A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law

Anonymous authors
Paper under double-blind review

Abstract

In the fast-evolving domain of artificial intelligence, large language models (LLMs) such as GPT-3 and GPT-4 are revolutionizing the landscapes of finance, healthcare, and law: domains characterized by their reliance on professional expertise, challenging data acquisition, high-stakes, and stringent regulatory compliance. This survey offers a detailed exploration of the methodologies, applications, challenges, and forward-looking opportunities of LLMs within these high-stakes sectors. We highlight the instrumental role of LLMs in enhancing diagnostic and treatment methodologies in healthcare, innovating financial analytics, and refining legal interpretation and compliance strategies. Moreover, we critically examine the ethics for LLM applications in these fields, pointing out the existing ethical concerns and the need for transparent, fair, and robust AI systems that respect regulatory norms. By presenting a thorough review of current literature and practical applications, we showcase the transformative impact of LLMs, and outline the imperative for interdisciplinary cooperation, methodological advancements, and ethical vigilance. Through this lens, we aim to spark dialogue and inspire future research dedicated to maximizing the benefits of LLMs while mitigating their risks in these precision-dependent sectors. To facilitate future research on LLMs in these critical societal domains, we also initiate a reading list that tracks the latest advancements under this topic, which will be released and continually updated.

Keywords: large language models, GPT-4, interdisciplinary research, finance, healthcare, law, ethics

1 Introduction

The advent of large language models (LLMs) such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) marks a significant milestone in the evolution of artificial intelligence. The research integrating LLMs with various disciplines, i.e., LLM+X, such as math, science, finance, healthcare, law, etc., is starting as a new epoch powered by collaborative endeavors spanning diverse communities. In this survey paper, we offer an exploration of the methodologies, applications, challenges, ethics, and future opportunities of LLMs within **critical societal domains**, including **finance, healthcare**, and **law**. These domains are major cornerstones of societal function and well-being, each playing a critical role in the fabric of daily life and the broader economic and social systems. They are frequently discussed together due to shared characteristics, including the *reliance on extensive professional expertise*, *highly confidential data*, *extensive multimodal documents*, *high legal risk and strict regulations*, and the *requirement for explainability and fairness.*

Reliance on Professional Expertise. These domains require extensive professional knowledge and experience. The finance domain involves complex financial analysis, investment strategies, and economic forecasting, necessitating deep knowledge of financial theories, market behavior, and fiscal policy (Benninga, 2014; Fridson & Alvarez, 2022; Roberts, 1959; Geels, 2013; Franses, 1998; Easterly & Rebelo, 1993). Healthcare requires specialized knowledge in medical sciences, patient care, diagnostics, and treatment planning, and professionals are trained for years in their specific fields (Melnick et al., 2002; Gowda et al., 2014; P. Collins, 1998; Brauer & Ferguson, 2015; Thomas et al., 2022). The legal domain demands a thorough understanding of legal principles, statutes, case law, and judicial procedures, with practitioners spending extensive periods in legal education and training (Hart & Green, 2012; Dworkin, 1986; Sunstein, 2018; Epstein & Sharkey, 2020;

Friedman, 2005; Chemerinsky, 2023; MacCormick, 1994).The need for profound professional expertise in these domains presents significant challenges in equipping LLMs with the requisite knowledge and capabilities (Li et al., 2023d; Xie et al., 2024b; Islam et al., 2023; Nori et al., 2023a;b; Tan et al., 2023; Yu et al., 2022; Choi et al., 2021; Iu & Wong, 2023).

Highly Confidential Data. Unlike many other domains where data might be more public or less sensitive, finance, healthcare, and law deal with information that is mostly personal and confidential. This brings unique challenges for LLM-based research, which is essentially data-driven. LLMs must be trained and tested in a manner that prevents data breaches or inadvertent disclosures. This necessitates research challenges such as training data synthesis, encryption techniques, secure data handling practices, transfer learning, etc.

Extensive Multimodal Documents. The complexity and multimodal nature of documents in these sectors mark another unique challenge. Financial documents may contain not only text but also tables and charts in diverse structures (Chen et al., 2021b; Bhatia et al., 2024). Healthcare data may contain text and various medical imaging modalities Gee et al. (2004); Wood et al. (2020); Yan et al. (2023c), such as X-ray Radiography, Ultrasound, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Legal documents may contain text, images of evidence, audio recordings of testimonies, or video depositions (Matoesian & Gilbert, 2018; He et al., 2023a). Developing LLMs that can accurately interpret and correlate information across modalities is crucial, demanding innovative approaches to model architecture and data processing.

High Legal Risk and Strict Regulations. Considering the potential serious consequences of actions in finance, healthcare, and law, the regulatory landscape in these domains is more complex and stringent than in many other fields. They must adhere to rigorous standards and laws from the outset to protect client welfare and ensure compliance. Such requirements pose unique challenges for developing LLM-based applications, as researchers need to design models carefully to ensure compliance with regulations (Meskó & Topol, 2023; Minssen et al., 2023; Zhang et al., 2023g; Ong et al., 2024; Hacker et al., 2023; Gilbert et al., 2023). LLMs must incorporate mechanisms to ensure regulatory compliance to not only achieve accuracy but also be extraordinarily aware of legal and regulatory nuances.

Requirement for Explainability and Fairness. Explainability and fairness have emerged as vital components of AI (Adadi & Berrada, 2018; Burkart & Huber, 2021; Mehrabi et al., 2021; Caton & Haas, 2020; Dong et al., 2023), ensuring transparent decision-making processes and guarding against biased outcomes. Particularly in knowledge-intensive and high-stakes domains like finance, healthcare, and law, decision-making often involves professional expertise and complicated processes. Furthermore, these decisions can directly influence people's lives in significant aspects (e.g., economic status, health, and legal rights). These facts necessitate a higher standard of transparency and bias mitigation of model design in these critical societal domains to maintain public trust and compliance with ethical guidelines Beauchamp & Childress (2001); Cranston (1995); Yamane (2020); Svetlova (2022). Developing LLM-based applications that offer transparent reasoning and minimize bias is vital for any real-world deployment in these domains.

Focusing on finance, healthcare, and law, this paper explores the extensive spectrum of LLM applications, underscoring LLM's transformative effects across these critical societal sectors. This exploration sheds light on how LLMs are reshaping traditional research methodologies in these important fields, and fostering the innovation, efficiency, and social impact of next-generation AI. More specifically, the rest of paper organization is as follows. We discuss related surveys in Section §2. We investigate the domains of finance, healthcare, and law in Section §3, §4, and §5, respectively. In Section §6, we consider a series of ethical concerns regarding adopting LLMs in these domains. Finally, we make conclusions in Section §7.

2 Related Surveys

Along with the rapidly evolving LLM research, there is a surge of LLM-related survey literature that explores a wide range of perspectives and aspects of LLM development. In addition to the surveys investigating the overall development of LLMs (Zhao et al., 2023a; Min et al., 2023a), recent surveys include fine-grained areas such as alignment (Shen et al., 2023a; Wang et al., 2023d; Liu et al., 2023d), augmentation (Mialon et al., 2023; Gao et al., 2023b), instruction tuning (Zhang et al., 2023e), reasoning (Huang & Chang, 2022; Qiao

et al., 2022), compression (Zhu et al., 2023a), evaluation (Chang et al., 2023), explainability (Zhao et al., 2024), and hallucination (Zhang et al., 2023l; Huang et al., 2023), as well as bias, fairness, and safety (Gallegos et al., 2023; Navigli et al., 2023; Li et al., 2023e; Weidinger et al., 2021; Yao et al., 2024; Shayegani et al., 2023). Surveys on LLM-based research in NLP tasks are also prevalent, such as text generation (Li et al., 2022a; Zhang et al., 2023c), code generation (Zan et al., 2022), information retrieval (Zhu et al., 2023b), recommendation (Wu et al., 2023c), etc. As the study of LLM+X becomes increasingly popular, surveys in this direction have also begun to emerge in domains such as robotics (Zeng et al., 2023), education (Yan et al., 2024b), software engineering (Fan et al., 2023), causal inference (Liu et al., 2024c), etc. In contrast with existing surveys that mostly focus on LLM integration for NLP tasks or STEM disciplines, our survey investigates LLMs in three critical societal sectors: finance, healthcare, and law.

In the finance domain, there are existing surveys on AI, machine learning, or deep learning in finance (Cao, 2022; Cao & Zhai, 2022; Maple et al., 2023; Ozbayoglu et al., 2020; Rundo et al., 2019), as well as general NLP techniques in finance (Xing et al., 2018; Fisher et al., 2016; Gao et al., 2021b; Gupta et al., 2020; Kumar & Ravi, 2016). In contrast, our survey focuses on cutting-edge LLM development in finance. The works by Li et al. (2023g) and Lee et al. (2024) address LLM techniques in finance. However, they primarily introduce general LLM techniques, financial-specific LLMs, and financial tasks. In contrast, our survey on finance not only covers more thorough explorations of financial tasks and financial-specific LLMs, but also investigates performance comparisons and analysis for LLMs, offering insights and guidance for future research. Furthermore, our survey explores LLM-based methodologies and adjacent research, concluding with a broad discussion of future prospects, emphasizing their implications for critical societal sectors from a comprehensive range of viewpoints.

