024.findings-acl.348/`.

[160] Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516, 2023.

[161] Samir D Ledade, Shishir N Jain, Ankit A Darji, and Vinodkumar H Gupta. Narrative writing: Effective ways and best practices. *Perspectives in clinical research*, 8(2):58, 2017.

[162] Ariel Lee, Cole Hunter, and Nataniel Ruiz. Platypus: Quick, Cheap, and Powerful Refinement of Llms. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

[163] H. Lee. The rise of ChatGpt: exploring its potential in medical education. *Anatomical Sciences Education*, 2023. doi: 10.1002/ase.2270.

[164] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBerT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

[165] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, 2022.

[166] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[167] Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. MedDm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*, 2023.

[168] Changmao Li and Jeffrey Flanigan. Task contamination: Language models may not be few-shot anymore. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18471–18480, 2024.

[169] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive Decoding: Open-ended Text Generation as Optimization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

[170] Yucheng Li. Estimating contamination via perplexity: Quantifying memorisation in language model evaluation. *arXiv preprint arXiv:2309.10677*, 2023.

[171] Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, 2024.

[172] Chin-Yew Lin. ROugE: A package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, 2004.

[173] Charlotta Lindvall, Chih-Ying Deng, Nicole D Agaronnik, Anne Kwok, Soujanya Samineni, Renato Umeton, Warren Mackie-Jenkins, Kenneth L Kehl, James A Tulsky, and Andrea C Enzinger. Deep learning for cancer symptoms monitoring on the basis of electronic health record unstructured clinical notes. *JCO Clinical Cancer Informatics*, 6:e2100136, 2022.

[174] 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. Evaluating the comparability of patient-level social risk data extracted from electronic health records: A systematic scoping review. *J. Health Inform.*, 29(3):14604582231200300, 2023. doi: 10.1177/14604582231200300.

[175] Todd Lingren, Louise Deleger, Katalin Molnar, Haijun Zhai, Jareen Meinzen-Derr, Megan Kaiser, Laura Stoutenborough, Qi Li, and Imre Solti. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Journal of the American Medical Informatics Association*, 21(3):406–413, 09 2013. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-001837. URL https://doi.org/10.1136/amiajnl-2013-001837.

[176] Chuang Liu, Renren Jin, Mark Steedman, and Deyi Xiong. Evaluating Chinese Large

Language Models on Discipline Knowledge Acquisition via Memorization and Robustness Assessment. In Oscar Sainz, Iker Garcıa Ferrero, Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yanai Elazar, and Yoav Goldberg, editors, *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 1–12, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.conda-1.1.

[177] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In Eneko Agirre, Marianna Apidianaki, and Ivan Vulić, editors, *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10/.

[178] Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang SU, Kao-Jung Chang, Wenting Chen, Haoliang Li, Linlin Shen, and Michael Lyu. Medchain: Bridging the Gap Between Llm Agents and Clinical Practice through Interactive Sequential Benchmarking, 2024. URL https://arxiv.org/abs/2412.01605.

[179] Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, et al. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*, 2024.

[180] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg Evaluation using Gpt-4 with Better Human Alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023.

Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

[181] Qiuhao Lu, Dejing Dou, and Thien Nguyen. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, 2022.

[182] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 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, August 2021. ACL. doi: 10.18653/v1/2021.acl-long.217. URL https://aclanthology.org/2021.acl-long.217.

[183] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*, 2022.

[184] Harrison C. Lucas, Jeffrey S. Upperman, and Jamie R. Robinson. A systematic review of large language models and their implications in medical education. *Medical Education*, 2024. doi: 10.1111/medu.15402.

[185] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRed: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282, 2022.

[186] Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, et al. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association*, page ocae037, 2024.

[187] Man Luo, Sharad Saxena, Swaroop Mishra, Mihir Parmar, and Chitta Baral. Biotabqa: Instruction learning for biomedical table question answering. *arXiv preprint arXiv:2207.02419*, 2022.

[188] Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. Extracting CovId-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *Journal of Biomedical Informatics*, 117:103761, 2021.

[189] Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631, 2021. doi: 10.1016/j.jbi.2020.103631.

[190] Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 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, April 2023. doi: 10.1093/jamia/ocad073. URL https://doi.org/10.1093/jamia/ocad073.

[191] Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. The 2022 n2c2/Uw shared task on extracting social determinants of health. *J Am Med Inform Assoc*, 30(8): 1367–1378, 04 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad012. URL https://doi.org/10.1093/jamia/ocad012.

[192] Chong Ma, Zihao Wu, Jiaqi Wang, Shaochen Xu, Yaonai Wei, Zhengliang Liu, Fang Zeng, Xi Jiang, Lei Guo, Xiaoyan Cai, et al. An iterative optimizing framework for radiology report summarization with ChatGpt. *IEEE Transactions on Artificial Intelligence*, 5(8):4163–4175, 2024.

[193] Mingyu Derek Ma, Alexander Taylor, Wei Wang, and Nanyun Peng. 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, July 2023. ACL. doi: 10.18653/v1/2023.acl-long.886. URL https://aclanthology.org/2023.acl-long.886.

[194] Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! *arXiv preprint arXiv:2303.08559*, 2023. doi: https://doi.org/10.48550/arXiv.2303.08559.

[195] Inbal Magar and Roy Schwartz. Data Contamination: From Memorization to Exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, 2022.

[196] Diwakar Mahajan, Jennifer J Liang, Ching-Huei Tsou, and Özlem Uzuner. Overview of the 2022 n2c2 shared task on contextualized medication event extraction in clinical notes. *Journal of biomedical informatics*, 144:104432, 2023.

[197] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOfu: A task of Fictitious Unlearning for Llms, 2024.

[198] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. LLm Dataset Inference: Did you train on my dataset? *The 1st Workshop on Data Contamination (CONDA)*, 2024.

[199] Andrew M Malty, Sandeep K Jain, Peter C Yang, Krysten Harvey, and Jeremy L Warner. Computerized approach to creating a systematic ontology of hematology/oncology regimens. *JCO clinical cancer informatics*, 2:1–11, 2018.

[200] Marc Marone and Benjamin Van Durme. Data Portraits: Recording Foundation Model Training Data. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 15121–15135. Curran Associates,

Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3112ee706d21d734c15532c1239773e1-Paper-Datasets_and_Benchmarks.pdf.

[201] Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership Inference Attacks against Language Models via Neighbourhood Comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, 2023.

[202] R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RavEn. *Transactions of the Association for Computational Linguistics*, 11:652–670, 06 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00567. URL https://doi.org/10.1162/tacl_a_00567.

[203] R Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11:652–670, 2023.

[204] Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 2369–2385, 2024.

[205] Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre de Montjoye. Copyright Traps for Large Language Models. In *Forty-first International Conference on Machine Learning*, 2024.

[206] Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, and Yves-Alexandre de Montjoye. Sok: Membership Inference Attacks on Llms are Rushing Nowhere (and How to Fix It). *arXiv preprint arXiv:2406.17975*, 2024.

[207] Behzad Mehrbakhsh, Dario Garigliotti, Fernando Martınez-Plumed, and Jose Hernandez-Orallo. Confounders in Instance Variation for the Analysis of Data Contamination. In Oscar Sainz, Iker Garcıa Ferrero, Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yanai Elazar, and Yoav Goldberg, editors, *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 13–21, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.conda-1.2.

[208] Antonio Miranda-Escalada, Farrokh Mehryary, Jouni Luoma, Darryl Estrada-Zavala, Luis Gasco, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. Overview of DrugProt task at BioCreative Vii: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations. *Database*, 2023: baad080, 2023.

[209] Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.570. URL https://aclanthology.org/2022.emnlp-main.570.

[210] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning*, pages 24950–24962. PMLR, 2023.

[211] Alexander A Morgan, Zhiyong Lu, Xinglong Wang, Aaron M Cohen, Juliane Fluck, Patrick Ruch, Anna Divoli, Katrin Fundel, Robert Leaman, Jörg Hakenberg, et al. Overview of BioCreative Ii gene normalization. *Genome biology*, 9:1–19, 2008.

[212] Jennifer Morone. An Integrative Review of Social Determinants of Health Assessment and Screening Tools Used in Pediatrics. *J Pediatr Nurs*, 37:22–28, 2017. ISSN 0882-5963.

doi: https://doi.org/10.1016/j.pedn.2017.08.022. URL https://www.sciencedirect.com/science/article/pii/S0882596317302920. Special Issue: Social Determinants of Health.

[213] Amol S Navathe, Feiran Zhong, Victor J Lei, Frank Y Chang, Margarita Sordo, Maxim Topaz, Shamkant B Navathe, Roberto A Rocha, and Li Zhou. Hospital readmission and social risk factors identified from physician notes. *Health Serv. Res.*, 53(2):1110–1136, 2018. doi: 10.1111/1475-6773.12670.

[214] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

[215] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, Kia Khezeli, and Parisa Rashidi. Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine*, 154:102900, 2024. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2024.102900. URL https://www.sciencedirect.com/science/article/pii/S0933365724001428.

[216] An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 299. NIH Public Access, 2017.

[217] Satoshi Nishioka, Masaki Asano, Shuntaro Yada, Eiji Aramaki, Hiroshi Yajima, Yuki Yanagisawa, Kyoko Sayama, Hayato Kizaki, and Satoko Hori. Adverse event signal extraction from cancer patients' narratives focusing on impact on their daily-life activities. *Scientific reports*, 13(1):15516, 2023.

[218] Office of the National Coordinator for Health Information Technol-

ogy. National Trends in Hospital and Physician Adoption of Electronic Health Records, 2021. URL https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records. Accessed: 2025-04-28.

[219] Office of the National Coordinator for Health Information Technology. Electronic Health Records, 2025. URL https://www.healthit.gov/playbook/electronic-health-records/. Accessed: 2025-04-28.

[220] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 Olmo 2 Furious. 2024. URL https://arxiv.org/abs/2501.00656.

[221] Mahmud Omar, Dana Brin, Benjamin Glicksberg, and Eyal Klang. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *American Journal of Infection Control*, 2024.

[222] Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. Interrater disagreement resolution: A systematic procedure to reach consensus in annotation tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, 2021.

[223] OpenAI. Models - OpenAi Api. URL https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed on Oct 6, 2024.

[224] OpenAI. Hello Gpt-4o, may 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed on Oct 6, 2024.

[225] OpenAI. GPt-4o mini: advancing cost-efficient intelligence, july 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed on Oct 6, 2024.

[226] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel

Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech

Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPt-4 Technical Report, 2024. URL https://arxiv.org/abs/2303.08774.

[227] Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving Test Set Contamination in Black-Box Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

[228] World Health Organization et al. The Icd-10 classification of mental and behavioral disorders. *Clinical descriptions and diagnostic guidelines*, 1992.

[229] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[230] Medha Palavalli, Amanda Bertsch, and Matthew Gormley. A taxonomy for Data Contamination in Large Language Models. In Oscar Sainz, Iker Garcıa Ferrero, Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yanai Elazar, and Yoav Goldberg, editors, *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 22–40, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.conda-1.3.

[231] Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. Privacy Risks of General-Purpose Language Models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331, 2020. doi: 10.1109/SP40000.2020.00095.

[232] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLm Evaluators Recognize and

Favor Their Own Generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4NJBV6Wp0h.

[233] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[234] Namu Park, Kevin Lybarger, Giridhar Kaushik Ramachandran, Spencer Lewis, Aashka Damani, Ozlem Uzuner, Martin Gunn, and Meliha Yetisgen-Yildiz. A novel Corpus of Annotated Medical Imaging Reports and Information Extraction Results Using BerT-based Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1280–1292, 2024.

[235] Braja G Patra, Mohit M Sharma, Veer Vekaria, et al. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *textitJ Am Med Inform Assoc*, 28(12):2716–2727, 2021. doi: 10.1093/jamia/ocab170.

[236] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BerT and Elmo on Ten Benchmarking Datasets. *BioNLP 2019*, page 58, 2019.

[237] Sundar Pichai and Demis Hassabis. Our next-generation model: Gemini 1.5, feb 2024. URL https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/. Accessed on Oct 6, 2024.

[238] Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. The RooTs Search Tool: Data Transparency for Llms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, 2023.

[239] Xiao-Ou Ping, Yi-Ju Tseng, Yufang Chung, Ya-Lin Wu, Ching-Wei Hsu, Pei-Ming Yang, Guan-Tarn Huang, Feipei Lai, and Ja-Der Liang. Information extraction for tracking liver cancer patients'statuses: from mixture of clinical narrative report types. *TELEMEDICINE and e-HEALTH*, 19(9):704–710, 2013.

[240] Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. Semeval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 54–62, 2014.

[241] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[242] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res*, 21(1):5485–5551, 2020. doi: https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

[243] Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nic Dobbins, Ozlem Uzuner, and Meliha Yetisgen. Prompt-based Extraction of Social Determinants of Health Using Few-shot Learning. In Tristan Naumann, Asma Ben Abacha, Steven Bethard, Kirk Roberts, and Anna Rumshisky, editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 385–393, Toronto, Canada, July 2023.

Association for Computational Linguistics. doi: 10.18653/v1/2023.clinicalnlp-1.41. URL https://aclanthology.org/2023.clinicalnlp-1.41.

[244] Federico Ranaldi, Elena Sofia Ruzzetti, Dario Onorati, Leonardo Ranaldi, Cristina Giannone, Andrea Favalli, Raniero Romagnoli, and Fabio Massimo Zanzotto. Investigating the Impact of Data Contamination of Large Language Models in Text-to-Sql Translation. *arXiv preprint arXiv:2402.08100*, 2024.

[245] Prashanth Rawla. Epidemiology of prostate cancer. *World journal of oncology*, 10(2): 63, 2019.

[246] Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of Pretraining Term Frequencies on Few-Shot Numerical Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, 2022.

[247] Olesya Razuvayevskaya, Irene Lopez, Ian Dunham, and David Ochoa. Why clinical trials stop: the role of genetics. *medRxiv*, pages 2023–02, 2023.

[248] Ehud Reiter and Robert Dale. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[249] Russell Richie, Victor M Ruiz, Sifei Han, Lingyun Shi, and Fuchiang Tsui. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. *J Am Med Inform Assoc*, page ocad046, 2023. doi: 10.1093/jamia/ocad046.

[250] Martin Riddell, Ansong Ni, and Arman Cohan. Quantifying contamination in evaluating code generation capabilities of language models. *arXiv preprint arXiv:2403.04811*, 2024.

[251] Brian Romanowski, Asma Ben Abacha, and Yadan Fan. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *J Am Med Inform Assoc*, page ocad071, 2023. doi: 10.1093/jamia/ocad071.

[252] Arpita Roy and Shimei Pan. Incorporating medical knowledge in BerT for clinical relation extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 5357–5366, 2021.

[253] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of Gemini Models in Medicine, 2024.