In the medical domain, previous studies have extensively explored applications of machine learning (Garg & Mago, 2021; Shehab et al., 2022) and deep learning (Piccialli et al., 2021; Egger et al., 2022; Miotto et al., 2018), with specific emphasis on NLP within medical contexts (Chary et al., 2019; Wu et al., 2020; Liu et al., 2022; Kalyan & Sangeetha, 2020). Our survey broadens the scope by including Large Language Models (LLMs) and their diverse applications in the medical field. Concurrently, Zhou et al. (2024a) investigates LLMs in the medical domain focusing primarily on single modality applications, while another work Hartsock & Rasool (2024) focused on medical Visual Question Answering (VQA) and report generation specifically. Our work encompasses a broader spectrum, surveying various applications in both the pure NLP domain and multimodal scenarios. We also discuss recent novel tasks such as medical instruction following and medical imaging classification via natural language.

In the law domain, there are existing surveys on AI in law (Chalkidis & Kampas, 2019; Cui et al., 2023b; Dias et al., 2022), our work improves the focus on current developments in LLMs within the legal area. While the studies by Katz et al. (2023) and Sun (2023) provide an overview of LLM techniques in legal contexts, they primarily discuss generalized LLM techniques alongside legal-specific LLMs and tasks. Our analysis extends beyond these initial explorations, offering a comprehensive examination of legal tasks catered to by legal-specific LLMs and conducting in-depth performance comparisons and analytical reviews of LLMs. This affords pivotal insights and directional guidance for burgeoning research. Furthermore, our survey explores LLM-based methodologies and allied areas of study, ultimately leading to an expansive discourse on future prospects. We place particular emphasis on the augmentation of datasets and the consideration of non-structured knowledge. Additionally, we underscore the importance of enhancing LLM interpretability and the integration of ancillary tools, which together forecast a revolutionary impact on key societal legal institutions.

With regard to ethics, there have been a few systematic surveys (Khan et al., 2022; Kaur et al., 2022; Mehrabi et al., 2021; Caton & Haas, 2020; Dong et al., 2023) and specifically ethics in LLMs (Liu et al., 2023d; Sun et al., 2024; Ray, 2023; Yao et al., 2024). In contrast to most of these surveys, which introduce ethics or trustworthiness in a general sense, in this work we focus on the ethics in the three critical domains of finance, healthcare, and law. More specifically, we highlight several ethical principles and considerations that are considered most important in these domains, showcase unique definitions and examples of ethics from different domains under these ethical principles, and summarize the progress of existing domain-specific studies for LLM ethics in the three domains respectively.

3 Finance

In this Section, we introduce the existing NLP tasks in the finance domain, including task formulations and datasets. In §3.2, we investigate various Pre-trained Language Models (PLMs) and LLMs developed for finance. In §3.3, we study the evaluations and analysis of the performance of various LLMs. In §3.4, we study various LLM-based methodologies developed for financial tasks and challenges. Finally, we summarize insights, make conclusions, and discuss potential future directions.

3.1 Tasks and Datasets in Financial NLP

In this section we introduce existing financial tasks and datasets studied extensively using LLM-related methods, including sentiment analysis, information extraction, question answering, text-enhanced stock movement prediction, and others. We also discuss additional financial NLP tasks that are mostly under-explored for LLM-based methods, suggesting future research opportunities. Figure 1 provides a summary of existing financial NLP tasks.

Sentiment Analysis (SA). The task of financial sentiment analysis aims at analyzing textual data related to finance, such as news articles, analyst reports, and social media posts, to gauge the sentiment or mood conveyed about specific financial instruments, markets, or the economy as a whole. An automatic analysis of the sentiments can help investors, analysts, and financial institutions to make more informed decisions by providing insights into market sentiment that might not be immediately apparent from quantitative data. The task of financial sentiment analysis is often formulated as a classification problem, with the input as the text to be analyzed and the target label as sentiment orientations such as positive, negative, or neutral. The Financial Phrase Bank dataset (Malo et al., 2014) is based on company news (in English), with the target sentiment categories from the investor's perspective. The FiQA[1] Task 1 focuses on aspect-based financial sentiment analysis, where the target is given as continuous numeric values. TweetFinSent (Pei et al., 2022) is another dataset based on stock tweets. The authors propose a new concept of sentiment labels indicating the opinion of stock movement forecasting. FinSent (Guo et al., 2023) is another sentiment classification dataset based on sentences from analyst reports of S&P 500 firms. In the evaluation of the financial language model BloombergGPT (Wu et al., 2023d), the authors also propose a set of sentiment analysis datasets.

Information Extraction (IE). Information extraction involves several key tasks that are essential for analyzing and understanding financial texts. **Named Entity Recognition (NER)** targets identification and classification of key entities in text, such as company names, stock symbols, financial metrics, and monetary values. In (Alvarado et al., 2015), a NER dataset is proposed, aiming at extracting fields of interest for risk assessment in financial agreements. In BloombergGPT (Wu et al., 2023d), the internal financial datasets proposed include NER over various sources. **Relation Extraction (RE)** focuses on identifying and categorizing finance-specific semantic relationships between the entities, such as *cost_of*, *acquired_by*. REFinD (Kaur et al., 2023) is a large-scale RE dataset built upon the 10-X filings from the Securities and Exchange Commission (SEC), focusing on common finance-specific entities and relations. **Event Detection** involves identification of significant financial occurrences like acquisitions, earnings reports, or stock repurchases, from sources such as news or social media. The EDT (Zhou et al., 2021) dataset focuses on detecting corporate events from news articles, aiming at predicting stock movements. In (Oksanen et al., 2022), the authors propose building knowledge graphs over SEC filings. These information extraction tasks play a fundamental role in transforming raw financial texts into structured, actionable insights, aiding professionals in more effective and efficient analysis.

Question Answering (QA). Question answering (QA) in finance involves building systems to answer finance-specific queries, typically from large volumes of financial data such as online forums, blogs, news, etc. Such QA systems can aid professionals in performing efficient financial analysis and decision-making. The FiQA[3] Task 2 is an early financial QA dataset targeting opinion-based QA over microblogs, reports, and news. Due to the fact that company financial documents contain large amounts of numeric values, which are of great importance for analysis and decision-making, later works have begun to explore complex QA involving

[1] https://sites.google.com/view/fiqa/home/
[2] https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/shared-task-finsbd
[3] https://sites.google.com/view/fiqa/home/

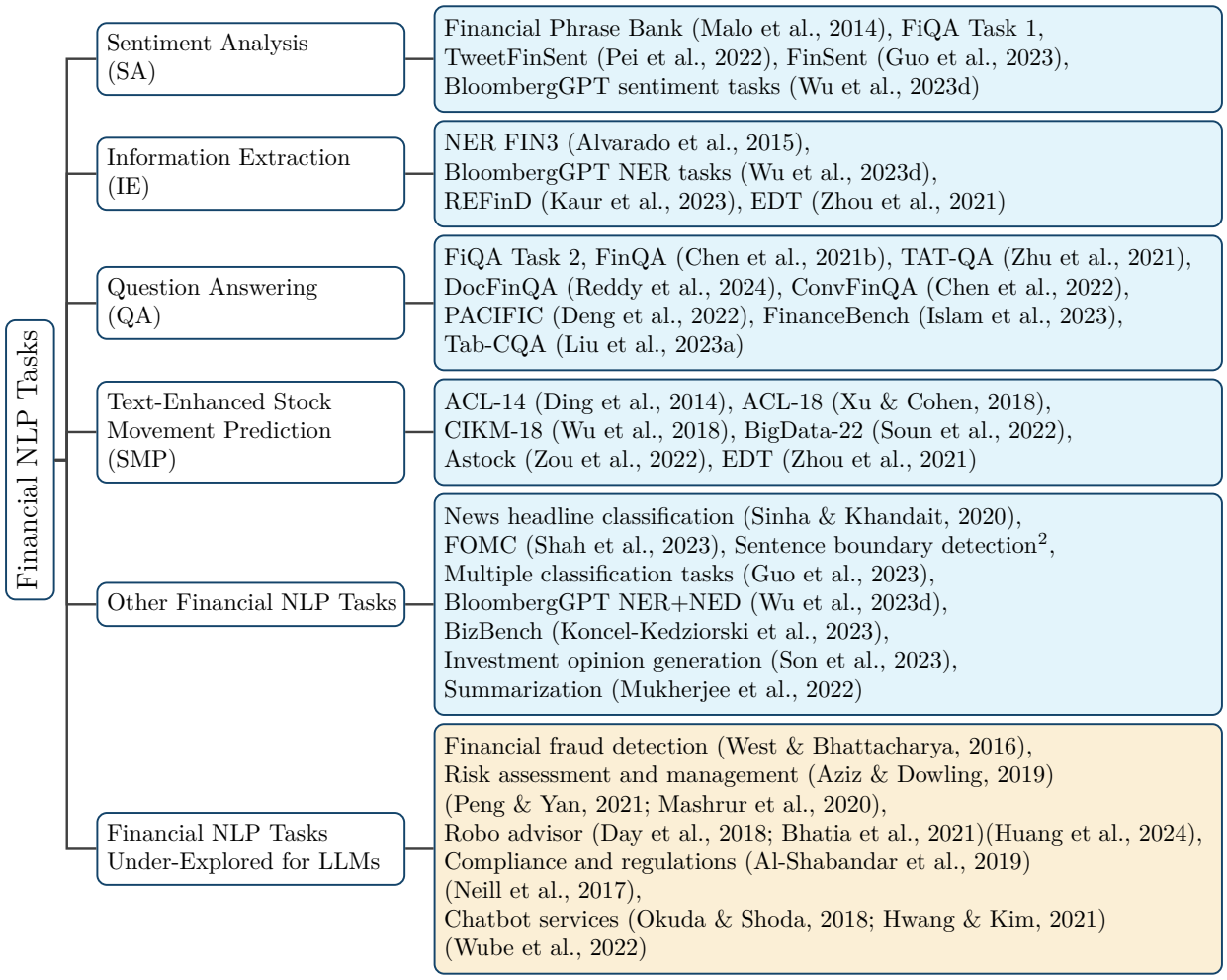

Figure 1: A summarization of existing financial NLP tasks and representative datasets. The yellow field shows the tasks relatively under-explored for LLMs.

numerical reasoning. The FinQA (Chen et al., 2021b) dataset offers expert annotated QA pairs over earnings reports of S&P 500 companies. The questions require complex numerical reasoning over both the textual and tabular contents in the reports to answer. TAT-QA (Zhu et al., 2021) is another large-scale QA dataset over financial reports containing both textual and tabular contents. In addition to numerical reasoning questions with arithmetic expressions to answer, the dataset also contains extractive questions whose ground truth answers are span or multiple spans from the input report. DocFinQA (Reddy et al., 2024) extends FinQA to the long-context setting, aiming to explore long-document financial QA. ConvFinQA (Chen et al., 2022) extends FinQA to a conversational QA setting involving numerical reasoning over the entire conversation history to capture long-term dependencies. PACIFIC (Deng et al., 2022) is another conversational QA dataset built upon TAT-QA (Zhu et al., 2021), focusing on building proactive assistants that ask clarification questions and resolve co-references. FinanceBench (Islam et al., 2023) is a large-scale QA dataset focusing on the open-book setting covering diverse sources and scenarios. Tab-CQA (Liu et al., 2023a) is a tabular conversational QA dataset created from Chinese financial reports of listed companies in a wide range of sectors.

Text-Enhanced Stock Movement Prediction (SMP). Text-enhanced stock movement prediction involves analyzing financial texts like news, reports, and social media to forecast stock price trends and market behavior. The task is mostly formulated to predict stock movement for a target day based on the financial texts and historical stock prices in a time window. In (Ding et al., 2014), the authors propose using extracted events to make stock predictions, and they construct a dataset for Standard & Poor's 500 stock

(S&P 500) based on financial news from Reuters and Bloomberg. In (Xu & Cohen, 2018; Soun et al., 2022; Wu et al., 2018), datasets are proposed based on stock-specific tweets from Twitter. Astock (Zou et al., 2022) is a Chinese dataset providing stock factors for each stock, such as Price to Sales ratio, turnover rate, etc. In the EDT (Zhou et al., 2021) dataset, the authors propose to make stock predictions immediately after a news article is published and to perform tradings. The proposed EDT dataset includes minute-level timestamps and detailed stock price labels.

Other Financial NLP Tasks. In (Sinha & Khandait, 2020), the authors propose a dataset for classification of news headlines regarding the gold commodity PRICE? into semantic categories such as price up and price down. In (Shah et al., 2023), the authors propose a dataset on hawkish-dovish classification based on monetary policy pronouncements by the Federal Open Market Committee (FOMC). The FinSBD-2019 Shared Task[4] proposes the task of financial sentence boundary detection, with the goal of extracting well-segmented sentences from financial prospectuses. In (Guo et al., 2023), the authors propose an evaluation framework including a set of proprietary classification datasets. In BloombergGPT (Wu et al., 2023d), one of the internal evaluation tasks proposed is NER+NED, namely NER followed by named entity disambiguation (NED). The goal is first to identify company mentions in financial documents and then to generate the corresponding stock ticker. In the recent BizBench (Koncel-Kedziorski et al., 2023) benchmark, the authors aim to evaluate financial reasoning abilities and propose eight quantitative reasoning datasets. In (Son et al., 2023), the authors propose the task of financial investment opinion generation based on analyst reports to evaluate the LLMs' ability to conduct financial reasoning for investment decision-making. In (Mukherjee et al., 2022), the authors propose a dataset for bullet-point summarization from long earnings call transcripts.

Financial NLP Tasks Under-Explored for LLMs. The above four categories summarize the current tasks and datasets covered in LLM-related studies. The overall financial NLP space is broader and still has many existing tasks under explored for LLMs. Financial fraud detection is a critical issue with severe consequences in financial activities. There are numerous studies spanning data mining and NLP techniques for financial fraud detection (West & Bhattacharya, 2016; Throckmorton et al., 2015; Seemakurthi et al., 2015; Goel & Uzuner, 2016; Pandey, 2017; Chen et al., 2017; Boulieris et al., 2023; Craja et al., 2020; Calafato et al., 2016), such as detecting fraud in transactions, financial statements, annual reports, tax, etc. Research studies on financial fraud detection using LLM-based methods are largely under-explored. Other tasks remaining relatively open for LLM research include financial risk assessment and management (Aziz & Dowling, 2019; Peng & Yan, 2021; Mashrur et al., 2020; Cheng et al., 2021; Li et al., 2020a; Zou et al., 2017; Giudici, 2018), robo advisor (Day et al., 2018; Bhatia et al., 2021; Huang et al., 2024), compliance and regulations (Al-Shabandar et al., 2019; Neill et al., 2017), chatbot services (Okuda & Shoda, 2018; Hwang & Kim, 2021; Wube et al., 2022), etc. These tasks mostly lack well-defined formulations and well-established public datasets due to their complexity. Given their importance in the financial sector, they are all valuable future directions for incorporating LLM-based research.

3.2 Financial LLMs

Since the invention of BERT (Devlin et al., 2019), there have been numerous efforts to build PLMs and LLMs specialized for finance. Consistent with the evolving modeling paradigms of general PLMs and LLMs, early financial PLMs adopted the pre-training followed by downstream task fine-tuning paradigm, and train relatively small language models. Recent works scale the model sizes up and conduct instruction fine-tuning, with the evaluation covering broader sets of financial tasks. Most existing financial LLMs are in single text modality, either in English or Chinese. Table 1 summarizes the PLMs and LLMs for the financial domain.

Pre-Training and Downstream Task Fine-Tuning PLMs. FinBERT-19 (Araci, 2019) was an early attempt to build financial pre-trained language models targeting the task of financial sentiment analysis. The authors first conducted further pre-training on BERT (Devlin et al., 2019) using a financial corpus, followed by fine-tuning using task training data. FinBERT-20 (Yang et al., 2020) is another PLM on sentiment analysis using similar training strategies, including further pre-training on BERT and pre-training from scratch. FinBERT-21 (Liu et al., 2021) is pre-trained on a large scale of both general domain and financial domain corpus from scratch based on BERT architecture, with a set of self-supervised pre-training

[4]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp/shared-task-finsbd

Model Name	Model Architecture	Evaluation Tasks	Languages	Size	Year
Pre-training & Downstream task fine-tuning approaches					
FinBERT-19 (Araci, 2019)	BERT	SA	English	110M	2019
FinBERT-20 (Yang et al., 2020)	BERT	SA	English	110M	2020
FinBERT-21 (Liu et al., 2021)	BERT	SA, QA, Others	English	110, 340M	2021
Mengzi-BERTbase-fin (Zhang et al., 2021)	RoBERTa	IE, Others	Chinese	103M	2021
FLANG (Shah et al., 2022)	BERT, ELECTRA	SA, IE, QA, Others	English	110M	2022
BBT-Fin (Lu et al., 2023)	T5	SA, IE, QA, Others	Chinese	220M, 1B	2023
Pre-training approaches					
BloombergGPT (Wu et al., 2023d)	BLOOM	SA, IE, QA, Others	English	50B	2023
Instruction fine-tuning approaches					
FinMA (Xie et al., 2023)	LlaMA	SA, IE, QA, SMP, Others	English	7B, 30B	2023
Instruct-FinGPT (Zhang et al., 2023a)	LlaMA-7B	SA	English	7B	2023
InvestLM (Yang et al., 2023a)	LlaMA-65B	SA, QA, Others	English	65B	2023
FinGPT (Wang et al., 2023c)	Six 7B models	SA, IE, Others	English	6B, 7B	2023
CFGPT (Li et al., 2023c)	InternLM-7B	Others	Chinese	7B	2023
DISC-FinLLM (Chen et al., 2023b)	Baichuan-13B	SA, IE, QA, Others	Chinese	13B	2023
FinMA-ES (Zhang et al., 2024b)	LlaMA2-7B	SA, IE, QA, SMP, Others	English, Spanish	7B	2024
FinTral (Bhatia et al., 2024)	Mistral-7B	SA, IE, QA, SMP, Others	English	7B	2024

Table 1: Summary of financial pre-trained language models. For evaluation tasks, we have **SA** for sentiment analysis, **IE** for information extraction, **QA** for question answering, **SMP** for text-enhanced stock movement prediction, and **Others** for other tasks out of the above three major categories. For the time of release, we report the initial release year of each work.

tasks. Mengzi-BERTbase-fin (Zhang et al., 2021) is a Chinese model based on the RoBERTa (Liu et al., 2019) architecture, pre-trained with general web corpus and financial domain corpus. FLANG (Shah et al., 2022) is another English model based on BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) using pre-training techniques in ELECTRA. BBT-Fin (Lu et al., 2023) is a Chinese pre-trained language model based on the T5 (Raffel et al., 2020) architecture and pre-training schema. The authors propose a large pre-training financial corpus in Chinese as well as a set of Chinese financial benchmarks, including classification and generation tasks.