[254] Satya S Sahoo, Joseph M Plasek, Hua Xu, Özlem Uzuner, Trevor Cohen, Meliha Yetisgen, Hongfang Liu, Stéphane Meystre, and Yanshan Wang. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *Journal of the American Medical Informatics Association*, 31(9):2114–2124, 2024.

[255] Sunil Kumar Sahu, Ashish Anand, Krishnadev Oruganty, and Mahanandeeshwar Gattu. Relation extraction from clinical texts using domain invariant convolutional neural network. *arXiv preprint arXiv:1606.09370*, 2016.

[256] Oscar Sainz, Jon Campos, Iker Garcıa-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLp Evaluation in trouble: On the Need to Measure Llm Data

Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, 2023.

[257] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. Did ChatGpt cheat on your test?, Jun 2023. URL `https://hitz-zentroa.github.io/lm-contamination/blog/`. Accessed: 2024-09-09.

[258] Oscar Sainz, Iker Garcıa-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. *arXiv preprint arXiv:2310.03668*, 2023.

[259] Guergana K Savova, Philip V Ogren, Patrick H Duffy, James D Buntrock, and Christopher G Chute. Mayo clinic Nlp system for patient smoking status identification. *J Am Med Inform Assoc*, 15(1):25–28, 2008. doi: 10.1197/jamia.M2437.

[260] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTakEs): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[261] Volker Schirrmacher. From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment. *International journal of oncology*, 54(2):407–419, 2019.

[262] Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583,

Tokyo, Japan, September 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.inlg-main.44/.

[263] Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. *The 1st Workshop on Data Contamination (CONDA)*, 2024.

[264] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLeuRt: Learning Robust Metrics for Text Generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

[265] Shahid Munir Shah and Rizwan Ahmed Khan. Secondary use of electronic health record: Opportunities and challenges. *IEEE access*, 8:136947–136965, 2020.

[266] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting Pretraining Data from Large Language Models. In *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[267] Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. A systematic review of large language model (Llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117, 2025.

[268] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: a cancer journal for clinicians*, 72(1), 2022.

[269] Ravindra Singh and Naurang Singh Mangat. Stratified Sampling. In *Elements of Survey Sampling*, pages 102–144. Springer, 1996.

[270] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[271] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL https://doi.org/10.1038/s41591-024-03423-7.

[272] G.G.R. Sng, J.Y.M. Tung, D.Y.Z. Lim, and Y.M. Bee. Potential and pitfalls of ChatGpt and natural-language artificial intelligence models for diabetes education. *Diabetes Care*, 46(5):e103–e105, 2023. doi: 10.2337/dc23-0197.

[273] Congzheng Song and Vitaly Shmatikov. Auditing Data Provenance in Text-Generation Models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 196–206, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330885. URL https://doi.org/10.1145/3292500.3330885.

[274] Ali Soroush, Benjamin S. Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W. Charney, Girish N Nadkarni, and Eyal Klang. Large Language Models Are Poor Medical Coders —Benchmarking of Medical Code Querying. *NEJM AI*, 1(5): AIdbp2300040, 2024. doi: 10.1056/AIdbp2300040. URL https://ai.nejm.org/doi/abs/10.1056/AIdbp2300040.

[275] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. CLamP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336, 2018.

[276] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[277] Paul Stallard, Philip Norman, Sarah Huline-Dickens, Emma Salter, and Jan Cribb. 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, 2004. doi: 10.1177/1359104504039767.

[278] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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, April 2012. ACL. URL `https://aclanthology.org/E12-2021`.

[279] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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 Association for Computational Linguistics*, pages 102–107, 2012.

[280] Shane Storks, Qiaozi Gao, and Joyce Yue Chai. Recent Advances in Natural Language Inference: A survey of Benchmarks, Resources, and Approaches. *arXiv: Computation and Language*, 2019. URL `https://api.semanticscholar.org/CorpusID:213613608`.

[281] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for

de-identification: The 2014 i2b2/Uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

[282] Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 09 2019. ISSN 1527-974X. doi: 10.1093/jamia/ocz163. URL https://doi.org/10.1093/jamia/ocz163.

[283] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813, 04 2013. ISSN 1067-5027. doi: 10.1136/amiajnl-2013-001628. URL https://doi.org/10.1136/amiajnl-2013-001628.

[284] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the ShaRe/CleF eHealth evaluation lab 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4*, pages 212–231. Springer, 2013.

[285] Sandra Susanibar-Adaniya and Stefan K Barta. 2021 update on diffuse large B cell lymphoma: a review of current data and potential applications on risk stratification and management. *American journal of hematology*, 96(5):617–629, 2021.

[286] Wilson L. Taylor. "Cloze Procedure": A new Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433, 1953.

[287] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy,

Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov,

Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Am-

brose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost

van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang

Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael

Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal,

Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani

Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang,

Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of Highly Capable Multimodal Models, 2024. URL `https://arxiv.org/abs/2312.11805`.

[288] Wen\text-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. The aci demo corpus: An open dataset for benchmarking the state-of-the-art for automatic note generation from doctor-patient conversations. *Accepted by Nature Scientific Data*, 2023.

[289] Wen\text-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and

Fei Xia. DermaVqa: A multilingual Visual Question Answering Dataset for Dermatology. *the 27th International Conference on Medical Image Computing and Computer Assisted Intervention*, 2024.

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

[291] Ross Thompson, Paul H Dworkin, Georgina Peacock, Mary Ann McCabe, John K Iskander, Phoebe Thorpe, and Susan Laird. Addressing health disparities in early childhood, 2016. URL `https://www.cdc.gov/grand-rounds/pp/2016/20160315-childhood-development.html`.

[292] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

[293] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[294] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BioAsq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16:1–28, 2015.

[295] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan

Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct Distillation of Lm Alignment, 2023.

[296] Grace Turner, J Chang, N Dorvall, et al. Domain adaptation of a deep learning symptom extractor for different patient populations and clinical settings. *AMIA 2021 Informatics Summit*, 2022.

[297] Özlem Uzuner. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570, 2009.

[298] Özlem Uzuner, Yuan Luo, and Peter Szolovits. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5):550–563, 2007.

[299] Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24, 2008.

[300] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[301] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/Va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[302] Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5): 786–791, 2012.

[303] Erik M. van Mulligen, Annie Fourrier-Reglat, David Gurwitz, Mariam Molokhia, Ainhoa Nieto, Gianluca Trifiro, Jan A. Kors, and Laura I. Furlong. The EU-Adr corpus: Annotated drugs, diseases, targets, and their relationships. *Journal of Biomedical Informatics*, 45(5):879–884, 2012. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2012.04.004. URL https://www.sciencedirect.com/science/article/pii/S1532046412000573. Text Mining and Natural Language Processing in Pharmacogenomics.

[304] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[305] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[306] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIder: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[307] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *arXiv preprint arXiv:2306.07899*, 2023.

[308] Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Robert Stewart, Rashmi Patel, and Sumithra Velupillai. Annotating Temporal Relations to Determine the Onset of Psychosis Symptoms. In *MedInfo*, pages 418–422, 2019.

[309] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[310] Chenguang Wang, Xiao Liu, Zui Chen, Haoyun Hong, Jie Tang, and Dawn Song. DeepStruct: Pretraining of language models for structure prediction. *arXiv preprint arXiv:2205.10475*, 2022.

[311] Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. *arXiv preprint arXiv:2401.16107*, 2024.

[312] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGpt a Good Nlg Evaluator? A preliminary Study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.newsum-1.1. URL https://aclanthology.org/2023.newsum-1.1/.

[313] Shuhe Wang, Shengyu Zhang, Jie Zhang, Runyi Hu, Xiaoya Li, Tianwei Zhang, Jiwei Li, Fei Wu, Guoyin Wang, and Eduard Hovy. Reinforcement learning enhanced llms: A survey. *arXiv preprint arXiv:2412.10400*, 2024.

[314] Siyuan Wang, Zheng Liu, and Bo Peng. A self-training Framework for Automated Medical Report Generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16443–16449, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1024. URL https://aclanthology.org/2023.emnlp-main.1024/.

[315] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li,

and Yixuan Yuan. A survey of Llm-based Agents in Medicine: How far are we from Baymax?, 2025. URL https://arxiv.org/abs/2502.11211.

[316] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. InstructUie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

[317] Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. Automated extraction of substance use information from clinical texts. In *AMIA Annu Symp Proc*, volume 2015, page 2121. AMIA, 2015. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4765598/.

[318] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2017.11.011. URL https://www.sciencedirect.com/science/article/pii/S1532046417302563.

[319] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. MedSts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72, 2020. ISSN 1574-0218. doi: 10.1007/s10579-018-9431-1. URL https://doi.org/10.1007/s10579-018-9431-1.

[320] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pages 804–823. PMLR, 2023.

[321] Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, and Zhiyong Lu.

tmVar 2.0: integrating genomic variant information from literature with dbSnp and ClinVar for precision medicine. *Bioinformatics*, 34(1):80–87, 2018.

[322] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[323] Johnny Wei, Ryan Wang, and Robin Jia. Proving membership in Llm pretraining data via data watermarks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 13306–13320, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.788. URL https://aclanthology.org/2024.findings-acl.788.

[324] Yixuan Weng, Minjun Zhu, Shizhu He, Kang Liu, and Jun Zhao. Large language models are reasoners with self-verification. *arXiv preprint arXiv:2212.09561*, 2022. doi: https://doi.org/10.48550/arXiv.2212.09561.

[325] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

[326] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. PMc-LlaMa: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.

[327] Jinge Wu, Rowena Smith, and Honghan Wu. Adverse Childhood Experiences Identification from Clinical Notes with Ontologies and Nlp, 2022.

[328] Jinge Wu, Rowena Smith, and Honghan Wu. Ontology-Driven Self-Supervision for Adverse Childhood Experiences Identification Using Social Media Datasets, 2022.

[329] Stephen Wu and Hongfang Liu. Semantic characteristics of Nlp-extracted concepts in clinical notes vs. biomedical literature. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1550. American Medical Informatics Association, 2011.

[330] Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470, 2020.

[331] Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Zhenxi Lin, Jie Yang, Shuang Zhao, and Yefeng Zheng. MedJourney: Benchmark and Evaluation of Large Language Models over Patient Clinical Journey. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 87621–87646. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9f80af32390984cb709cdeb014d0df41-Paper-Datasets_and_Benchmarks_Track.pdf.

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

System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

[333] Fei Xia and Meliha Yetisgen-Yildiz. Clinical corpus annotation: challenges and strategies. In *Proceedings of the third workshop on building and evaluating resources for biomedical text mining (BioTxtM'2012) in conjunction with the international conference on language resources and evaluation (LREC), Istanbul, Turkey*, pages 21–27, 2012.

[334] Qian-Wen Xie, Xiangyan Luo, Roujia Chen, and Xudong Zhou. Associations Between Parental Employment and Children's Screen Time: A longitudinal Study of China Health and Nutrition Survey. *Int. J. Public Health*, 67:1605372, 2023. doi: 10.3389/ijph.2022.1605372.

[335] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Huan He, Lucila Ohno-Machido, Yonghui Wu, Hua Xu, and Jiang Bian. Me LlaMa: Foundation Large Language Models for Medical Applications, 2024.

[336] Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark Data Contamination of Large Language Models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

[337] Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. Exploring the relationship between electronic health records and provider burnout: a systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021, 2021.

[338] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. Towards Interpretable Mental Health Analysis with Large Language Models. *arXiv preprint arXiv:2304.03347*, 2023. doi: https://doi.org/10.48550/arXiv.2304.03347.

[339] Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking Benchmark and Contamination for Language Models with Rephrased Samples, 2023.

[340] Shunyu Yao. The second half. https://ysymyth.github.io/The-Second-Half/, 2025. Accessed: 2025-05-20.

[341] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed Levitated Marker for Entity and Relation Extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4904–4917. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.337.

[342] Meliha Yetisgen and Lucy Vanderwende. 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, 2017. URL https://link.springer.com/chapter/10.1007/978-3-319-59758-4_18.

[343] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023.

[344] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, and Meliha Yetisgen. To err is human, how about medical large language models? comparing pre-trained language models for medical assessment errors and reliability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16211–16223, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1409/.

[345] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, and Meliha Yetisgen. To Err Is Human, How about Medical Large Language Models? Comparing Pre-trained Language Models for Medical Assessment Errors and Reliability. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16211–16223, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1409.

[346] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*, 2020.

[347] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*, 2023.

[348] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. In *International Conference on Machine Learning*. PMLR, 2024.

[349] Zehao Yu, Xi Yang, Chong Dang, et al. 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, 2021. URL https://pubmed.ncbi.nlm.nih.gov/35309014/.

[350] Zehao Yu, Xi Yang, Yi Guo, et al. Assessing the Documentation of Social Determinants of Health for Lung Cancer Patients in Clinical Narratives. *textitFront Public Health*, 10, 2022. doi: 10.3389/fpubh.2022.778463.

[351] Weizhe Yuan, Graham Neubig, and Pengfei Liu. BArtScore: Evaluating Generated Text as Text Generation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Informa-*

tion Processing Systems, volume 34, pages 27263–27277. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf.

[352] Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

[353] Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor R¨uhle, Andrew Paverd, Olga Ohrimenko, Boris K¨opf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 363–375, 2020.

[354] Zexian Zeng, Xiaoyu Li, Sasa Espino, Ankita Roy, Kristen Kitsch, Susan Clare, Seema Khan, and Yuan Luo. Contralateral breast cancer event detection using nature language processing. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1885. American Medical Informatics Association, 2017.

[355] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating Factual Consistency with A unified Alignment Function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

[356] Haijun Zhai, Todd Lingren, Louise Deleger, Qi Li, Megan Kaiser, Laura Stoutenborough, and Imre Solti. Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing. *Journal of Medical Internet Research*, 15(4):e73, Apr 2013. ISSN 1438-8871. doi: 10.2196/jmir.2426. URL https://doi.org/10.2196/jmir.2426.

[357] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *Advances in Neural Information Processing Systems*, 36:39321–39362, 2023.

[358] Mozhi Zhang, Hang Yan, Yaqian Zhou, and Xipeng Qiu. Promptner: A prompting method for few-shot named entity recognition via k nearest neighbor search. *arXiv preprint arXiv:2305.12217*, 2023.

[359] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[360] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

[361] Jin Zhao, Chao Liu, Jiaqing Liang, Zhixu Li, and Yanghua Xiao. A novel Cascade Instruction Tuning Method for Biomedical Ner. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11701–11705. IEEE, 2024.

[362] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[363] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Asso-

ciation for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.131. URL https://aclanthology.org/2022.emnlp-main.131/.

[364] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

[365] Sitong Zhou, Kevin Lybarger, Meliha Yetisgen, and Mari Ostendorf. Generalizing through Forgetting-Domain Generalization for Symptom Event Extraction in Clinical Notes. *AMIA Summits on Translational Science Proceedings*, 2023:622, 2023.

[366] Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. UniversalNer: Targeted Distillation from Large Language Models for Open Named Entity Recognition. 2023.