Pre-training LLMs. BloombergGPT (Wu et al., 2023d) is a large English financial language model built by Bloomberg. As rich financial resources can be obtained within Bloomberg, there is a vast amount of well-curated company financial documents employed for model pre-training, including web content, news articles, company filings, press releases, Bloomberg-authored news, and other documents such as opinions and analyses. Together with three public general domain datasets - The Pile (Gao et al., 2021a), C4 (Raffel et al., 2020) and Wikipedia, the authors created a large training corpus with over 700 billion tokens. The model architecture is based on BLOOM (Scao et al., 2022). The authors note that the common approach of greedy sub-word tokenization (Sennrich et al., 2016; Wu et al., 2016) is not efficient for financial tasks, as it does not handle numeric expressions very well. Instead, they use a unigram-based approach inspired by (Kudo, 2018). In addition, as a pre-processing step, they separate numeric and alphabetic unigrams. To assess the model performance, the authors conducted evaluations based on both external tasks with public datasets and internal tasks with datasets annotated by Bloomberg financial experts. All models are evaluated using vanilla few-shot prompting. For external tasks, including sentiment analysis, headline classification, NER, and conversational QA, BloombergGPT achieved significant improvements for all tasks except NER, over baseline LLMs of similar size. For internal tasks, including sentiment analysis, NER, and NER+NED, BloombergGPT outperforms baseline LLMs for most datasets except NER. For NER, BloombergGPT slightly underperforms the larger 176B BLOOM (Scao et al., 2022) model. Notably, despite increasing the vocabulary size, BloombergGPT's tokenization method improves the efficiency of token representations, and the model outperforms open-domain LLMs on financial QA tasks that involve numerical reasoning. Due to data leakage concerns, BloombergGPT has not been released for public usage.

Instruction Fine-Tuning LLMs. FinMA (Xie et al., 2023) is an open-sourced financial LLM built from instruction fine-tuning on LlaMA (Touvron et al., 2023a). The authors note that financial data is often

expressed in multimodal contexts such as tables and time-series representations. They develop FLARE, a large instruction-tuning dataset that covers 136 thousand examples from a collection of diverse financial tasks, and contains instructions over tabular and time-series data. For evaluation, FinMA-30B outperforms BloombergGPT and GPT-4 in classification tasks like sentiment analysis and headline classification. Despite being tuned on FLARE, FinMA falls short of GPT-4 (Achiam et al., 2023) on quantitative reasoning benchmarks such as FinQA (Chen et al., 2021b) and ConvFinQA (Chen et al., 2022). The reason could be the lack of proper numerical reasoning process generation data in the instruction-tuning dataset. Instruct-FinGPT (Zhang et al., 2023a) is a model built over LlaMA-7B (Touvron et al., 2023a) using instruction fine-tuning, specifically targeting the task of financial sentiment analysis. CFGPT (Li et al., 2023c) is a Chinese model based on InternLM with continued pre-training and instruction fine-tuning. Following the Superficial Alignment Hypothesis (Zhou et al., 2023), InvestLM (Yang et al., 2023a) is trained with instruction fine-tuning using a well-curated set of 1.3k examples over LlaMA-65B. The resulting model achieves comparable performance and sometimes surpasses the proprietary models like GPT-3.5. DISC-FinLLM (Chen et al., 2023b) is a Chinese model based on instruction fine-tuning on Baichuan-13B[5], that performs instruction fine-tuning on separate LoRA (Hu et al., 2022) modules for each type of task. FinGPT (Wang et al., 2023c) is a series of 6B/7B models trained with instruction fine-tuning. In (Liu et al., 2023c), the authors propose a framework FinGPT consisting of data collection and processing pipeline from diverse sources, as well as financial LLM fine-tuning using reinforcement learning with stock prices. Most recently, researchers have begun to explore financial LLMs in broader settings. In (Zhang et al., 2024b), the authors propose instruction datasets, fine-tuned model FinMA-ES, and evaluation benchmarks in a bilingual setting of Spanish and English. FinTral (Bhatia et al., 2024) is a series of multimodal LLMs based on Mistral-7B (Jiang et al., 2023). The authors incorporate tool usage (Schick et al., 2023), RAG (Lewis et al., 2020), and visual understanding based on CLIP (Radford et al., 2021a). This allows the authors to explore multimodal contexts, and to include visual reasoning tasks such as question answering over charts and graphs. Despite being pre-trained and instruction-tuned on multimodal data, the FinTral-DPO model underperforms on visual reasoning tasks compared to other SotA multimodal LLMs such as Qwen-VL-Plus (Bai et al., 2023), and GPT-4V (Achiam et al., 2023), but is on par or better than open-source LLMs of similar size. These challenges point to the need for cross-disciplinary research at the intersection of Multimodal LLMs (MLLMs), quantitative reasoners, and financial LLMs.

3.3 Evaluation and Analysis

Performance Evaluation and Analysis for Popular Financial Tasks. In (Li et al., 2023d; Xie et al., 2023; Yang et al., 2023a), the authors conducted experiments on several popular financial datasets using an array of methods. Table 2 summarizes the performance of various methods on two sentiment analysis datasets, one headline classification task, and one NER dataset. For sentiment analysis, GPT-4 and the recent instruction fine-tuning models like FinMA (Xie et al., 2023) achieve similar performance as the best fine-tuning methods. For headline classification, FinMA also slightly surpasses the best fine-tuning method. We anticipate that the performances on such datasets have nearly reached saturation. As suggested by (Li et al., 2023d), adopting generalist LLMs could be an easy choice for relatively simple financial tasks. For IE tasks like NER, there is still a gap between LLMs and fine-tuning methods. On the relation extraction dataset REFinD (Kaur et al., 2023), CPT-3.5 and GPT-4 still largely fall behind the fine-tuning model (Li et al., 2023d).

Figure 2 shows the performance comparisons between various methods on the FinQA dataset (Chen et al., 2021b). Large, general LLMs like GPT-4 still achieve the leading performances with simple prompts, due to the strong knowledge and reasoning ability achieved during pre-training. Domain-specific fine-tuning models follow behind. Instruction fine-tuning models fall far behind compared to the former. Note that in the construction of instruction fine-tuning data in the FinMA (Xie et al., 2023) model, the authors did not include the generation of reasoning programs in the FinQA dataset (Chen et al., 2021b), which could be a major reason for the inferior performance on the FinQA dataset. Therefore, we anticipate that there is still

[5]https://github.com/baichuan-inc/Baichuan-13B
[7]https://sites.google.com/view/fiqa/home/

Model	FPB		FiQA-SA	Headline	NER FIN3
	Accuracy	F-1	Weighted F-1	Weighted F-1	Entity F-1
Fine-tuning	0.86	0.84	0.87	0.95	0.83
BloombergGPT (Wu et al., 2023d) (few-shot)	-	0.51	0.75	0.82	0.61
LlaMA-65B (zero-shot)	-	0.38	0.75	-	-
InvestLM (Yang et al., 2023a) (zero-shot)	-	0.71	0.90	-	-
FinMA-30B (Xie et al., 2023) (zero-shot)	0.87	0.88	0.87	0.97	0.62
GPT-3.5 (zero-shot)	0.78	0.78	0.76	0.72	0.29
GPT-3.5 (few-shot)	0.79	0.79	0.78	0.75	0.52
GPT-4 (zero-shot)	0.83	0.83	0.87	0.84	0.36
GPT-4 (few-shot)	0.86	0.86	0.88	0.86	0.57

Table 2: Performance comparisons for sentiment analysis tasks (FPB dataset (Malo et al., 2014) and FiQA-SA dataset[7]), headline classification task (Headline dataset (Sinha & Khandait, 2020)), and IE task (NER FIN3 dataset (Alvarado et al., 2015)). The few-shot setting is five shots for FPB, FiQA-SA, and Headline, and twenty shots for NER FIN3. For *Fine-tuning* the model, we select the best performance achieved through fine-tuning models in each dataset respectively. We aggregate the results from (Li et al., 2023d; Xie et al., 2023; Yang et al., 2023a). Note that some results reported for the same model in the above three papers differ, mostly for GPT-3.5 and GPT-4, which may need further verifications.