[367] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022. doi: https://doi.org/10.48550/arXiv.2212.09561.

[368] Wenhong Zhu, Hongkun Hao, Zhiwei He, Yun-Ze Song, Jiao Yueyang, Yumeng Zhang, Hanxu Hu, Yiran Wei, Rui Wang, and Hongyuan Lu. CLeaN–EvaL: Clean Evaluation on Contaminated Large Language Models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 835–847, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.53. URL https://aclanthology.org/2024.findings-naacl.53.

[369] Yongshuo Zong, Tingyang Yu, Bingchen Zhao, Ruchika Chavhan, and Timothy Hospedales. Fool your (vision and) language model with embarrassingly simple permutations. *arXiv preprint arXiv:2310.01651*, 2023.

Appendix A
NLP BACKGROUND

In this chapter, we first present a comprehensive overview of important tasks in clinical NLP. We then introduce two widely adopted categories of transformer-based systems: Bidirectional Encoder Representations from Transformers (BERT), a model known for its strong performance on language understanding tasks, and decoder-only generative large language models (LLMs), which are designed to produce coherent and contextually appropriate text based on input prompts. Finally, we examine the importance of benchmark development, outlining key considerations such as dataset representativeness, task diversity, and evaluation metrics, which are essential for assessing and comparing model performance in clinical applications.

A.1 Clinical NLP Tasks

In this section, we categorize the clinical NLP tasks into two categories: Natural Language Understanding (NLU) and Natural Language Generation (NLG).

A.1.1 Natural Language Understanding (NLU) Tasks in the Clinical Domain

Clinical narratives contain fine-grained, context-rich information beyond standardized, structured EHR components [153, 138]. For example, those unstructured EHRs - such as discharge summaries, progress notes, and radiology reports - often include symptom descriptions [365], temporal relationships [283], and complex clinical reasoning [31]. However, this semantic variation and highly specialized language in free-text clinical narratives pose challenges for large-scale data retrieval and analysis [153].

To harness the rich information embedded in both clinical narratives, medical Natural

Language Understanding (NLU) employs NLP techniques to interpret and extract meaningful data from unstructured texts [254, 1]. Here, we introduce some fundamental NLU tasks included in the medical Language Understanding and Reasoning Benchmark (BLURB), a widely used dataset in biomedical NLU [102], as well as the *Speech and Language Processing* textbook [141]

Named Entity Recognition (NER) is a task to label meaningful text spans (named entities) with their type [141]. NER can identify medication mentions [300] for pharmacovigilance and person names [298] for de-identification. For example, in the sentence, "*Anna is taking ibuprofen for pain relief after removing her wisdom teeth.*", an NER system would detect "*Anna*" as a person name and *"ibuprofen"* as a medication. A related task is medical **event extraction**, which extracts named entities as events, and assigns fine-grain attributes to them [196, 114].

Relation Extraction (RE) is the task of identifying meaningful relationships between the recognized entities in text [141]. For example, important relations include temporal relations for treatment timeline [283], condition-treatment relations [301], as well as adverse drug reactions [106]. For example, from the sentence, "*Acetaminophen heals his fever.*", a relation extraction system should infer a drug–disease treatment relationship between "*Acetaminophen*" and "*fever*".

Sentence and Document Classification is the task of assigning a meaningful label to a text snippet [141]. For example, it can be used to identify a patient's smoking status [299], recognize obesity and its co-morbidities [297], or determine eligibility for inclusion in a research cohort based on clinical text [282].

Multiple-Choice Question Answering (MCQA) is the task of selecting the correct answer based on a given question, context, and a set of candidate options. This task is commonly used to evaluate an NLP system's ability to comprehend information, perform reasoning, and possibly demonstrate its internal medical knowledge. Examples of popular MCQA datasets are PubMedQA [137], derived from PubMed abstracts, and MedQA-USMLE [136], based on U.S. medical licensing examinations.

Semantic Text similarity is the task of evaluating how closely two text snippets align in meaning by assigning a semantic similarity score [319]. In the medical domain, this task is useful for reducing redundancy and retrieving relevant clinical or research records [319].

Natural Language Inference (NLI) is the task that involves determining the logical relationship between two text snippets, typically whether one sentence *entails*, *contradicts*, or is *neutral* with respect to the other [141]. NLI is a common benchmark for evaluating language understanding and reasoning abilities.

A.1.2 Natural Language Generation (NLG) Tasks in the Clinical Domain

Natural Language Generation (NLG) is the task of producing human-like text from a given input representation [248, 72]. In this section, we discuss several key tasks in clinical NLG. We focus on NLG applications where the input is only in natural language. Meanwhile, all NLU tasks can be translated into NLG tasks, by expressing the desired labels or outputs in natural language. We will discuss more about NLU tasks in the NLG format in Chapter 5.

Clinical decision support can provide healthcare professionals with relevant, evidence-based information and insights [179], such as symptom analysis [311], risk assessment [53], diagnostic reasoning [331], and treatment recommendation [167].

Clinical report generation can help doctors in multiple ways. For example, the automatic translation of doctor-patient dialogues into structured clinical notes, can substantially reduce the time clinicians spend on documentation [288]. Summarization of existing clinical reports can decrease the cognitive load on healthcare providers by distilling essential information, thereby facilitating quicker decision-making [192]. Additionally, NLP systems can improve the quality of clinical text by detecting and correcting medical errors [4].

As summarized in a 2024 survey by Lucas et al. [184], **medical education** is another critical type of clinical NLG task. For instance, ChatGPT and similar LLMs have been explored for their potential to assist in curriculum design [163], explain complex medical concepts to students [8], and generate disease management instructions for patient education [272].

Recently, as NLP systems are getting more powerful, the NLG tasks have been expanded across the entire patient hospital journey. For example, the wide range of applications can include hospital reception and triage support [331], specialty referral [178], insurance-related QA [331], and retrieval of patient information and lab results from EHRs [133],

Recently, as NLP systems have become increasingly powerful, Natural Language Generation (NLG) tasks have expanded to cover the entire patient hospital journey. These applications now encompass a wide range of interactions, including hospital reception and triage support [331], specialty referral management [178], insurance-related question answering [331], and the retrieval and summarization of patient information and lab results from clinical narratives [133].

A.2 Transformer-based Language Models for Clinical NLP

Transformer-based NLP systems have advanced healthcare AI through a wide range of applications, including all tasks discussed in Section A.1.1 [305, 215]. The transformer is a type of deep neural network architecture, which was first introduced by Vaswani et al. in their landmark 2017 paper, "Attention Is All You Need" [305]. Its key contribution is the self-attention mechanism, which models the relative importance of each token within the input sequence and effectively captures contextual relationships across both short and long distances. Additionally, the use of multi-head attention facilitates parallelization and efficient batch computation, making transformers highly scalable for large datasets and complex tasks.

The transformer architecture contains two main components: the encoder and the decoder. The encoder processes input sequences to learn contextual representations, while the decoder generates output sequences based on both the encoder's output and previously generated tokens. The two components can be used independently, depending on the task.

In this section, we focus on the two most widely used types of transformer-based language models (LMs): (1) Bidirectional Encoder Representations from Transformers (BERT), which only the encoder is optimized for text understanding tasks, as described in Section A.1.1 [66]; and (2) generative large language models (LLMs), which are usually only the decoder,

trained on vast and diverse datasets and serve as foundation models capable of adapting to various downstream tasks, especially NLG tasks as described in Section A.1.2 [36]. We will discuss their architectures, training procedures, and applications.

A.2.1 BERT

The application of BERT-based LMs typically follows a three-step procedure: vocabulary construction, pre-training, and fine-tuning [66]. The vocabulary construction allows each word to be split into one or multiple tokens from a fixed token vocabulary, and can thus effectively handle infrequent, out-of-vocabulary words. While the unsupervised, pre-training phase enables the model to learn a fundamental understanding of semantics, the supervised, fine-tuning phase learns about a specific language understanding task.

In the **vocabulary construction** step, BERT employs a WordPiece tokenizer [332], which starts with a base vocabulary of all individual characters, and merges frequent pairs of characters or subwords in a corpus, until the desired vocabulary size is reached. The WordPiece tokenizer can balance coverage and efficiency. For example, the initial BERT model contains around 30,000 tokens [66].

In the **pre-training** step, the BERT is exposed to large-scale unlabeled corpora, with two learning objectives: masked language modeling (MLM) and next sentence prediction (NSP). The MLM involves randomly masking a subset of input tokens and training the model to predict these masked tokens based on their context, which is often referred to as the *Cloze* task [286]. This objective also relates to the *distributional hypothesis*, which assumes that similar words usually occur in the same context [139, 110, 82, 141]. On the other hand, the NSP is a binary classification task, where the model is presented with pairs of sentences and trained to predict whether the second sentence logically follows the first in the original text. This task allows BERT to learn the relation between two sentences, helping downstream tasks such as logical reasoning tasks, such as QA and NLI.

For example, PubMedBERT is pre-trained from scratch on 21 GB of PubMed abstracts to better model biomedical language [102]. Instead from pre-training from scratch, we can

adapt BERT for a specific domain by continuing pre-training on domain-specific data. For example, the Bio+ClinicalBERT [12] adapts BioBERT [164] for clinical NLP tasks, by further pre-training BioBERT on de-identified clinical notes from the MIMIC-III dataset [138].

After adding a task-specific neural layer on top, a pre-trained BERT model is then **fine-tuned** for specific tasks. This architecture can predict categorical or numerical labels for each input sequence or token. Thus, this supervised fine-tuning phase allows BERT to be adapted to any of the NLU tasks described in Section A.1.1. However, fine-tuning is typically performed for a single task or dataset, and the model may experience a severe performance drop when applied to data with different distributions. For instance, a BERT model fine-tuned for NER on clinical notes from one clinical specialty may perform poorly when applied to notes from another clinical specialty [365].

A.2.2 Decorder-only Large Language Models (LLMs)

Generative LLMs typically employ a decoder-only transformer architecture and generate natural text, token by token, in an auto-regressive manner. In this setup, each token is produced by conditioning on the previously generated tokens, allowing the model to capture logical dependencies in the prior context. This architecture has been widely deployed in LLMs such as ChatGPT [2], GPT-4 [226], Gemini [287], and LLaMA [20].

Similar to the development of BERT-based language models, the development of decoder-only LLMs can be divided into two main phases: pre-training [214] and post-training [157]. The pre-training phase is unsupervised, building foundational knowledge from a large corpus of text using a next-token prediction objective [214]. The post-training phase usually enhances LLMs' reasoning, accuracy, and alignment with particular user needs and ethics [157].

Following the widespread adoption of BERT-based applications in the clinical domain, the rise of decoder-only LLMs is closely tied with two components in the post training: instruction fine-tuning, and scaling law [54]. The **instruction fine-tunig** step fine-tunes a pre-trained LLM on a diverse collection of datasets, tasks, and prompt formats, and thus allows the LLM to follow instructions and generalize to unseen tasks and domains [54]. The

scaling law is an empirical observation, which finds the LLM's performance can improve with the increase in the size of the model, the volume of training data, and the diversity of fine-tuning tasks [54, 145]. This law suggests that larger LLMs can potentially learn complex semantic patterns, and thus perform more sophisticated tasks. Building upon these two major components, there has also been more advanced training techniques [313, 26]. One prominent example is reinforcement learning from human feedback (RLHF) [51, 24]. This approach aligns the LLM' responses better with human values and expectations, and thus makes the LLM a more helpful assistant [51, 24].

In addition to advances in training, significant progress has also been made for LLM inference. For example, the LLMs are found to be in-context learners, which can adapt to new tasks by learning examples provided in the input context, without explicit parameter updates [41]. Similarly, Retrieval-Augmented Generation (RAG) incorporates relevant information retrieved from an external knowledge base into the input context, and is particularly beneficial in domains requiring up-to-date or specialized knowledge [166]. Afterwards, techniques like Chain-of-Thought (CoT) prompting have been introduced, guiding models to generate intermediate reasoning steps, and thus improving performance on complex tasks [322].

Those advances in system development allow LLMs to serve as foundation models, which are capable of adapting to various downstream tasks [36]. An important general-domain foundation model, ChatGPT [2] has demonstrates its strong performance through the landmark paper by Gilson et al. in 2023, by achieving a passing score on the United States Medical Licensing Examination (USMLE) equivalent to that of a third-year medical student [97]. Afterwards, a growing number of medical foundation models have emerged [150], such Med-PaLM [271], MEDITRON [50], and Biomistral [159]. Those foundation models have demonstrated impressive capabilities across multiple benchmark tasks, particularly in areas requiring in-depth medical knowledge and complex diagnostic reasoning.

More recent research has revolutionized the landscape of LLM applications in the healthcare domain, with the introduction of agent-style LLMs [315]. Those LLM agents are integrated with advanced capabilities such as memory retrieval, strategic planning, collaboration with

other models, and tool usage. Unlike traditional models that respond passively, these agents can analyze problems, formulate actionable plans, and execute solutions by interacting with external systems and resources. This paradigm shift toward active problem-solving has paved the way for more diverse and high-quality applications in healthcare, such as more accurate clinical decision support, automated documentation, medical education, and healthcare service optimization [315].

A.3 Evaluating Clinical NLP Systems

As the field of clinical NLP has been evolving quickly, there has been a growing need for reliable ways to evaluate and compare across different systems. Benchmarks help with this. They include standardized datasets and scoring metrics, measuring how well a system works on a specific task [280]. Benchmarks are also important for guiding system development [152]. When there are gaps in a benchmark, those gaps often show where new tools or improvements are needed.

In Section A.1.1 and A.2, we have discussed some widely used NLP systems and their applications. In this section, we will discuss how to create benchmark datasets, commonly used evaluation metrics in clinical NLP benchmarks, and some challenges for the development of reliable benchmarks.

A.3.1 Dataset and Label Curation in Benchmarks

Dataset and Label Curation are foundational to the benchmark development. Ideally, the benchmark dataset should resemble the real-world samples of the target task, so that the systems trained and tested on this dataset can generalize well in real-world scenarios. During this process, stratified sampling is usually employed to ensure that each sub-population (aka class) within dataset maintains the same distribution as in the overall population, which is particularly important in cases of class imbalance [269].

Reliable ground-truth labels are essential to evaluate system performance. In the domain of clinical NLP, these labels can be derived through various methods:

Adaptation from Existing Knowledge Bases: In tasks like medical question answering (e.g., MedQA [136]) or diagnosis prediction [289], labels are often sourced from established knowledge bases such as medical licensing exams or online forums.

Human Expert Annotation: Due to the complexity and nuance inherent in clinical narratives, annotations are typically performed by clinical experts [149]. The quality of these annotations is assessed using inter-annotator agreement metrics, which serve as indicators of dataset reliability and the potential upper bound of model performance [19]. To enhance annotation quality, it's a common practice to have the test set doubly annotated, with a third expert resolving any disagreements. On the other hand, NLP systems can . For example, Lybarger et al. deploy active learning to identify and prioritize uncertain samples for social determinants of health (SDoH) annotation [189].