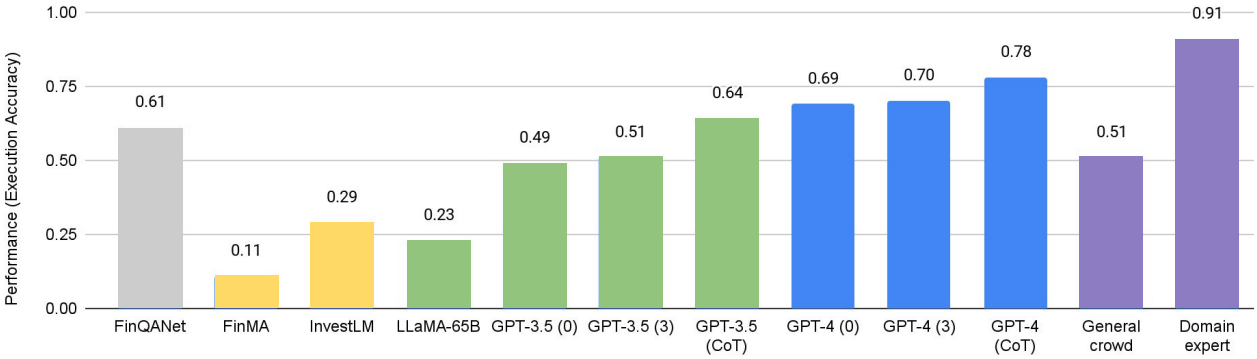

Figure 2: Performance comparison on the FinQA dataset (Chen et al., 2021b). We compare the execution accuracy following the evaluation standard in the original paper. The fine-tuning method FinQANet is the RoBERTa-based model in (Chen et al., 2021b); The instruction fine-tuning methods include FinMA (Xie et al., 2023) and InvestLM (Yang et al., 2023a); The general-purpose LLMs include LlaMA-65B, GPT-3.5 and GPT-4, with zero-shot (0), few-shot (3 shots), and Chain-of-Thought (CoT) prompting; We also list the human expert and general crowd performances. Results are sourced from (Li et al., 2023d; Xie et al., 2023; Yang et al., 2023a).

ample room for developing open-source instruction fine-tuning models to improve on tasks requiring complex reasoning abilities.

As demonstrated by (Li et al., 2023d), in most financial tasks, GPT-4 can achieve an over 10% performance increase over ChatGPT. Except for the QA task on FinQA (Chen et al., 2021b) and ConvFinQA (Chen et al., 2022), ChatGPT and GPT-4 perform either comparable or less effective than task-specific fine-tuned models. For FinQA (Chen et al., 2021b) and ConvFinQA (Chen et al., 2022), the authors argue that the reasoning complexity involved is still deemed as basic in financial analysis, but ChatGPT and GPT-4 still make simple

errors. Significant improvement is required to adopt these LLMs as trustworthy financial analyst agents in real-world industry usages.

New Evaluation Frameworks and Tasks. In (Guo et al., 2023), a financial language model evaluation framework, FinLMEval, is proposed consisting of a set of classification tasks and NER. The authors compare the performances of fine-tuned encoder-only models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and zero-shot decoder-only models, such as GPT-4 (Achiam et al., 2023) and FinMA (Xie et al., 2023). Though achieving considerable performance, the zero-shot decoder-only models mostly fall behind fine-tuned encoder-only models on these tasks. The performance gap between the fine-tuned encoder-only models and zero-shot decoder-only models is larger on their proposed proprietary datasets than on public datasets. The authors conclude that there remains room for enhancement for more advanced LLMs in the financial NLP field. Most recently, in (Xie et al., 2024b), the authors propose a large collection of evaluation benchmarks for financial tasks containing 35 datasets across 23 financial tasks. They conclude that GPT-4 mostly performs the best in quantification, extraction, understanding, and trading tasks, and the recent Google Gemini (Team et al., 2023) leads in generation and forecasting tasks.

In (Son et al., 2023), the authors propose the task of financial investment opinion generation based on analyst reports, to evaluate the ability of various LLMs, with and without instruction fine-tuning, to conduct financial reasoning for investment decision-making. The authors conducted experiments on a series of 2.8B to 13B models and concluded that the ability to generate coherent investment opinion first emerges at 6B models, and obtains improvements with instruction-tuning or larger datasets. In (Lopez-Lira & Tang, 2023), the authors study the ability of LLMs to predict stock market returns. It is found that GPT-4 outperforms other LLMs being studied on forecasting returns and delivers the highest Sharpe ratio, suggesting the great potential of advanced LLMs in the investment decision-making process. In (Zhou et al., 2024b), the authors proposed the Financial Bias Indicators (FBI) framework to evaluate the financial rationality of LLMs, including belief bias and risk preference bias. They find that model rationality increases with model size and is often influenced by the temporal bias in the financial training data. Prompting methods, such as instructional and Chain of thought (CoT) (Wei et al., 2022a), can also mitigate the biases. In (Islam et al., 2023), the authors propose the task of open-book QA to test the model's ability to handle long context. They conclude that current strong LLMs like GPT-4-Turbo still fall far behind satisfactory performances, either with a retrieval system or using a long context model. In (Callanan et al., 2023), the authors evaluate the ability of GPT-3.5 and GPT-4 in passing the first two levels of the Certified Financial Analyst (CFA) exam. Expectedly, GPT-4 outperforms GPT-3.5 in both levels, but both models struggle with longer contexts, sophisticated numerical reasoning, and tabular information, especially in Level II. The authors also demonstrate that chain of thought prompting offers limited improvement over zero-shot settings, but in-context learning with 2 or more examples produces the best results. A detailed error analysis reveals that a lack of domain knowledge leads to the majority of errors for both models, especially in Level II exams.

3.4 LLM-based Methodologies for Financial Tasks and Challenges

This section addresses LLM-based methodologies that have been proposed to tackle some of the key challenges in Financial NLP, including the scarcity of high-quality data in the public domain, the multimodal nature of many financial documents, the challenge of quantitative reasoning, the lack of domain knowledge in LLMs, and the importance of detecting or preventing hallucinations.

Confidentiality and Scarcity of High-Quality Data. Due to the confidential nature of data in the financial domain, clean and high quality datasets can be difficult to obtain (Assefa et al., 2020; Zhang et al., 2023b). In (Aguda et al., 2024), the authors assess the efficacy of LLMs in annotating data for a financial relation extraction task. While larger LLMs such as GPT-4 (Papailiopoulos, 2023) and PaLM-2 (Anil et al., 2023) outperform crowdsourced annotations, they fall far behind expert annotators, demonstrating that domain knowledge plays a crucial role. Other studies have tackled the scarcity of non-English financial training datasets (Zhang et al., 2024b; Hu et al., 2024).

Quantitative Reasoning. Reasoning over numerical data is a major component of QA and IE tasks in the financial domain. Several recent studies have proposed prompting strategies that enhance the quantitative reasoning capabilities of LLMs in financial QA tasks.

In (Wang et al., 2024b), the authors introduce ENCORE, a method that decomposes the numerical reasoning steps into individual operations, and grounds each operand within the input context. When used as a few-shot prompting strategy, ENCORE improves the performance of SotA LLMs on TAT-QA (Zhu et al., 2021) and FinQA (Chen et al., 2021b) by an average of 10.9% compared to standard Chain-of-Thought (CoT) prompting (Wei et al., 2022a). In (Chen et al., 2023c), the authors propose the Program-of-Thoughts (PoT) prompting approach that improves the numerical reasoning capability of LLMs on financial datasets, including FinQA and ConvFinQA (Chen et al., 2022). PoT explicitly prompts the model to frame its calculations as a program, using programming languages as tooling. On TAT-QA, despite better performance against other prompting strategies such as Chain-of-Thought (CoT), PoT prompting falls short of the SotA performance. An error analysis reveals that the majority of errors are due to incorrect retrieval. This may be the result of the complex structure of tabular data in the TAT-QA dataset, which does not include standardized tabular structures. In (Wang et al., 2024a), the authors show that using equations (rather than programs) as intermediate meaning representations can enhance the numeric reasoning capability of LLMs. By decomposing the prompts into sub-prompts that correspond to single equations, the authors demonstrate that the equations can better follow the logical order of calculations implied in the prompts, as opposed to programs, which have certain ordering constraints (e.g. a variable cannot be mentioned prior to being defined). Their equation-as-intermediate-meaning method, known as BRIDGE, outperforms other methods, including PoT, on math word problems. In (Zhu et al., 2024), the authors introduce TAT-LLM, a specialized LLM based on the Llama2-7B base (Touvron et al., 2023b) that can perform quantitative reasoning over text/tabular data. The authors instruction-tune the base model using a step-wise strategy that prompts the model to retrieve relevant evidence from the context, generate the reasoning steps, and produce the final answer accordingly. TAT-LLM outperforms SotA LLMs as well as SotA fine-tuned models on FinQA, TAT-QA, and TAT-DQA datasets. It improves the exact-match accuracy over the best-performing baseline (GPT-4) by an average of 2.8%.

In (Srivastava et al., 2024), the authors analyze the performance of LLMs on quantitative reasoning tasks over text/tabular contexts. They identify three common failure modes: 1) incorrect extraction of relevant evidence from the input, 2) incorrect generation of the reasoning program, and 3) incorrect execution of the program. An analysis of four financial QA datasets shows that reasoning and calculation errors dominate in datasets that provide standardized tabular data (FinQA (Chen et al., 2021b) and ConvFinQA (Chen et al., 2022)), whereas in datasets with complex tabular structures (TAT-QA (Zhu et al., 2021) and MultiHiertt (Zhao et al., 2022)), extraction errors are more common.