Pre-annotation deploys a baseline system to generate initial ground-truth labels of good or mediocre quality, which are then reviewed and revised by human annotators. It can expedite the annotation process by allowing human annotators to focus on the most complex and critical tasks, without compromising the annotation quality [175]. However, it is crucial to deploy pre-annotation cautiously. Over-reliance on machine-generated labels can lead to the propagation of errors and may cause annotators to overlook mistakes, especially if the pre-annotations are of low quality.

On the other hand, there are other frequently used automated label generation approaches, such as rule-based systems [144], knowledge distillation [103], and self-training [314]. However, due to the specialized nature of medical language and rigorous data quality required for evaluation benchmarks, these methods are typically used to augment training data rather than to construct gold-standard evaluation datasets. Meanwhile, another approach, crowd sourcing, breaks the annotation task into simpler subtasks and collects annotations from a large group of non-experts [356]. However, due to the sensitive nature of patient records, crowdsourcing is deployed more frequently in the biomedical domain [151].

A.3.2 Evaluation Metrics in the Benchmark Dataset

In this section, we will present two types of evaluation metrics used in benchmark datasets: automatic and human evaluation.

Automatic Evaluation

In this section, we present some commonly used automatic evaluation metrics, as summarized by Schmidtova et al. (2024) [262], who analyzed evaluation methods across 110 papers published at the International Natural Language Generation Conference (INLG) and the Association for Computational Linguistics (ACL) conference in 2023.

Word-overlap metrics measure surface-level similarity between the predicted text and a reference, by comparing the n-gram (e.g., unigrams, bigrams) overlap. Common examples include ROUGE [172], BLEU [233], METEOR [27], and CIDEr [306]. These metrics are fast, easy to implement, and intuitive. However, they often fail to capture deeper semantic meaning and cannot account for synonyms or paraphrases, potentially penalizing predictions that are semantically correct but lexically different from the reference.

Semantic similarity metrics assess how closely the predicted and reference texts align in meaning using embeddings from pre-trained language models. For example, BERTScore computes similarity by aligning words from the reference to the most similar words in the prediction (recall) and vice versa (precision), using contextual embeddings from BERT [360]. Similarly, BARTScore uses BART to measure generation likelihood from one text to another [351]. While these metrics better capture semantic nuances, they may be sensitive to domain shifts, especially when the embedding models are not trained on domain-specific text.

Match-based metrics focus on comparing standardized labels extracted from the predicted and reference texts. Accuracy is commonly used for tasks like multiple-choice question answering (MCQA), sentence classification, and document classification. For tasks with class imbalance, such as named entity recognition (NER), precision, recall, and F1-score are preferred, as they provide a more nuanced view of performance across different entity

types.

Text property-based metrics evaluate specific characteristics of the generated text. For example, readability metrics like the Flesch Reading Ease score assess how easily a text can be understood [83, 33]. Other properties include n-gram diversity [202], which measures lexical richness, and n-gram repetition [169], which captures redundancy in generated output.

Classifier-based metrics use learned models to predict how well the generated text aligns with human judgments. For instance, BLEURT is a regression model built on top of BERT that is fine-tuned with thousands of annotated examples to estimate the quality of a generation [264]. UniEval is a unified framework for evaluation that leverages pretrained classifiers to assess multiple dimensions such as coherence and relevancy [363].

Fact-based metrics aim to measure the factual correctness of the generated content, either directly or indirectly [262]. For example, our proposed metric, MedCon, assesses factuality by comparing the sets of medical concepts extracted from both the reference and predicted texts using the Unified Medical Language System (UMLS) [288]. Indirect approaches include AlignScore, which aligns key facts between text pairs [355], and NLI-based metrics, which determine if the prediction logically entails, contradicts, or is neutral with respect to the reference statement [262, 143].

GPT-Eval metrics use generative LLMs, such as GPT-4 [226], to evaluate specific aspects of predicted text by providing the model with an evaluation rubric [86, 312]. This approach is highly flexible and, in some cases, shows better alignment with human judgments compared to traditional metrics discussed above [180]. However, recent studies have also identified biases in LLM-based evaluations, such as a tendency to favor text generated by other LLMs [232].

In summary, each metric has its own advantages and limitations. More complex metrics can potentially better align with semantic and factual similarities between text pairs, but they may be sensitive to domain shifts and are often less interpretable. As a result, it is now common practice to use a combination of metrics to evaluate different aspects of generated text.

Human Evaluation

Given the limitations of automatic evaluation metrics and the highly specialized and sensitive nature of the clinical domain, human evaluations are commonly employed for more reliable and explainable assessments. Below, we summarize some commonly used human evaluation approaches in clinical NLP.

Comparison-based evaluation compares a model's answers to gold-standard references. For example, medical experts can help assess the factual correctness of open-ended responses from LLMs to clinical exams [344].

Reference-free evaluation does not rely on predefined answers. Instead, experts directly judge the quality of outputs using criteria like accuracy, safety, completeness, and relevance. In the Med-PaLM paper, physicians and lay users rated model-generated answers based on agreement with clinical knowledge, reasoning quality, potential harm, and helpfulness [270].

Ranking-based evaluation involves comparing multiple outputs for the same question. Evaluators rank them by preference or usefulness. For instance, in Med-PaLM 2 paper, doctors rank their preferences among the LLM's answers with those from other physicians, showing doctor preference over the LLM's outputs on some key clinical dimensions [271].

Distinguishability tests ask evaluators to judge whether an answer came from a human or a model. This helps assess how human-like the responses are. In Med-PaLM 2 evaluations, specialists were often unable to tell the difference between the model's answers and those written by physicians, showing the model's ability to generate realistic clinical responses [271].

A.3.3 Challenges in Benchmark Development

In this section, we discuss some challenges about benchmark development for clinical NLP tasks.

Clinical alignment is one of the major challenges in developing useful benchmarks for healthcare. As many benchmarks are developed by NLP experts, they may not align with the clinical significance and practical needs of healthcare providers. For example, a 2023 survey

by Blagec et al. found that existing benchmarks at that time fail to cover many critical needs of clinical experts, such as those related to routine documentation and administration tasks [35]. A more recent survey published in 2025 by Wang el al, summarizes LLMs' promise in addressing these gaps by integrating external knowledge, planning, and tool use to support clinical workflows [315]. However, comprehensive evaluations are necessary to assess their reasoning abilities and ensure alignment with diverse clinical perspectives [315].

Another challenge is the **data quality** concerns, such as missing data and the reliability of ground-truth labels. For example, EHRs often capture only the information relevant to a specific hospital course. Patients may receive care at different hospitals, leading to fragmented records. If one wants to construct a predictive benchmark dataset for a disease, the EHR might not contain all the necessary information [96]. On the other hand, to construct the ground-truth labels, disagreements between human experts are common [289]. These disagreements could stem from human errors or insufficient input contexts for making accurate decisions. Therefore, understanding the reasons behind human disagreements and assessing the quality of the data is important.

As LLMs are trained on vast datasets, **data contamination**, when the training and test sets overlap, is another significant concern [88]. This overlap can lead to inflated performance metrics and misrepresent the model's true capabilities on real-world tasks. Meanwhile, data contamination detection approaches can be used to identify the use of private or sensitive information used in model training. In the Chapter 6, we discuss our study on how to effectively detect data contamination, and our findings detection approaches for data contamination can easily fail. To mitigate data contamination, it is especially important in the clinical domain, to study the data use agreements for patient records. For example, even for the de-identified EHR dataset, MIMIC-III, and its derived NLP systems, information cannot be freely distributed and must be shared under specific agreements [138].

Appendix B

CACER

B.1 Data Set Statistics

Table. B.1 demonstrates the data set statistics and patient demographics.

Type	Subtype	Train	Valid	Test
# Notes	-	400	60	115
# Unique patients	-	306	43	115
Annotation	Double	0	47	115
	Single	400	13	0
Cancer	DLBCL	193	22	56
	Prostate	207	38	61
Gender	Male	314	51	92
	Female	86	9	23

Table B.1: Note statistics and patient demographics.

B.2 Prompts

All task prompts are provided as '*user*' messages in GLMs.

Event Extraction

"You are a medical expert. Extract all drug and medical problem events from the following clinical note. All events constraints span-only arguments and/or valued arguments. Span-only arguments must use the span original from the clinical note. A medical problem event contains required arguments as a trigger span and an assertion value (present, absent, possible,

conditional, hypothetical, not_patient), as well as optional arguments as at most one anatomy span, at most one duration span, at most one frequency span, characteristics spans, change value (no_change, improving, worsening, resolved), severity value (mild, moderate, severe). The drug event contains a required argument as a trigger span."

Relation Extraction, Marker Format

"Extract all relations related to drug and medical problems from this clinical note.

Clinical Note: '..."'

Relation Extraction, QA Format 1 for Drug-Problem relations

"What is the relationship between $<A>$(Drug) and $<B>$(Problem) in the following clinical notes?

Clinical Note: '...'

Options:

(A) A is given as a treatment for B, but the outcome is not mentioned in the sentence.

(B) A is not given or discontinued because of the B that it does not cause.

(C) A is given as a treatment for B, but A does not cure the B, does not improve the B, or makes the B worse.

(D) A is not given as a treatment for B, but it causes B.

(E) A improves, cures, stablize B.

(F) None of the above. "

Relation Extraction, QA Format 2 for Problem-Problem relations

"What is the relationship between $<A>$(Problem) and $<B>$(Problem) in the following clinical notes?

Clinical Note: '...'

Options:

(A) A causes, describes or reveals aspects of B.

(B) B causes, describes or reveals aspects of A.

(C) None of the above."

B.3 Condensed Annotation Guideline

You are a medical expert. Extract all drug and medical problem events from the following clinical note. All events contains a trigger, span-only arguments and/or valued arguments. Trigger and span-only arguments must use the original span from the clinical note, and the shortest span possible. Valued arguments must be chosen from a pre-defined list. For every note, output None, if the span or value does not exist. Output the events by the order of trigger occurrence from clinical note. If there are multiple arguments of the same type, separate them by $< s >$. For example, 'congestive $< s >$ progressive' Multiple Drug and Problem events are separated by [SEP]

This is an example format: <Problem>span <Assertion>value <Anatomy>span $< s >$.. $< s >$ span <Duration>span <Frequency>span <Characteristics>span $< s >$.. $< s >$ span <Change>value <Severity>value [SEP] <Drug>span [SEP] ...

The drug event contains only required argument as a trigger span, which is the shortest span possible indicating a drug or treatment name.

A medical problem event contains required arguments as a trigger span and an assertion value (present, absent, possible, conditional, hypothetical, not_patient). The problem trigger span is be the shortest span possible. The problem trigger is a span that contains observations made by patients or clinicians about the patient's body or mind that are thought to be abnormal or caused by a disease. Generally, the trigger span should not include anatomical information or characteristics of the problem, as this information is captured through separate Anatomy and Characteristics arguments. They are loosely based on the UMLS semantic types of pathologic functions, disease, or syndrome, mental or behavioral dysfunction, cell or molecular dysfunction, congenital abnormality, acquired abnormality, injury or poisoning, anatomic abnormality, neoplastic process, virus/bacterium, sign or symptom, but are not limited by UMLS coverage.

1. present: patient experienced or is experiencing
2. absent: patient has not or is not experiencing

3. possible: patient may be experiencing (denoted by terms like "probably" or "likely")
4. conditional: patient only experiences under specific conditions
5. hypothetical: patient may experience in the future
6. not_patient: not associated with the patient

The problem event has the following optional arguments:

1. Anatomy (span): indicates the body part or region of the body associated with the problem.

2. Duration (span): how long a problem has persisted or when the problem started.

3. Frequency (span): how often a problem occurs (e.g. occasionally, intermittently, chronic, daily, hourly, persistent, etc.)

4. Characteristics (span): problem descriptors, including descriptions of color, consistency, sound, pain, diffuse/localized, etc. A single event (trigger) may have multiple Characteristics spans. For example, a cough could be described through two Characteristics spans, like "dry non-productive" and "painful."

5. Change (value): captures explicit descriptions of changes in the state of the problem. Choose from no_change, improved, worsened, and resolved.

6. Severity (value): Choose from mild, moderate, severe. Severity can be direct description of the patient status such as 'mild fever', or inferred by the treatment plan as (1). No treatment needed - mild, (2). treatment needed - moderate (3). hospitalization needed - severe

B.4 Relation Extraction Context Window

The RE model can be constrained by their context length, ranging from 512 tokens in BERT, 1024 tokens in Flan-T5, to 32k tokens for GPT-4. In this work, we consider a context window

for a certain relation as the minimum continuous sentence span that contains both head and tail events. Our RE approaches are restricted by context windows no longer than 5 sentences and 400 Bio+Clinical BERT tokens, which contain the majority (98.7%) of the relations. Intra-sentence relations constitute 70.7% of the relations from this data set.

GLMs classify each possible relation pair within its corresponding context window. On the other hand, BERT-based models process each unique context window only once. During the training phase, the context window encompasses all events and relations. During inference, to ensure that each relation is predicted only once, predictions are considered valid only when meeting one of the following criteria: (1) the head trigger appears in the beginning sentence and the tail trigger in the ending sentence, (2) the tail trigger appears in the beginning sentence and the head trigger in the ending sentence, or (3) the context window consists of a single sentence.

Appendix C

BIOMISTRAL-NLU

C.1 Unified Prompt Format

Utilizing the unified prompt format outlined in Section 5.1.1, we developed (1) the MNLU-Instruct dataset based on the collection of datasets detailed in Table C.1 and C.2; and (2) the evaluation dataset from BLUE and BLURB utilizing the labels from Table 5.2 and 5.3. In this section, we provide detailed information on dataset creation and examples of the input-output format for each task type.

Task	dataset	# instances	Labels
NER	i2b2 2006DeID	5,608	Location, ID, Date, Hospital, Doctor, Contact, Name, Age
	i2b2 2011	25,689	Person, Treatment, Test, Problem
	i2b2 2012	7,446	Test, Problem, Frequency, Time, Date, Occurrence, Treatment, Duration, Clinical department
	i2b2 2014	52,462	ID, Contact, Age, Name, Location, Profession, Date
	GENIA	15,023	RNA, DNA, Cell type, Protein, Cell line
	linnaeus	11,935	Species
	tmVar	5,351	Cell Line, SNP, Gene, Protein Mutation, Protein Allele, Species DNA Allele, DNA Mutation, Other Mutation, Acid Change,
	DrugProt	17,274	Organism Taxon, Disease Or Phenotypic Feature, Cell Line, Gene Or Gene Product, Sequence Variant, Chemical
	BioRed	13,706	Chemical, Gene
	GNorm	4,006	Family Name, Domain Motif, Gene
	NLM-Gene	5,048	Gene, Gene reference into function (function of a gene), Domain, Steroidogenic acute regulatory protein (a protein coding gene)
	ClinicalIE_Med	105	Route, Duration, Reason, Dosage, Frequency, Medication
	ClinicalIE_Status	105	Neither medications, Discontinued medications, Active medications
	BC4CHEMD	30,682	Chemical
	PubMed PICO	1,961	Species, Comparator, Outcome, Intervention, Strain, Induction
	PICO-Data	36,224	Participants, Intervention, Outcome
EE	i2b2 2009	117,446	Medication (Dosage, Route, Frequency, Duration, Reason, Context)
	i2b2 2018	155,716	Drug, ADE (Strength, Frequency, Reason, Form, Route, Dosage)
	n2c2 2022	36,359	Alcohol, Drug, Tobacco, Employment, Living (time, duration, history, type, amount, frequency)

Table C.1: NER and EE Task labels and number of instances in the MNLU-Instruct dataset. For EE tasks, labels inside () refer to event arguments.

Task	dataset	# instances	Labels
DC	i2b2 2006Smoke	398	Current smoker/Past smoker/Non-smoker/Unknown
	i2b2 2008	17,242	10 obesity commodities (Asthma, Depression, ...)
	n2c2 2018	2,626	Different selection criteria for 13 cohorts (Abdominal, English, ...)
	2024 SemEval2	1,700	Adverse Events, Eligibility, Results, Intervention
	TrialStop	3,747	17 reasons to stop a study (Study staff moved, Another study, ...)
	MTSamples	3,206	48 medical specialties or domains (Bariatrics, Nephrology, ...)
RE	i2b2 2011	25,689	Refers to
	i2b2 2012	7,446	Ends by, Happens during, Happens before and overlap, Begins by, Happens before, Happens simultaneously with, Happens after, Overlaps with,
	EUADR	318	Gene-disease association
	DrugProt	35,624	Antagonist, Agonist, Indirect upregulator, Part of, Agonist activator, Substrate, Activator, Inhibitor, Direct regulator, Agonist inhibitor, Product of, Substrate product of, Indirect downregulator
	BioRed	4,328	Drug interaction, Positive correlation, Cotreatment, Comparison, Bind, Conversion, Association, Negative correlation
NLI	Multi-NLI	785,404	Entailment, Contradiction, Neutral
	SNLI	1,098,734	Entailment, Contradiction, Neutral
	BioNLI	23,704	Entailment, Contradiction, Neutral
STS	SIS-B	11,018	6 similarity scales
NLG	PubMedSum	1,407	Article summarization
	CDSR	436	Article simplication
	AciDemo	204	Dialogue to note summarization

Table C.2: Task labels and number of instances in the MNLU-Instruct datasets, excluding the NER and EE subsets.

C.1.1 Named Entity Recognition (NER)

We conduct NER at the sentence level, because most NER datasets comprise pre-split sentences. For NER datasets where the medical text is an entire document, we use the SpaCy

tokenizer[1] to split the document into sentences.

Below is an example of the NER input-output pair. The example is from the n2c2 2022 dataset [191], a shared task focused on extracting social determinants of health from clinical notes.

- NER Input -

Extract all relevant medical named entities faithfully from the medical text below. Focus on identifying the following entities: Living status, Tobacco, Drug, Employment, Alcohol.

Medical text: Denies any IV drug use or any recreational drug use.

- NER Output -

Living status: None

Tobacco: None

Drug: IV drug use ... recreational drug use

Employment: None

Alcohol: None

C.1.2 Event Extraction (EE)

The EE task is composed of event trigger extraction (ETE), event argument extraction (EAE), and event argument classification (EAC). ETE uses the same prompting formats as NER. In EAE and EAC, we additionally include two adjacent sentences to provide more context information. Below are examples of the EAE and EAC input-output pairs from the n2c2 2022 dataset [191]

- EAE Input -

According to the medical text, what is the Method attribute of the Drug event 'IV drug use' in the medical text below? Extract the attribute faithfully from the medical text.

Medical text: ... Currently admits to five drinks of alcohol per week. Denies any IV drug use or any recreational drug use. Divorced with no children. ...

---

[1]https://spacy.io/api/sentencizer

- EAE Output -

Drug - Method: IV

- EAC Input -

According to the medical text, what is the Status time attribute of the Drug event 'IV drug use' in the medical text below? Choose from the following options.

Medical text: ... Currently admits to five drinks of alcohol per week. Denies any IV drug use or any recreational drug use. Divorced with no children. ...

Options: (A) none (B) past (C) future (D) current

- EAC Output -

Drug - Status time: (A) none

C.1.3 Document Classification (DC)

Our document classification task involves classifying a document or sentence into one or multiple pre-defined categories.

In the i2b2 2006Smoke [299] and i2b2 2008 [297] dataset, where the input document is a lengthy clinical note, we first deploy BioMistral to summarize the document. We use the prompt format, 'Summarize the {*type*} from the following clinical note.', where *type* is the corresponding DC type label, such as smoking status or asthma status.

The MTSamples dataset aims to classify a medical report into one of 48 medical specialties or domains [3]. The large number of possible categories results in lengthy prompts. Instead, in each instance, we include the correct category along with 12 randomly selected negative categories in our prompts for more efficient training.

Below is an example of the DC input-output pair from the TrialStop dataset [247].

- DC Input -

According to the medical text below, which options best describe reason to stop the study? Choose from the following options. Multiple options can be true.

Medical text: 13 of 15 patients recruited. Study patients responded with no safety signals. Recruitment's slow, timely end of study necessary to keep development timelines.

Options: (A) Insufficient enrollment (B) Logistics resources (C) Business administrative (D) Insufficient data (E) Endpoint met (F) Negative (G) Study success (H) Regulatory (I) Interim analysis (J) Ethical reason (K) Invalid reason (L) Study design (M) No context (N) Another study (O) Covid19

- DC output -

(A) Insufficient enrollment (C) Business administrative

C.1.4 Relation Extraction (RE)

The RE task focuses on classifying the relation between any possible entity pairs within the same sentence. We adapt the relation labels from the original publications into descriptive language. We additionally include two adjacent sentences to provide more context information. Below is an example from the i2b2 2011 for coreference resolution on clinical named entities [302]:

- RE Input -

According to the Medical text below, what is the co-reference relationship between the Person entity 'Mr. Andersen' and the Person entity 'who'? Choose from the following options.

Medical text: ... History of Present Illness: Mr. Andersen is a 71-year-old male with worsening anginal symptoms who underwent catheterization that showed severe three-vessel disease. He is presenting for revascularization . ... Options: (A) 'Mr. Andersen' refers to 'who' (B) None of the above.

- RE Output -

(A) 'Mr. Andersen' refers to 'who'

C.1.5 Multi-choice Question-Answering (QA)

The QA task aims to answer a research question regarding the medical text within a predefined answer set. The PubMedQA dataset consists of research questions about PubMed abstracts, with answers categorized as yes, no, or maybe [137]. The BioASQ includes biomedical questions with answers classified as yes or no [294].

Directly applying our sequence classification prompt format for the QA task results in single-word multi-choice answers like *yes* or *no*. Instead, we transform the single-word options into descriptive sentences so that the QA output format is more straight-forward. We utilize one-shot learning with BioMistral to combine the question and each answer into a single statement. The one-shot example is randomly chosen from the PubMedQA train split, and the example output is written by human.

Below is an example of the QA input-output pair from the PubMedQA dataset, with descriptive multi-choice options.

- QA Input -

According to the medical literature below, Is there a connection between sublingual varices and hypertension? Choose from the following options. Only one option can be true.

Medical literature: BACKGROUND: Sublingual varices have earlier been related to ageing, smoking and cardiovascular disease. The aim of this study was to investigate whether sublingual varices are related to presence of ...

Options: (A) The answer is not mentioned in the text (*maybe*). (B) There is a connection between sublingual varices and hypertension (*yes*). (C) There is not a connection between sublingual varices and hypertension (*no*).

- QA Output -

(B) There is a connection between sublingual varices and hypertension (*yes*).

C.1.6 Natural Language Inference (NLI)

The NLI task utilizes a similar multi-choice prompt format to other sequence classification tasks. Below is an example from the BioNLI dataset [28]

- NLI Input -

What is the relationship of the hypothesis with respect to the premise? Choose from the following options.

Premise: The administration of heparin with or without ACTH significantly decreased hepatic cholesterol content in catfish. In serum, heparin alone produced first hypercholesterolemia

which was followed by hypocholesterolemia whereas it potentiated hypercholesterolemic action of ACTH three hours after administration.

Hypothesis: It is concluded that heparin inhibits the cholesterol-lowering action of ACTH in catfish.

Options: (A) neutral (B) entailment (C) contradiction

- NLI Output -

(C) contradiction

C.1.7 Semantic Text Similarity (STS)

We adapt the scoring criteria from the original publications and translate the numerical similarity scores into a descriptive sentences. Below is an example from the STS-B dataset [309]

- STS Input -

How similar are the two sentences below? Choose from the following options.

Sentence 1: A plane is taking off.

Sentence 2: An air plane is taking off.

Options: (A) The two sentences are completely dissimilar. (B) The two sentences are not equivalent, but are on the same topic. (C) The two sentences are not equivalent, but share some details (D) The two sentences are roughly equivalent, but some important information differs / missing. (E) The two sentences are mostly equivalent, but some unimportant details differ. (F) The two sentences are completely or mostly equivalent, as they mean the same thing.

- STS Output -

(F) The two sentences are completely or mostly equivalent, as they mean the same thing.

C.1.8 Natural Language Generation (NLG)

We focus on the abstractive summarization task from NLG. Although abstractive summarization is different from our focus on NLU, it also requires in-depth document understanding,

and thus we believe it can help improve NLU performance. We include three NLG datasets in the MNLU-Instruct: PubMedSum [57], CDSR [105], and AciDemo [288]. PubMedSum has the input as the complete PubMed articles and the output as their abstracts. CDSR is a text simplification task that translates domain-specialized summaries into lay-user summaries. AciDemo is a task that summarizes doctor-patient dialogues into clinical note sections. Because the PubMedSum and AciDemo documents can be very lengthy, we only include instances with less than 800 words. Additionally, we restrict the output in PubMedSum to be at most half of its corresponding input word count to ensure that the PubMedSum splits contain high-quality summaries.

Below is an example from the AciDemo dataset.

- NLG Input -

Summarize the relevant medical information from a dialogue between a doctor and a patient. The summary should be the objective exam section from the clinical note. Output None if no relevant information is found.

Dialogue:[doctor] hi alan , how are you ?

[patient] hi , good to see you .

[doctor] good to see you as well . are you ready to get started ?

[patient] ...

- NLG Output -

PHYSICAL EXAMINATION

Neck: Supple.

No jugular venous distension.

Respiratory: Slight expiratory wheezing bilaterally.

Cardiovascular: Regular rate and rhythm. No murmurs.

Musculoskeletal: Trace edema in the bilateral lower extremities.

C.2 Baseline System with ICL for NER Tasks

Generalized LLMs do not automatically extract named entities in a unified format. To avoid confounding factors from different output formats and simplify NER evaluation, we utilize the same NER input-output format as described in Appendix C.1.1. Additionally, we include a descriptive paragraph at the beginning of the input prompt to specify the output format: "Your answer should use the following format, with one entity type per line. The span refers to the original text span from the Medical text. Output None if there is no such span. Use '...' to separate multiple spans."

We also include two in-context examples to ensure the baseline system adheres to the desired output format. For each inference query, the 2-shot examples are randomly selected from the training split of each dataset. We ensure the outputs from the 2-shot examples are different from each other, to prevent bias towards a specific extraction response.

Appendix D

DATA CONTAMINATION

D.1 Risks and Mitigation Approaches for Data Contamination

During our paper collection, we identify the relevant research on the risks and mitigation strategies for data contamination, which does not involve proposing or evaluating existing detection approaches for data contamination. While excluding those papers from the main text of this chapter, we provide their citations in Table D.1 in this Appendix.

Citation	Content
[364]	Impact of direct data contamination on test performance
[62]	Impact of indirect data contamination on test performance
[125]	Strategies to prevent contamination in benchmark datasets
[368]	Strategies to mitigate contamination in benchmark datasets
[107]	Mitigating data contamination in benchmarks through retrospectively creating held-out datasets.
[158]	Proposing an evaluation pipeline in Systematic Literature Review to mitigate data contamination.
[203]	Evaluating the novelty of LM-generated text
[197]	Studying unlearning methods to make LMs forget specific training data
[38]	Studying contributing factors behind data poisoning, with corrupted or malicious training data.
[210]	Differentiates human vs. machine-generated text using probability curvature

Table D.1: Seleted relevant work to risks and mitigation approaches for data contamination.

D.2 Table of Notations

We present the notations used in Chapter 6 in Table D.2.

Notation	Definition
x	A language instance, as a sequence of tokens.
(x_p, x_s)	A prefix-suffix instance pair, which is a common data format in NLG tasks.
(x_c, x_k)	A context-key instance pair from slot-filling tasks, as defined in Requirement R5.
M	A language model.
$M.(x)$	M's output respect to an input x, given a decoding setup $\cdot$. If $\cdot$ is not specified, we consider it as a fixed, but unknown decoding state.
D	A dataset, as a set of language instances.
D_M	M's training set.
$b(x, x')$	Binary indicator function for instance-level contamination, which takes two instances x and x' as inputs, and returns *False* (0) or *True* (1), based on the instance similarity.
$S(x, x')$	A function accessing the similarity between two instances, x and x' and outputs a real value.
$f(M, x)$	Gold standard for instance-level contamination, as defined in Equation 6.1.
$P_M(x)$	The probability of the instance x given an LM M.
Min p% Token	The average probabilities of top p% least likely tokens in an instance x, based on a given LM M.
τ	The contamination threshold for functions.
Var($\{M.(x_p)\}$)	The measure of variations of outputs produced by M under diverse, different sampling strategies. given x_p
PPL_k	The perplexity from an instance's first k tokens.
Entropy k	The entropy of the top k most likely tokens for the next position, defined by Equation D.1.
Eval(M, D)	The evaluation result of an LM M on a dataset D.

Table D.2: Table of notations.

D.3 Case Study for Instance Similarity

To assess the applicability of instance similarity-based detection approaches (see Section 6.2.1), we analyzed how frequently their requirements, R1 and R2, are met. We reviewed the top 10 models from the Vellum LLM leaderboard[1]. As demonstrated in Table D.3, none of the models fulfilled R1, the most basic requirement. However, some models disclose their cut-off date for collecting training data [5, 223].

Model	**Citation**	**Meet Require.**	
		R1	R2
Claude 3.5 Sonnet	[15]	No	-
Claude 3 Opus	[17]	No	-
Gemini 1.5 Pro	[237]	No	-
GPT-4	[5]	No	-
Llama 3 Instruct - 70B	[20]	No	-
Claude 3 Haiku	[16]	No	-
GPT-3.5	[223]	No	-
Mixtral 8x7B	[132]	No	-
GPT-4o	[224]	No	-
GPT-4o mini	[225]	No	-

Table D.3: None of the top 10 LMs, in the LLM Leaderboard by Vellum meet the requirements of disclosing pre-training corpora (R1).