Multimodal Understanding. As noted above, the complex structure of tabular data can complicate numerical reasoning. Documents with visually rich content and complex layouts are common in financial domains, and studies such as (Ye et al., 2023a) and (Wang et al., 2023a) have demonstrated that the incorporation of visual and spatial features in the representation of text can enhance the performance of LLMs on tabular and visual reasoning tasks. In (Yue et al., 2024), the authors propose a framework for LLM-based information extraction from long documents containing hybrid text/tabular content. By serializing tabular data into text, dividing the documents into segments, retrieving relevant segments, and summarizing each retrieved segment, they show substantial improvement over basic prompt-based information extraction from financial documents. In (Ouyang et al., 2024), the authors demonstrate that the fusion of multimodal information (text, audio, video), with domain knowledge represented in a knowledge graph can lead to better predictions of the movement and volatility of financial assets. Notably, they use a Graph Convolutional Network (Kipf & Welling, 2017) as a universal fusion mechanism across modalities, and show the representations learned by the GCN can be used to instruction-tune an LLM to obtain superior performance to SotA approaches.

LLMs & Time-Series Data. Another data modality that is prominent in financial applications is time-series data. Research into time-series modeling has shown that pre-trained LLMs can be "patched" to model time-series data (Jin et al., 2024; Chang et al., 2024). Consistent with (Wu et al., 2023d), Gruver et al. (2023) demonstrate that number-aware tokenization enhances the performance of LLMs on time-series forecasting tasks, even in zero-shot settings. Yu et al. (2023e) demonstrate the ability of LLMs to perform explainable stock return prediction by combining time-series data with news and company metadata. The

GPT-4 model, combined with Chain-of-Thought (CoT) prompting outperforms other methods, including an instruction-tuned model based on Open LLaMA-13B.

LLMs & Hallucination. In (Kang & Liu, 2024), the authors analyze and profile the hallucination behaviors of SotA LLMs including GPT-4 and Llama-2, when applied to financial tasks. They demonstrate that a high rate of hallucination can occur with tasks that require financial domain knowledge or retrieval from pretraining data. They demonstrate that Retrieval Augmented Generation (RAG) can mitigate hallucinations on knowledge-sensitive tasks. For smaller LLMs, they recommend tuning and prompt-based tool learning as a mitigation strategy, such as data retrieval via APIs.

Modeling Domain Knowledge. A major challenge in financial applications is to fill the knowledge gap that exists between models trained on open-domain datasets, and tasks that require financial domain knowledge (Chen et al., 2021b; Aguda et al., 2024; Kang & Liu, 2024). Zhang et al. (2023b) demonstrate that instruction-tuning LLMs on domain-specific data and using Retrieval Augmented Generation can enhance their performance on financial sentiment analysis. In (Deng et al., 2023b), the authors use the Chain-of-Thought (CoT) (Wei et al., 2022a) prompting to augment an LLM with financial domain knowledge. The LLM is used to generate weak labels over social media posts, which are in turn used to train a small LM to detect market sentiment from social media. This method can be used to tackle several challenges in conventional approaches to market sentiment detection, including the scarcity of labeled data and the specificity of social media jargon. In (Zhao et al., 2023b), the authors introduce KnowledgeMath, a Math Word Problem benchmark for finance that requires college-level proficiency in the domain. They demonstrate that while knowledge augmentation techniques such as Chain-of-Thought (Wei et al., 2022a) and Program-of-Thought (Chen et al., 2023c) can enhance LLM performance by as much as 34%, the best-performing LLM achieves an overall score of 45.4, far from the human baseline of 94.

LLM Agents. The complexity of some tasks in the financial domain has motivated research into agent-based systems. In (Li et al., 2024b), the authors introduce FinMem, a multi-agent system for financial trading. The authors propose a multi-layered memory mechanism that helps LLM agents retrieve the most recent, relevant, and important events for a given trading decision. In addition, a profiling mechanism enables agents to emulate various trading personas and domains. When tested on five companies sampled from diverse trading sectors, FinMem yields significantly higher cumulative return compared to baselines. In (Park, 2024), the authors introduce a framework for anomaly detection in financial data using LLM agents. By casting agents as task experts, the author proposes a pipeline through which the agents undertake different sub-tasks such as data manipulation, detection, and verification. The output produced by these agents is then presented to "manager" agents, which will use an interactive debate mechanism (Li et al., 2023a) to derive the final output. Xing (2024) criticize the standard approach to the development of debating systems among homogeneous agents. Instead, the author proposes heterogeneous agents that can emulate roles or personas, arguing that a debate among heterogeneous agents can lead to better outcomes on semantically challenging tasks such as sentiment analysis.

3.5 Future Prospects

Tasks and Datasets. As a longstanding challenge for almost every interdisciplinary area, the knowledge gap between the NLP community and the finance community hinders the progress of financial NLP. This divide has led to the predominance of tasks focusing on shallow semantics and basic numerical reasoning, such as those illustrated by the FinQA dataset (Chen et al., 2021b), where questions typically require only rudimentary calculations like percentage changes. Such tasks, though valuable, are recognized as elementary within the realm of financial analysis (Li et al., 2023d). As shown in §3.3, current methods nearly reached saturated performance in relatively simple tasks like sentiment analysis and headline classification. In addition to the knowledge gap, another barrier is the difficulty and high cost of acquiring high-quality, real-world data. Apart from confidential data sources, creating financial datasets requires a high level of expertise. For example, the construction of the FinQA dataset (Chen et al., 2021b) involved the recruitment of eleven financial professionals for data annotation, which cost over 20k US dollars[8]. Here, we highlight some potential directions for future research:

[8]We contacted the authors of FinQA for this information.

- **Exploring Realistic Tasks.** Moving beyond surface-level tasks to embrace more sophisticated and realistic challenges is imperative. This involves the formulation of tasks that demand intricate reasoning as mentioned in §3.1, such as multi-step financial analyses, fraud detection, risk assessment, etc, that require building language agents (Wang et al., 2024c; Shinn et al., 2024; Sumers et al., 2023) capable of nuanced planning and decision-making processes in real-world financial analysis.
- **Incorporating Multimodal Documents.** Current financial NLP tasks still mostly target text or simple tables. Real-world financial documents may involve richer modalities and structures, such as complex tables with nested structures and charts of various structures. Understanding and performing financial reasoning on multiple modalities is under-explored.
- **Fostering Interdisciplinary Collaboration and Learning.** The development of high-fidelity financial NLP solutions necessitates closer collaboration between the NLP and finance sectors. In order to bridge the conceptual and methodological gaps between these fields, researchers should also take the initiative to acquire cross-disciplinary knowledge, with the goal of better understanding the imperative challenges in finance, as well as to achieve smoother and more effective communications with financial domain experts.

Methodologies. The development of LLM-based methods for financial tasks closely follows the general NLP community, from the early pre-training with downstream fine-tuning paradigm to the recent instruction fine-tuning paradigm. Task-specific fine-tuning could achieve good performance (Li et al., 2023d) but often incurs high costs. We believe that two major factors should still be emphasized when developing LLM-based methods for the financial domain: how to equip the LLMs with domain knowledge and reasoning skills, especially in a cost-effective setting. As shown in §3.3, current instruction fine-tuning approaches, especially relatively smaller models, still fall behind fine-tuning models or general LLMs in complex tasks. Though it is widely believed that large general LLMs will eventually lead to the best performance for broader domains, developing strong, lightweight domain-specific models is still a promising direction. Below are some potential future directions:

- **Knowledge-Intensive Instruction Fine-Tuning.** Beyond generic fine-tuning or instruction fine-tuning curated from existing datasets, we envision the development of novel paradigms that specifically enhance LLMs' understanding of complex financial concepts, terminologies, logics, and rules. This involves creating high-quality, finance-specific datasets for instruction fine-tuning that encapsulate the breadth and depth of the domain's knowledge.
- **Retrieval-Augmented Generation (RAG).** The RAG framework (Lewis et al., 2020) offers a compelling method for LLMs to dynamically integrate external domain knowledge into their generative processes. By adapting RAG for finance, LLMs can access and apply up-to-date market data, regulations, and financial theories, thereby enhancing their analytical and predictive capabilities.

Deployment and Applications. Despite the considerable amount of existing LLM research on finance, their real-world deployment remains scant. As suggested by (Li et al., 2023d), current LLMs excel primarily at straightforward financial NLP tasks; however, they falter when confronting more intricate challenges, failing to meet the rigorous standards of the industry. Given the high stakes of financial decision-making, where inaccuracies can precipitate substantial losses and legal entanglements, we believe the following dimensions are critical for transitioning from theoretical academia models to impactful real-world deployments:

- **Enhancing Accuracy and Robustness.** The systems cannot be deployed for real-world usages but are only limited to academic experiments until we reach a satisfactory level for industrial standards. Presently, academic benchmarks often lack the depth and realism to adequately prepare these models for practical tasks. Meanwhile, studying how to develop models that are robust to adversaries and attacks is also an important direction.
- **Evolving Human-AI Collaboration Paradigms.** How to design usage paradigms for real-world users is also an important future direction. Current systems predominantly operate under a paradigm

of user assistance, augmenting rather than replacing the expertise of financial professionals (Chen et al., 2021b; 2022). We expect that more future works could be explored, such as advanced collaboration frameworks that enhance decision-making efficacy, system transparency, and user engagement, while also embedding HCI principles to foster intuitive and efficient user experiences.