D.4 Entropy Calculation

In this section, we explain the procedure for verifying Assumption A6. In the context of casual language modeling, we consider an LM M with a given prefix sequence x_p. Assumption A6 assumes that given x_p, if M has seen an instance with the same prefix, it will generate similar responses, regardless of the sampling strategy used. Since the verification of this assumption can be influenced by various sampling strategies, we quantify the variance in the model's output by measuring the entropy of the token probabilities, which indicates the

[1] `https://www.vellum.ai/llm-leaderboard#model-comparison`. Accessed on Oct 6, 2024.

model's certainty about the next token generation.

To do this, we first compute the probability distribution of the next token over the model's vocabulary. Given that LLMs may contain vocabularies with over 50,000 tokens [34], most tokens have a very low likelihood of being sampled. Therefore, we focus on the Entropy among the top k most likely tokens (**Entropy** k).

At every token position x_p, we calculate the entropy based on the probabilities of the top k tokens, using the following formula:

$$\begin{aligned} &\text{Entropy}_k(M, x_p) \\ &\quad = -\sum_{i=1}^{k} P_i(M(x_p)) \log P_i(M(x_p)) \end{aligned} \tag{D.1}$$

Given an instance x with N tokens, the Entropy k for x is the average $\text{Entropy}_k(M, x_p)$ across all tokens x_p in x:

$$\text{Entropy}_k(M, x) = (\sum_{p=1}^{N} \text{Entropy}_k(M, x_p))/N \tag{D.2}$$

D.5 More Case Study Results

D.5.1 Within-Domain Detection with Different LMs

In this section, we present the detailed detection AUCs for all models: (1) different sizes of Pythia Models: Pythia-70m (Table D.4), Pythia-160m (Table D.5), Pythia-410m (Table D.6), Pythia-1.4b (Table D.7), Pythia-2.8b (Table D.8), Pythia-6.9b (Table D.9), Pythia-12b (Table D.10); (2) OLMo-2-7B (Table D.11); and (3) BioMistral-NLU-7B (Table D.12).

Similar to the results in Table 6.3, we observed close-to-random performance in the detection AUCs for all Pythia models and dataset domains.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.4	49.3	50.5	51.3	51.7	47.4	48.9	52.2	49.9
	PPL_100	50.3	50.2	51.5	50.8	50.5	46.2	48.1	52.0	50.8
	PPL_200	50.3	50.1	53.5	50.5	49.9	44.3	51.1	50.8	51.5
	Min 5% token	51.7	49.9	49.9	50.6	48.9	45.0	50.4	49.9	51.1
	Min 15% token	51.5	49.5	51.0	50.5	49.5	45.4	50.3	49.6	51.6
	Min 25% token	51.4	50.1	51.0	50.9	49.9	45.5	49.5	49.8	51.3
A4	Mem 5	47.6	49.2	49.5	51.5	50.8	48.8	49.5	50.0	51.2
	Mem 15	49.6	49.5	48.7	50.1	50.4	49.0	49.3	50.4	51.2
	Mem 25	49.6	49.9	48.3	50.6	50.3	48.2	49.6	49.6	51.0
A6	Entropy 5	50.8	49.7	48.2	52.1	49.9	47.4	49.3	50.4	50.2
	Entropy 15	50.6	49.8	48.9	52.4	49.4	47.5	49.3	50.5	50.3
	Entropy 25	50.6	49.9	49.2	52.4	49.0	47.7	49.3	50.2	50.2
Average AUC		50.1	49.7	49.8	51.2	50.1	47.3	49.4	50.3	50.9
PPL_200	Seen	11.1±12.7	13.1±8.9	29.5±21.4	28.9±13.5	46.0±23.5	36.7±21.1	42.9±16.7	45.3±41.3	50.2±33.7
	Unseen	11.5±20.9	14.2±25.3	31.7±22.4	29.0±13.1	45.4±22.5	33.7±18.2	43.7±18.8	44.3±28.7	51.5±35.6

Table D.4: Average contamination detection AUC for the **pythia-70m** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.7	49.0	50.3	51.2	51.9	47.6	48.6	52.3	50.0
	PPL_100	50.3	49.4	51.2	51.0	50.4	46.1	47.6	51.7	50.7
	PPL_200	50.1	49.4	53.0	50.7	49.6	44.2	50.5	50.7	51.7
	Min 5% token	51.9	49.9	51.0	50.5	48.6	45.0	48.6	49.4	51.3
	Min 15% token	51.4	49.8	50.9	50.8	49.5	45.1	50.1	49.3	51.4
	Min 25% token	51.3	49.7	51.3	51.3	49.6	45.6	49.4	49.2	51.4
A4	Mem 5	48.3	49.1	50.4	52.8	51.2	52.6	50.1	49.9	51.1
	Mem 15	48.2	49.2	50.1	51.4	51.5	48.0	49.1	49.3	50.0
	Mem 25	48.4	49.0	49.9	50.5	51.4	46.2	49.2	49.4	50.2
A6	Entropy 5	51.3	49.1	48.9	52.1	49.8	47.0	48.8	50.1	50.6
	Entropy 15	51.1	49.3	49.4	52.2	49.7	47.5	48.9	50.1	50.6
	Entropy 25	51.0	49.4	49.6	52.1	49.4	47.7	48.9	50.0	50.6
Average AUC		50.1	49.3	50.4	51.5	50.3	47.7	49.2	49.9	50.8
PPL_200	Seen	7.4±8.5	8.2±5.6	18.8±13.5	18.6±8.6	29.9±31.5	45.8±598.9	29.0±11.0	30.9±28.2	33.3±25.3
	Unseen	7.6±13.5	8.8±16.7	20.3±14.8	18.7±8.3	28.5±12.8	24.8±11.4	29.6±13.1	30.4±23.0	34.4±25.9

Table D.5: Average contamination detection AUC for the **pythia-160m** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.5	49.5	50.1	50.7	51.0	47.4	48.8	51.5	50.0
	PPL_100	50.5	49.2	51.3	50.9	49.7	45.5	47.6	51.0	50.3
	PPL_200	50.3	49.3	53.0	50.6	48.8	44.6	50.9	50.2	52.0
	Min 5% token	52.2	49.8	50.3	50.7	49.0	45.5	49.3	48.9	51.4
	Min 15% token	51.6	49.6	50.9	51.0	49.3	46.0	50.2	49.0	51.2
	Min 25% token	51.4	49.4	51.0	51.3	48.9	46.4	50.2	48.9	51.0
A4	Mem 5	47.5	49.2	51.3	53.0	50.5	50.3	49.4	49.7	50.3
	Mem 15	47.6	49.2	51.3	52.6	51.8	50.3	49.7	50.1	50.4
	Mem 25	48.2	49.2	51.1	51.2	51.9	52.6	50.5	49.2	51.2
A6	Entropy 5	51.2	49.1	49.4	52.2	50.5	47.4	49.0	49.5	50.2
	Entropy 15	51.0	49.3	49.9	52.0	49.7	48.0	48.9	49.5	50.2
	Entropy 25	50.9	49.4	50.1	51.9	49.4	48.2	48.9	49.2	50.3
Average AUC		50.0	49.4	50.8	51.9	50.1	48.4	49.5	49.7	50.7
PPL_200	Seen	4.7±5.0	5.5±3.4	11.8±8.2	12.7±6.0	18.8±9.7	20.0±9.1	19.8±7.5	20.6±22.2	22.3±17.0
	Unseen	4.8±7.2	5.9±11.4	12.9±9.6	12.7±5.7	18.2±8.3	18.5±7.9	20.4±9.3	20.0±14.8	23.0±17.8

Table D.6: Average contamination detection AUC for the **pythia-410m** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.5	49.0	49.8	50.7	50.7	48.1	48.5	51.0	49.8
	PPL_100	50.7	49.5	50.9	51.0	49.2	45.6	47.4	50.9	50.2
	PPL_200	50.8	49.8	51.9	50.9	48.7	44.4	50.9	49.9	52.0
	Min 5% token	51.8	50.5	49.7	51.0	49.2	46.2	48.9	48.4	51.1
	Min 15% token	51.5	49.7	50.3	51.3	49.6	47.0	50.1	48.7	51.1
	Min 25% token	51.4	49.4	50.5	51.5	49.1	47.8	50.0	48.6	51.2
A4	Mem 5	48.8	49.1	50.9	53.0	51.0	51.5	49.6	49.1	50.5
	Mem 15	48.0	49.5	50.8	51.4	51.2	51.7	49.8	48.9	50.4
	Mem 25	48.7	49.5	51.1	51.0	51.5	50.9	49.7	48.7	51.0
A6	Entropy 5	51.2	49.0	49.9	52.0	49.9	46.6	48.7	49.3	50.2
	Entropy 15	51.0	49.1	50.1	51.9	49.5	48.0	48.7	48.9	50.2
	Entropy 25	51.0	49.2	50.1	51.8	49.3	48.5	48.8	48.7	50.3
Average AUC		50.2	49.4	50.5	51.7	50.1	48.6	49.2	49.1	50.6
PPL_200	Seen	3.6±3.8	4.4±2.4	8.5±5.9	9.8±4.6	14.2±7.4	16.0±6.9	15.3±5.7	16.3±19.2	17.1±12.5
	Unseen	3.6±4.9	4.7±8.6	9.3±7.3	9.9±4.4	13.7±6.4	14.8±6.1	15.7±7.2	15.7±12.3	17.8±13.9

Table D.7: Average contamination detection AUC for the **pythia-1.4b** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.4	48.6	49.4	50.7	50.5	47.8	48.0	51.1	49.9
	PPL_100	50.5	49.2	50.6	50.9	49.0	45.9	47.5	51.0	50.4
	PPL_200	50.6	49.6	51.7	50.8	48.6	45.3	50.8	49.9	52.2
	Min 5% token	51.6	49.8	49.9	51.2	49.5	47.2	48.2	48.6	51.6
	Min 15% token	51.5	49.6	50.3	51.4	49.5	48.1	49.4	48.7	51.3
	Min 25% token	51.2	49.4	50.1	51.4	48.9	48.8	49.7	48.6	51.0
A4	Mem 5	48.7	49.1	50.7	52.8	50.9	51.5	49.5	50.2	50.9
	Mem 15	48.2	48.9	50.4	51.0	51.3	50.4	48.9	49.7	50.2
	Mem 25	48.9	48.5	50.5	50.8	51.4	50.0	48.7	49.3	50.2
A6	Entropy 5	50.9	49.1	49.5	51.9	50.0	47.5	48.9	49.0	50.6
	Entropy 15	50.8	49.2	49.8	51.9	49.5	48.7	49.0	48.9	50.5
	Entropy 25	50.8	49.2	50.0	51.8	49.3	49.1	49.0	48.8	50.6
Average AUC		50.1	49.2	50.2	51.6	50.0	48.8	49.0	49.4	50.8
PPL_200	Seen	3.2±3.3	3.9±2.1	7.2±5.1	8.7±4.1	12.5±6.5	14.0±6.2	13.3±4.9	14.4±17.3	15.0±10.2
	Unseen	3.2±4.4	4.2±7.8	7.9±6.3	8.7±3.9	12.1±5.8	13.1±5.5	13.6±6.1	13.9±11.3	15.7±12.2

Table D.8: Average contamination detection AUC for the **pythia-2.8b** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.7	48.9	49.5	50.7	50.3	47.7	48.5	50.7	49.8
	PPL_100	50.7	49.4	50.3	51.1	48.8	46.2	47.4	50.4	50.4
	PPL_200	50.8	49.4	51.1	50.9	48.4	46.7	50.7	49.6	52.1
	Min 5% token	51.7	49.8	49.2	51.3	49.7	48.2	47.8	48.5	50.9
	Min 15% token	51.6	49.7	49.8	51.3	49.1	49.5	49.7	48.6	51.1
	Min 25% token	51.3	49.3	50.0	51.4	48.8	50.2	49.8	48.5	51.1
A4	Mem 5	49.2	48.7	50.6	52.7	50.5	50.0	49.9	49.1	50.8
	Mem 15	48.9	48.9	51.0	51.9	51.8	50.6	49.2	48.6	50.7
	Mem 25	49.6	48.8	51.1	51.6	51.1	50.3	49.5	48.2	51.3
A6	Entropy 5	51.0	49.1	49.5	52.0	49.7	48.4	49.1	49.3	50.9
	Entropy 15	50.8	49.1	49.7	52.0	49.3	49.6	49.0	49.0	50.9
	Entropy 25	50.8	49.2	49.8	51.9	49.1	50.0	49.0	48.9	51.0
Average AUC		50.3	49.2	50.2	51.8	49.9	49.4	49.3	49.0	50.9
PPL_200	Seen	2.8±3.0	3.6±1.9	6.1±4.4	7.9±3.7	11.2±5.8	12.2±5.9	12.2±4.5	13.2±16.0	13.7±9.1
	Unseen	2.9±4.1	3.8±6.2	6.7±5.6	8.0±3.6	10.9±5.3	11.5±5.1	12.4±5.4	12.7±10.6	14.3±11.0

Table D.9: Average contamination detection AUC for the **pythia-6.9b** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		Github	FreeLaw	Enron-Emails	ArXiv	OpenWeb-Text2	Open-Subtitles	Hacker-News	Youtube-Subtitles	Pile-CC
A1	PPL_50	49.9	48.7	49.5	51.0	50.4	48.5	48.3	50.7	49.8
	PPL_100	50.9	49.6	50.2	50.9	48.9	47.4	47.3	50.2	50.3
	PPL_200	50.9	49.5	51.0	51.0	48.5	48.4	50.5	49.4	51.8
	Min 5% token	51.9	50.2	49.2	51.4	49.5	49.2	47.8	48.7	50.9
	Min 15% token	51.7	49.4	49.8	51.3	49.0	50.2	49.4	48.7	51.0
	Min 25% token	51.4	49.1	49.9	51.4	48.6	50.8	49.3	48.8	51.0
A4	Mem 5	49.0	49.0	50.9	53.9	51.0	52.6	49.6	48.5	50.9
	Mem 15	48.5	49.5	51.0	52.5	51.1	49.0	49.1	48.3	50.0
	Mem 25	49.5	49.2	50.6	52.7	50.7	48.6	49.2	48.2	51.4
A6	Entropy 5	51.0	49.0	49.5	52.0	50.0	49.2	48.8	49.3	51.0
	Entropy 15	50.9	49.1	49.7	51.9	49.6	50.6	48.8	49.1	50.9
	Entropy 25	50.9	49.2	49.8	51.8	49.4	51.0	48.8	49.0	50.9
Average AUC		50.3	49.2	50.2	52.1	49.9	49.9	48.9	48.9	50.8
PPL_200	Seen	2.6±2.8	3.4±1.7	5.4±4.0	7.5±3.5	10.4±5.4	11.0±5.7	11.2±4.0	12.3±15.0	12.9±8.4
	Unseen	2.7±3.7	3.6±5.6	5.9±5.0	7.5±3.4	10.1±5.0	10.6±5.0	11.5±5.0	11.8±10.0	13.4±10.1