- **Navigating Responsibility, Ethics, Regulations, and Legal Concerns.** The deployment of LLMs in high-stakes domains like finance necessitates a conscientious approach to design, underscored by ethical and legal foresight. Current work in academic settings rarely addresses these considerations in a comprehensive and systematic way. Issues such as fairness, accountability for misguided financial advice, and the ethical implications of AI-driven decision-making demand rigorous attention. Future developments must prioritize responsible AI frameworks that address these concerns head-on, ensuring that LLMs contribute positively and ethically to the financial ecosystem.

4 Medicine and Healthcare

NLP has made remarkable strides in the biomedical field, providing essential insights and capabilities for various healthcare and medical applications. The recent emergence of LLMs has brought significant advancements to the medical field, primarily by incorporating extensive medical knowledge during training. This section explores the impact of LLMs on diverse biomedical tasks, benchmarks, and real-world applications. It demonstrates not only the power of LLMs in the biomedical sphere but also highlights their potential in practical medical scenarios. The organization of this section is as follows: In §4.1, we give an overview of the tasks and benchmarks in the medical domain. In §4.2, we summarize the advance of LLMs in three aspects: (1) closed-source LLMs (e.g., GPT-4 (OpenAI, 2023) and ChatGPT (OpenAI, 2022)) and their performance for medical applications; (2) open-sourced LLMs in the medical domain, including their training strategies, data, and performance; (3) multimodal medical LLMs that bridge natural language with other modalities and being applied beyond text-only tasks. In §4.3, §4.4, §4.5, and §4.6, we will delve into some of the practical applications of LLMs for clinical applications. We will present and discuss performance comparison of various task-specific methods and LLMs. Finally, in §4.7, we summarize our insights and discuss potential future directions.

4.1 Tasks and Benchmarks for Medical NLP

Sentence Understanding A fundamental task in clinical NLP is to process sentences and documents, which could help extract meaningful information from clinical documents and assist clinicians in decision-making processes. Dernoncourt & Lee (2017) proposed a dataset for sequential sentence classification, where sentences in medical abstracts are labeled with one of the following classes: background, objective, method, result, or conclusion, which can help researchers to skim through the literature more efficiently. Abnormality detection (Harzig et al., 2019; He et al., 2023c) aims to detect abnormal findings in clinical reports, with a similar goal to reduce the workload of radiologists. Ambiguity classification (He et al., 2023d) has a different purpose to focus on patient care, where it aims to find ambiguious sentences written by doctors, that could cause the misleading from patients.

Clinical Information Extraction. In the biomedical NLP community, a primary goal is the extraction of key variables from biomedical texts for effective biomedical text analysis. Clinical sense disambiguation interprets medical abbreviations within their clinical context into specific terminology, or conversely, translating medical terminology into abbreviations. This is particularly crucial for understanding clinical notes, which are frequently filled with complex jargon and abbreviations (He et al., 2023d). For example, the abbreviation 'pt' could mean patient, physical therapy, or prothrombin time, etc. This task is usually formatted as a multiple-choice problem and evaluated by accuracy and F1 scores. Biomedical evidence extraction focuses on automatically parsing clinical abstracts to extract key information, such as interventions and controls, from clinical trials, aiding the adoption of evidence-based medicine by synthesizing findings across research studies (Nye et al., 2018). Coreference resolution is essential for accurately identifying and linking noun phrases that refer to the same entity, such as a person or a medical term. This process is crucial in clinical contexts, where it helps to distinguish between a patient's own medical history and that of their family members (Zheng

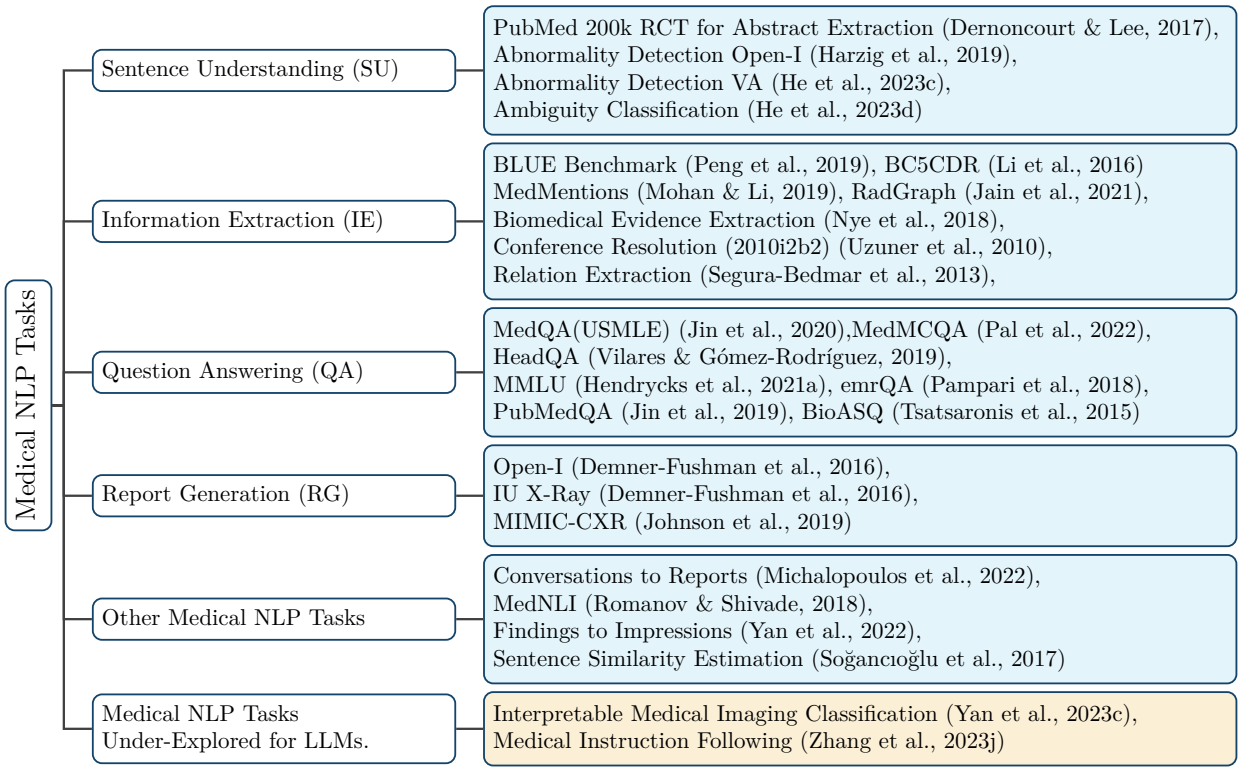

Figure 3: A summarization of medical NLP tasks and representative datasets. The yellow field shows the tasks relatively under-explored for LLMs.

et al., 2011; Chen et al., 2021a). This task has been largely evaluated on the 2010 i2b2/VA challenge, which consists of thousands of coreference chains (Uzuner et al., 2010).

Medical Question Answering. Question answering (QA) in the medical domain is a fundamental task in NLP, requiring language models to answer particular questions based on their internal medical knowledge. This task not only demands a deep understanding of clinical terminologies and concepts but also requires the capability to comprehend and interpret complex medical reasoning given the question. Medical QA tasks are mainly formed as multiple-choice questions providing a set of possible answers for each question, from which the correct one must be chosen. This format is particularly useful for testing the language model's ability to discriminate between related concepts and to understand nuances in medical knowledge. MedQA(USMLE) (Jin et al., 2020) evaluates professional biomedical and clinical knowledge through 4-way multiple-choice questions from the US Medical Licensing Exam. MedMCQA (Pal et al., 2022) is a large scale 4-way multiple-choice dataset from Indian medical school entrance exams. HeadQA (Vilares & Gómez-Rodríguez, 2019) offers multiple-choice questions from specialized Spanish healthcare exams between 2013 and 2017, with 2013 and 2014 featuring five-option tests and 2015 to 2017 having four-option tests. MMLU (Hendrycks et al., 2021a) includes a section of professional medicine questions with four-way multiple choices. PubMedQA (Jin et al., 2019) and BioASQ (Tsatsaronis et al., 2015) are reading comprehension datasets to answer yes/no/maybe based on a given passage.

In the following sections, we will discuss some of the representative tasks in the clinical setting, from abnormality detection and medical report generation, to some of the recently proposed tasks such as medical instruction evaluation and medical-imaging classification via natural language.

4.2 LLMs for Medicine and Healthcare

Close-sourced Medical LLMs. Close-sourced LLM pretraining for general proposes, such as ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), have shown strong medical capacity across both medical

benchmarks and real-world applications. Liévin et al. (2023) utilized GPT-3.5 with different prompting strategies, including Chain-of-thought, few-shot, and retrieval augmentation, for three medical reasoning benchmarks, to show the model's strong medical reasoning ability in the absence of specific fine-tuning. The evaluation of LLMs, such as ChatGPT, on medical exams, including US Medical Exams (Kung et al., 2023) and Otolaryngology–Head and Neck Surgery Certification Examinations (Long et al., 2023), indicates that they achieve scores close to or at the passing threshold. This suggests the potential of LLMs to support real-world medical usages such as medical education and clinical decision-making. Agrawal et al. (2022) views LLMs such as GPT-3 as clinical information extractors and shows the potential in different information extraction tasks. The MedPaLM models (Singhal et al., 2022; 2023), are a series of medical domain-specific LLMs, adapted from PaLM models(**?**Anil et al., 2023; Chowdhery et al., 2022), which have shown performance in answering medical questions on par with that of medical professionals (Singhal et al., 2022; 2023). GPT-4 (OpenAI, 2023) demonstrates strong medical capacities without specialized training strategies in the medical domain or engineering for solving clinical tasks (Nori et al., 2023a;b). When the scope is narrowed down to sub-domain domains, the performance of LLMs is variable. GPT-4 outperforms or performs on par with the current SOTA radiology models (Liu et al., 2023b) across radiology tasks, and it matches human performance on gastroenterology board exam self-assessments (Ali et al., 2023). Peng et al. (2023) examines ChatGPT and GPT-4 performance on Physical Medicine and Rehabilitation, and demonstrates their potential capabilities in the submedical field. However, in dementia, LLMs fail to surpass the performance of traditional AI tools (Wang et al., 2023f). Besides directly utilizing LLMs for different medical tasks, they also can be used as a knowledge base for "retrieving" informative context to auxiliary downstream tasks (Yu et al., 2023d). For example, Zhang et al. (2023h) employs ChatGPT as a medical knowledge base to generate medical knowledge for supporting downstream medical decision-making. Kwon et al. (2024) generates clinical rationales for patient descriptions and utilizes the rationale as an additional training signal to fine-tune student models in both unimodal and multimodal settings to improve diagnosis prediction performance.