Table D.10: Average contamination detection AUC for the **pythia-12b** model, under different domains within the Pile dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		cpp	python	Github-Lean	julia	tex	Github-Isabelle	fortran	Github-Coq	r
A1	PPL_50	51.1	51.7	49.9	50.2	49.4	50.9	50.3	49.1	49.5
	PPL_100	51.4	51.6	52.0	51.3	50.4	51.7	50.6	49.1	51.4
	PPL_200	51.8	50.8	51.2	51.6	50.0	50.8	51.1	49.0	53.0
	Min 5% token	50.1	51.1	49.0	52.0	49.4	51.2	48.1	49.4	50.9
	Min 15% token	50.2	50.4	50.7	51.2	48.6	50.4	48.1	49.0	51.5
	Min 25% token	50.3	49.8	51.3	51.0	48.7	49.9	48.2	48.9	52.4
A4	Mem 5	51.3	48.2	51.8	49.0	48.2	55.4	50.8	50.5	51.6
	Mem 15	52.0	49.6	50.7	48.3	48.5	54.6	50.6	50.4	52.0
	Mem 25	51.6	50.5	50.1	49.1	49.4	54.3	50.4	50.8	50.9
A6	Entropy 5	50.2	49.4	51.7	51.3	48.8	49.0	48.8	48.9	54.1
	Entropy 15	50.2	49.4	51.7	51.3	48.8	48.9	48.8	49.0	53.5
	Entropy 25	50.3	49.5	51.8	51.2	48.8	49.0	48.8	49.0	53.5
Average AUC		50.9	49.8	50.9	50.4	49.0	51.6	49.6	49.7	52.2
PPL_200	Seen	4.3±2.6	6.7±3.9	6.9±3.4	8.3±5.0	8.3±5.1	8.7±5.4	9.5±6.8	10.4±8.3	10.5±7.0
	Unseen	4.6±3.0	6.8±4.1	6.9±3.0	8.6±5.6	8.7±6.3	9.2±6.5	9.9±7.6	9.9±7.2	10.5±5.6

Table D.11: Average contamination detection AUC for the **OLMo-2-1124-7B** model, under different domains within the Algebraic Stack dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

Assumptions & Metric		RE-2012temp	STS-B	DC-MTSample	RE-2011coref	events-BioRed	events-NLMGene	events-2012temp	events-2006deid	events-BioASQ
A1	PPL_50	**62.5**	**83.3**	51.9	**60.5**	50.4	49.4	50.0	51.4	**67.6**
	PPL_100	**95.3**	**93.3**	60.0	**70.0**	50.0	48.8	50.4	59.3	52.6
	PPL_200	**99.4**	**96.8**	58.9	**70.5**	**89.4**	**87.4**	**76.2**	**79.0**	**66.1**
	Min 5% token	**92.9**	**93.4**	47.4	**72.6**	**61.4**	**76.1**	57.5	**78.0**	**74.9**
	Min 15% token	**93.3**	**93.4**	51.7	**70.1**	**88.3**	**85.3**	**71.2**	**82.3**	**69.4**
	Min 25% token	**93.4**	**92.1**	53.3	**68.5**	**94.1**	**80.5**	**74.8**	**76.5**	**64.7**
A4	Mem 5	41.4	48.0	47.9	46.6	49.8	52.0	51.0	52.5	49.2
	Mem 15	45.2	53.3	49.1	46.4	48.9	52.3	51.9	52.7	52.8
	Mem 25	48.9	55.9	50.3	48.4	46.6	52.1	52.6	53.1	52.5
A6	Entropy 5	**93.2**	**80.5**	55.4	**63.9**	**94.5**	**65.3**	**74.3**	**62.6**	43.1
	Entropy 15	**93.1**	**82.0**	54.8	**64.5**	**94.5**	**66.6**	**74.2**	**63.7**	46.0
	Entropy 25	**93.1**	**82.6**	54.1	**64.7**	**94.5**	**67.1**	**74.4**	**64.2**	47.6
Average AUC		**74.1**	**73.1**	52.0	59.0	**69.8**	**63.5**	**62.4**	**62.8**	55.1
PPL_200	Seen	1.4±0.1	1.5±0.1	1.7±0.2	2.1±0.8	2.8±0.3	3.1±0.4	3.2±0.4	3.3±0.6	8.1±2.3
	Unseen	3.0±1.6	2.0±0.3	1.8±0.2	3.6±2.4	3.8±0.9	4.3±1.0	4.1±1.1	4.6±1.6	9.4±2.7

Table D.12: Average contamination detection AUC for the **BioMistral** model, under different domains within the Medical-NLU dataset. 'PPL_200' represents the average perplexity ± STD, from the first 200 tokens within every instance. The color green represents AUCs higher than 60.

D.5.2 Metric Distribution in Histogram

In this section, we present the distributions of different metrics both within domain and across domains.

We compare the MIA performance between the GitHub and Pile-CC domains, from the Pythia-6.9b model. As shown in Figure D.1 & D.2, when the seen and unseen instances are from the same domain, their PPL_200 distributions are very similar. However, as shown in Figure D.3 & D.4, when the seen and unseen instances are from different domains, their PPL_200 distributions are very different. This indicates that the PPL_200 relates more to domain shifts, instead of the contamination status of individual instances.

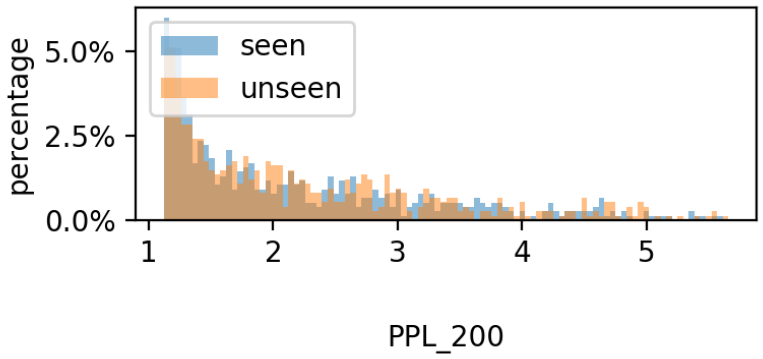

Figure D.1: The density plot of **PPL_200** from the Pythia-6.9b model, when both seen and unseen instances are from the **Github** domain.

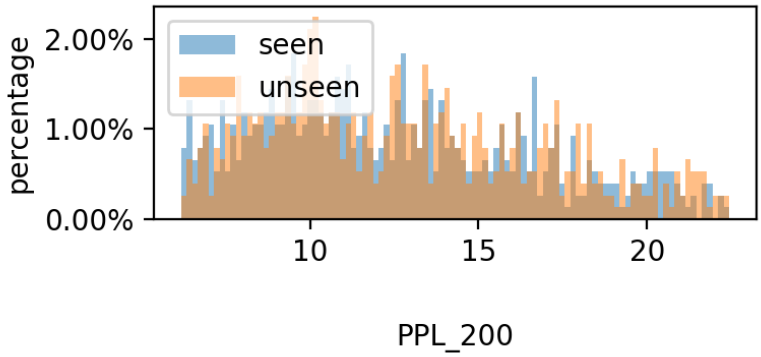

Figure D.2: The density plot of **PPL_200** from the Pythia-6.9b model, when both seen and unseen instances are from the **Pile-CC** domain.

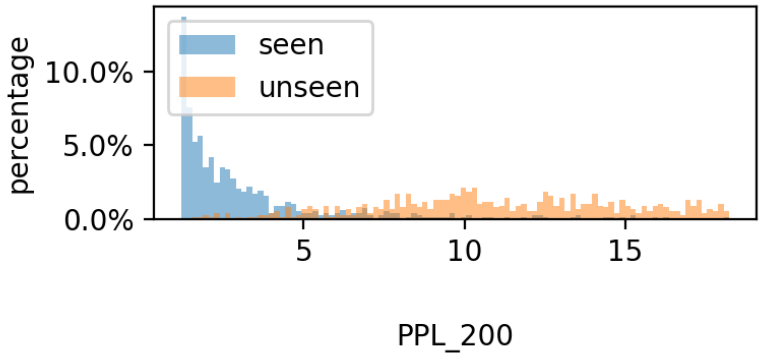

Figure D.3: The density plot of **PPL_200** from the Pythia-9.6b model, when the seen instances are from the **Github** domain, and the unseen instances are from the **Pile-CC** domain.

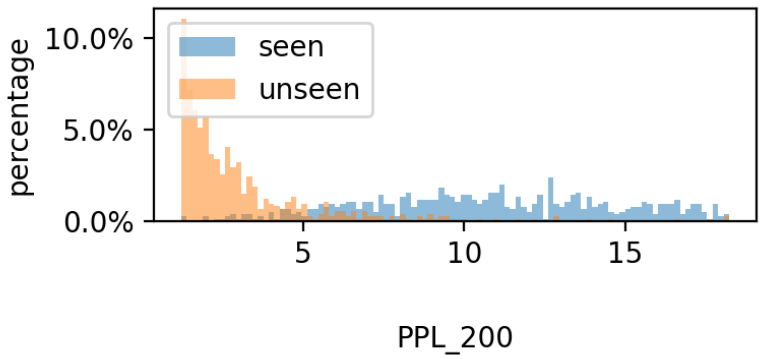

Figure D.4: The density plot of **PPL_200** from the Pythia-9.6b model, when the seen instances are from the **Pile-CC** domain, and the unseen instances are from the **Github** domain.

We observe a similar trend for other metrics: Min 25% Prob (Figure D.5, D.6, D.7, D.8), Mem 25 (Figure D.9, D.10, D.11, D.12), Entropy 25 (Figure D.13, D.14, D.15, D.16).

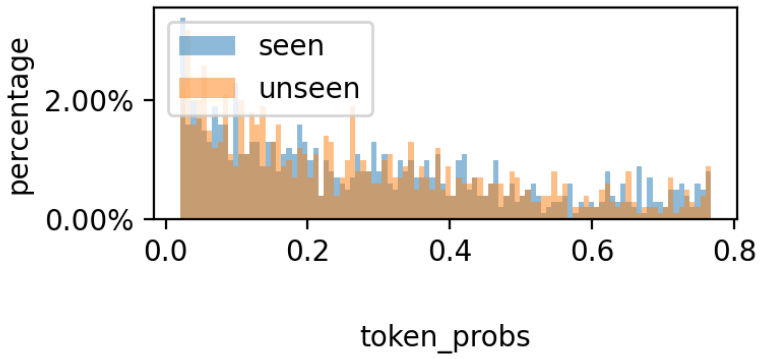

Figure D.5: The density plot of **Min 25% Prob** when both seen and unseen instances are from the **Github** domain.

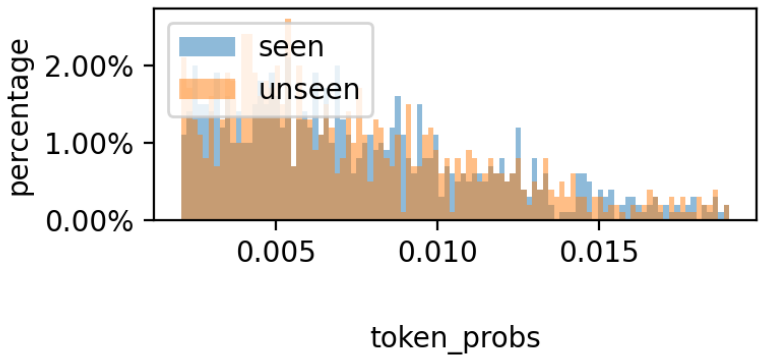

Figure D.6: The density plot of **Min 25% Prob** when both seen and unseen instances are from the **Pile-CC** domain.

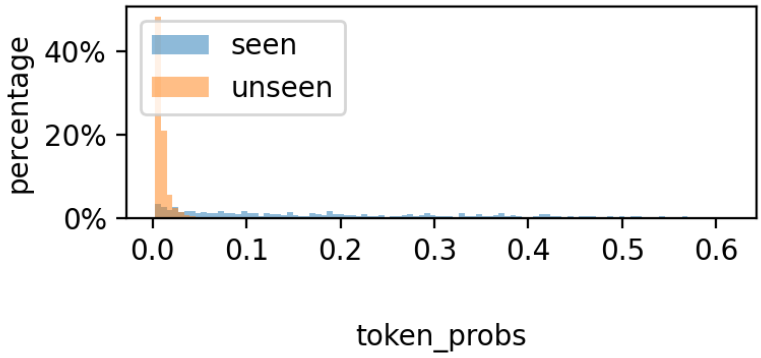

Figure D.7: The density plot of **Min 25% Prob** from the Pythia-9.6b model, when the seen instances are from the **Github** domain, and the unseen instances are from the **Pile-CC** domain.

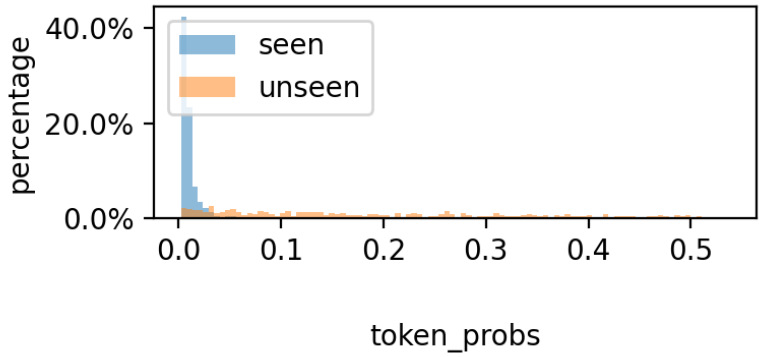

Figure D.8: The density plot of **Min 25% Prob** from the Pythia-9.6b model, when the seen instances are from the **Pile-CC** domain, and the unseen instances are from the **Github** domain.

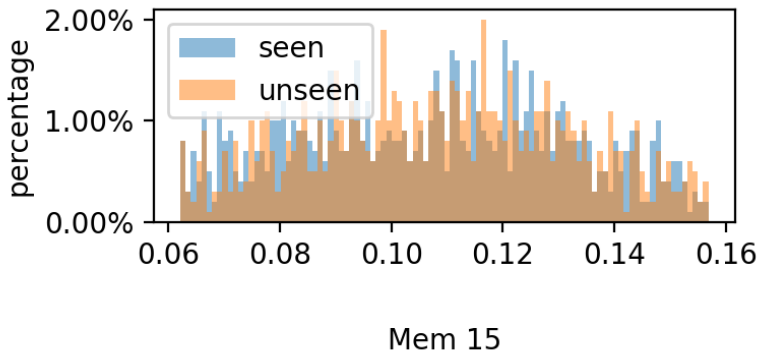

Figure D.9: The density plot of **Mem 25** from the Pythia-6.9b model, when both seen and unseen instances are from the **Github** domain.

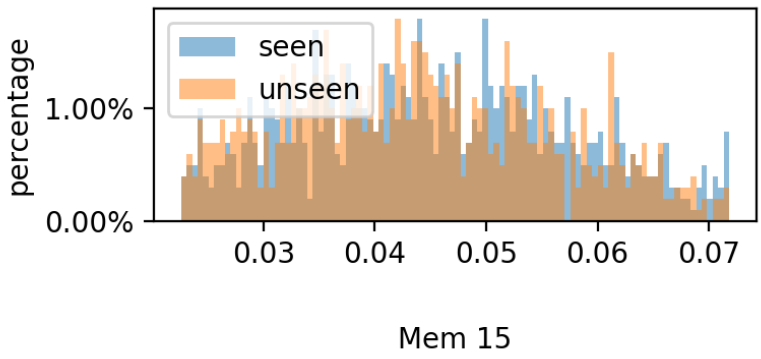

Figure D.10: The density plot of **Mem 25** from the Pythia-6.6.9b model, when both seen and unseen instances are from the **Pile-CC** domain.

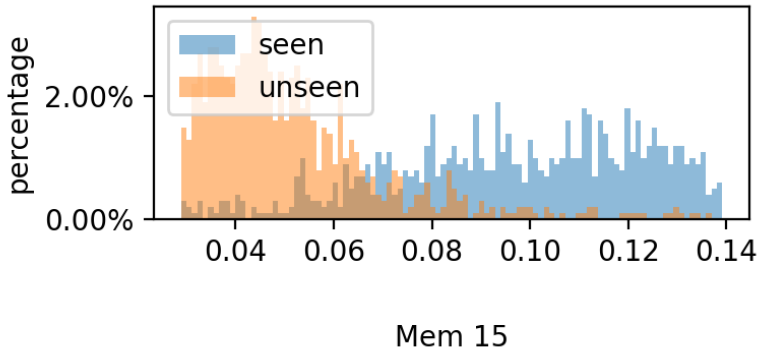

Figure D.11: The density plot of **Mem 25** from the Pythia-9.6b model, when the seen instances are from the **Github** domain, and the unseen instances are from the **Pile-CC** domain.

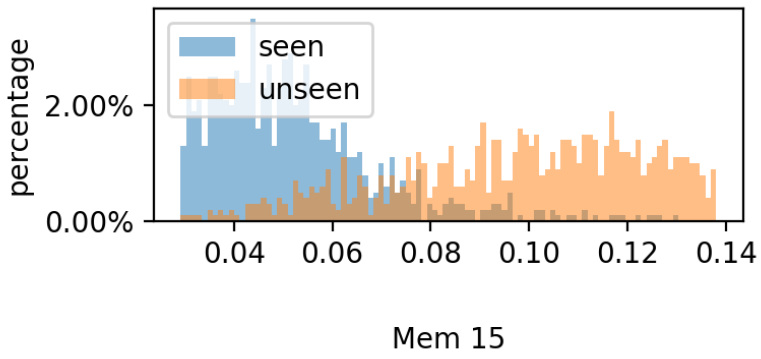

Figure D.12: The density plot of **Mem 25** from the Pythia-9.6b model, when the seen instances are from the **Pile-CC** domain, and the unseen instances are from the **Github** domain.

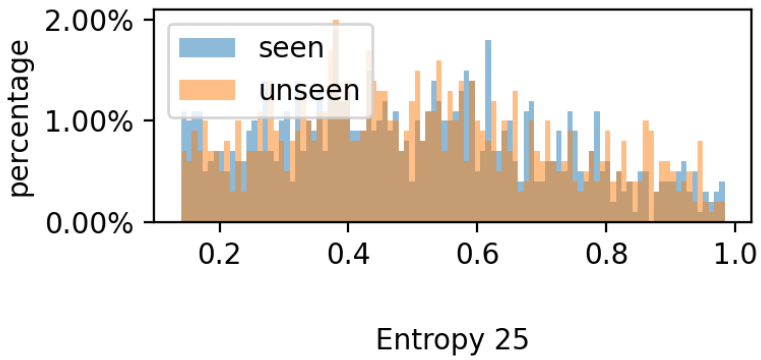

Figure D.13: The density plot of **Entropy 25** from the Pythia-6.9b model, when both seen and unseen instances are from the **Github** domain.

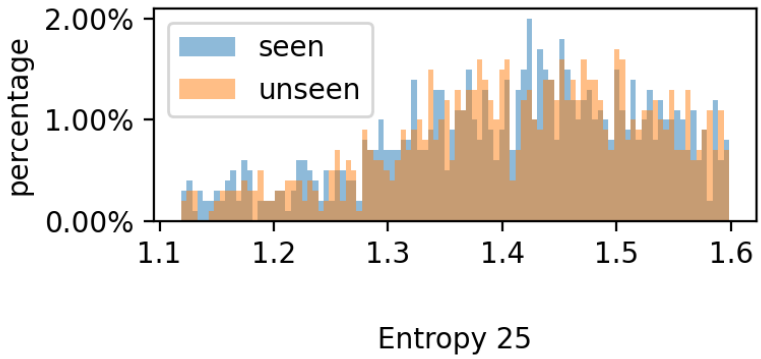

Figure D.14: The density plot of **Entropy 25** from the Pythia-6.9b model, when both seen and unseen instances are from the **Pile-CC** domain.

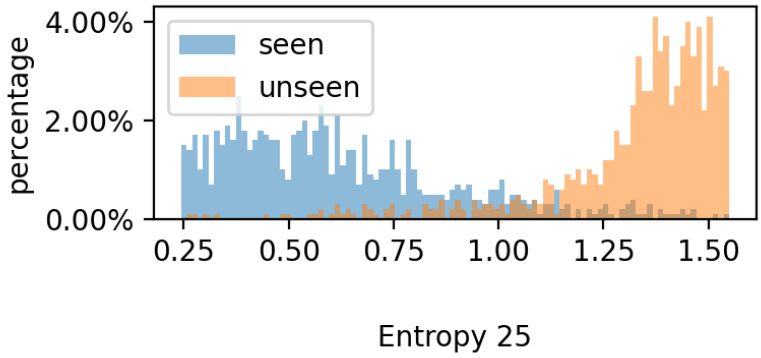

Figure D.15: The density plot of **Entropy 25** from the Pythia-9.6b model, when the seen instances are from the **Github** domain, and the unseen instances are from the **Pile-CC** domain.

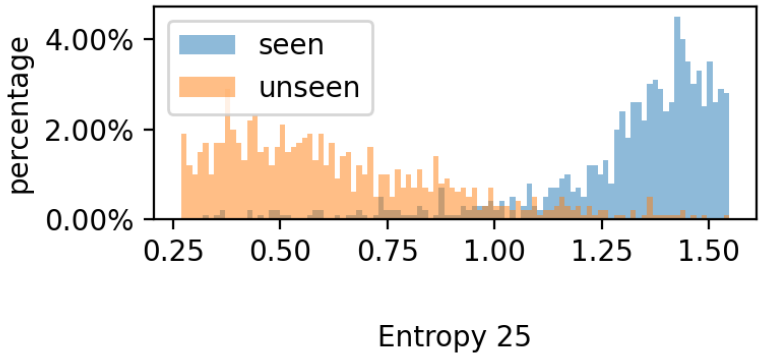

Figure D.16: The density plot of **Entropy 25** from the Pythia-9.6b model, when the seen instances are from the **Pile-CC** domain, and the unseen instances are from the **Github** domain.

D.5.3 Cross-Domain Detection with Different Metrics

In this section, we present AUC results with other metrics, when seen and unseen instances are from different domains, for the Pythia-6.9b model. The metrics include Min 25% token (Figure D.17), Mem 25 (Figure D.18), and Entropy 25 (Figure D.19). All metrics exhibit higher AUC values in the top-right corner and lower values in the bottom-left, while diagonal points approach random guessing.

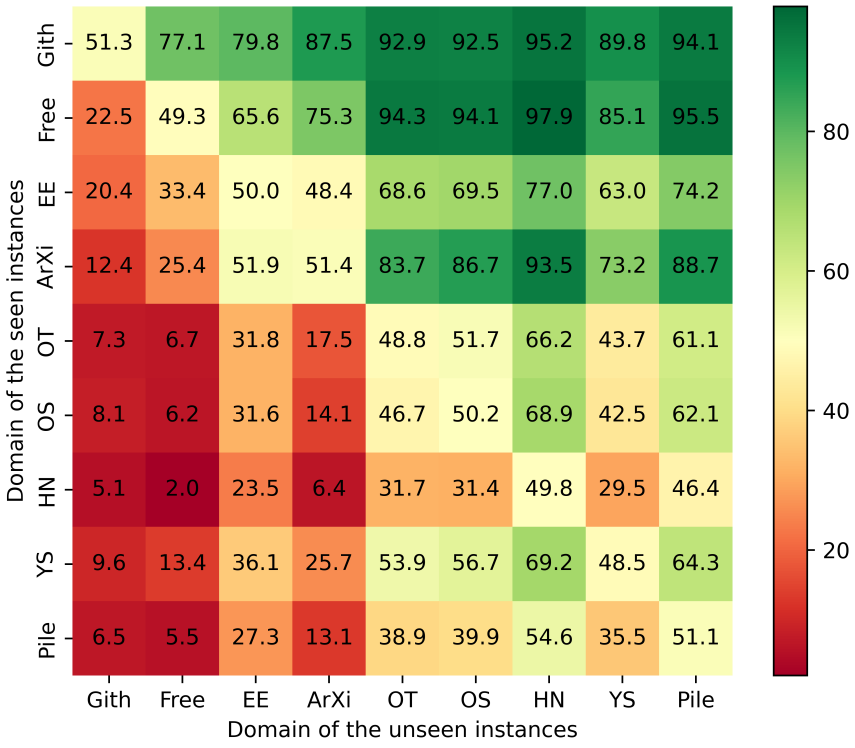

Figure D.17: Average contamination detection AUC for the Pythia-6.9b model with the metric, **Min 25% token**, when the seen and unseen instances are from different domains. The abbreviations represent the domains in Table D.9.

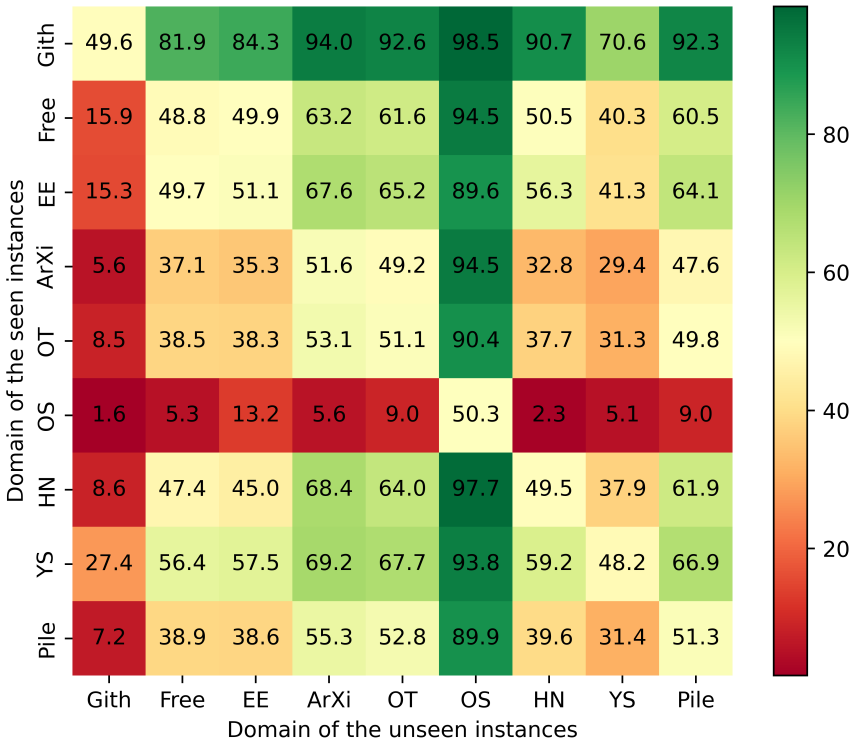

Figure D.18: Average contamination detection AUC for the Pythia-6.9b model with the metric, **Mem 25**, when the seen and unseen instances are from different domains. The abbreviations represent the domains in Table D.9.

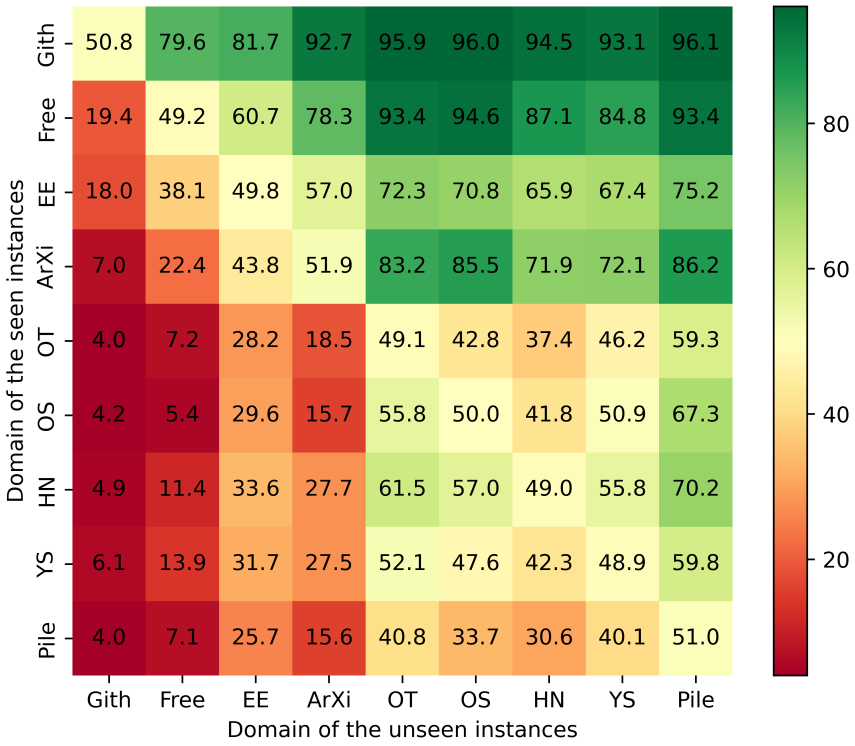

Figure D.19: Average contamination detection AUC for the Pythia-6.9b model with the metric, **Entropy 25**, when the seen and unseen instances are from different domains. The abbreviations represent the domains in Table D.9.