Open-Sourced Medical LLMs. Due to privacy concerns (Zhang et al., 2023h) and high costs, several open-source medical LLMs (Xu et al., 2023; Han et al., 2023; Li et al., 2023i; Wu et al., 2023b; Zhang et al., 2023j) have be built by tuning open-source base model, such as LLaMA (Touvron et al., 2023a;b), on medical corpus. These works mainly employ two different strategies: 1) continue pretraining followed by instruction-finetuning (Wu et al., 2023b; Xie et al., 2024a) and 2) direct instruction-finetuning (Xu et al., 2023; Han et al., 2023; Li et al., 2023i; Zhang et al., 2023j; Tran et al., 2023), as shown in 3. Specifically, the first approach involves continuously pretraining language models on biomedical corpora, including medical academic papers and textbooks, and then further fine-tuning the models with various medical instructional datasets to align with human intent for medical applications. The second approach directly conducts instruction fine-tuning on the base models to elicits the medical capabilities of base models directly. In the pretraining and fine-tuning schema, PMC-LLAMA (Wu et al., 2023b) employs a two-step training process that first extends LLaMA's training with millions of medical textbooks and papers, and then instructionally tunes the model on a dataset of 202 million tokens. Me LLaMA (Xie et al., 2024a) proposes a domain-specific base model by continuing to pretrain LLaMA-2 models with 13B and 70B parameters on a 129 billion token medical dataset, and then creates corresponding chat models by conducting instruction fine-tuning on 219k instances. On the other hand, ChatDoctor (Li et al., 2023i) collects 100k real online patient-doctor conversations and directly fine-tunes LLaMA on the dialogues dataset. MedAlpaca (Han et al., 2023) increases the instructional dataset to 230k including question-answer pairs and dialogues and conducts the fine-tuning procedure. Baize-Healthcare (Xu et al., 2023) employs about 100k dialogues from Quora and MedQuAD for instructional tuning by LoRA (Hu et al., 2022). AlpaCare (Zhang et al., 2023j) proposes a 52k diverse machine-generated medical instructional dataset, by distilling medical knowledge from robust closed-sourced LLMs (Li et al., 2022b). Then they fine-tune open-sourced LLMs on the dataset to show the importance of training data diversity on the model's ability to follow medical instructions while maintaining generalizability. BioInstruct (Tran et al., 2023) utilize GPT-4 to generate a 25k instruction dataset covering the topic in question-answering, information extraction and text generation to instruction tune LLaMA models (Touvron et al., 2023a;b) with LoRA (Hu et al., 2022). Their experimental results show consistent improvement in different medical NLP tasks compared to LLMs without instruction fine-tuning.

Multimodal Medical LLMs. Though LLMs have the potential to process and assist clinical NLP tasks, multimodal data (e.g., X-ray Radiology, CT, MRI, ultrasound) plays a vital role in medical and healthcare

Model Name	Model Architecture	Training Corpus	Size
MedAlpaca (Han et al., 2023)	LLaMA	QA rephrased from wikidoc & Flashcards, Stackexchange medical Q&A	7B,13B
ChatDoctor (Li et al., 2023i)	LLaMA	real patient-doctor conversations	7B
Doctor GLM (Xiong et al., 2023)	GLM	medical Dialogues	6.2B
Baize-Healthcare (Xu et al., 2023)	LLaMA	dialogues from Quora and MedQuAD	7B
AlpaCare (Zhang et al., 2023j)	LLaMA	Alpaca data, ChatGPT & GPT-4 generated instruction data	7B, 13B
BioInstruct (Tran et al., 2023)	LLaMA & LLaMA-2	GPT-4 generated instruction data	7B, 13B
Clinical Camel (Toma et al., 2023)	LLaMA-2	Dialogues, articles, Medical QA	13B, 70B
BioMedGPT (Luo et al., 2023)	LLaMA-2	BioMed Articles, PubChemQA, UniProtQA	7B, 10B
Meditron (Chen et al., 2023d)	LLaMA-2	PubMed articles, abstracts, medical guidelines	70B
PMC-LLAMA (Wu et al., 2024)	LLaMA	biomedical academic papers , textbook, medical QA, rationales, dialogues	7B, 13B
Me LLaMA (Xie et al., 2024a)	LLaMA	medical data, medical instruction samples	13B, 70B
BioMistral (Labrak et al., 2024)	Mistral	PubMed Central	7B
LLaVA-Med (Li et al., 2024a)	LLaVA	figure-caption data from PubMed Central GPT-4 generated instruction data	7B, 13B

Table 3: Summary of open-sourced medical LLMs.

applications. When making diagnosis and medical suggestions, it is important for the model to have access to and being capable of understanding clinical modalities beyond text. Hence, there is a strong need to build multi-modal LLMs (Zhu et al., 2022; Liu et al., 2024b; Yan et al., 2023a; 2024a) that are capable of connecting language with other modalities. A representative open-domain model is GPT-4V (Achiam et al., 2023), where researchers have explored its potential for understanding X-rays (Yang et al., 2023b; Wu et al., 2023a). LLaVA-Med (Li et al., 2024a) leverages a figure-caption dataset extracted from PubMed Central, and uses GPT-4 to self-instruct open-ended instruction-following data from the captions to train a visual AI assistant. Gao et al. (2023a) uses a similar recipe to train a multimodal LLM for ophthalmology. Zhang et al. (2023d) trained their model with image-only, text-only, and multi-modal tasks like image captioning and VQA for radiology tasks.

Other than medical chatbots, another important aspect of multimodal medical LLMs is to transform other modalities into language space. To this end, various CLIP-style models (Radford et al., 2021b; Zhang et al., 2022; Wang et al., 2022; Bannur et al., 2023) with two stream visual and text encoder have been proposed. A downstream application is interpretable medical image classification (Yan et al., 2023c), which tries to generate medical concepts with LLMs and concept bottleneck models (Yan et al., 2023b; Echterhoff et al., 2024). This line of work leverages language to explain model decisions while also being able to keep similar or even better classification performance than black-box vision models.

4.3 Abnormality and Ambiguity Detection

Abnormality detection (Harzig et al., 2019) aims to identify abnormal findings in a radiology report by classifying if a sentence reports normal or abnormal conditions. In this task, language models are used to automatically read medical reports and reduce the workload of doctors.

Ambiguity detection was first proposed in (He et al., 2023d), where it tries to detect ambiguous sentences appear in radiology reports that lead to mis-interpretation of reports. Accurate identification of such sentences is crucial, as they impede patients' comprehension of diagnostic decisions and may cause potential treatment delays and irreparable consequences. As a novel task proposed recently, existing LMs may not readily include

such a task into its pre-training stage. Therefore, evaluation of this task allows us to investigate how language models perform for unseen tasks.

Both tasks (He et al., 2023c) are sentence-level classification tasks. For comparison, we measured the classification performance of finetuned LMs (BERT (Devlin et al., 2019), RadBERT (Yan et al., 2022), BioBERT (Lee et al., 2020), ClinicalBERT (Huang et al., 2019), BlueBERT (Peng et al., 2019), BioMed-ReBERTa (Gururangan et al., 2020)) and prompted LLMs (GPT-3, ChatGPT, Vicuna, BioMed LM (Bolton et al., 2024)) by reporting their F1 scores, as shown in Table 4. One can observe that though general LLMs can make reasonable predictions and can improve their performance via few-shot learning, there is still a gap between finetuned LMs and prompted LLMs. Moreover, the novel task of ambiguity detection indeed raises challenges, and there is a need to improve the generalizabitily of LLMs to deal with unseen tasks.

Models		Chest		Miscellaneous Domains	
		Abnormality ↑	Ambiguity ↑	Abnormality↑	Ambiguity ↑
Fine-tuned LMs with task specific data	BERT	0.9791	**0.9893**	0.9607	0.9749
	RadBERT	0.9794	0.9869	0.9640	**0.9813**
	BioBERT	0.9791	0.9862	0.9614	0.9743
	ClinicalBERT	**0.9809**	0.9874	0.9588	0.9736
	BlueBERT	0.9803	0.9867	0.9601	0.9775
	BioMed-ReBERTa	0.9569	0.9758	**0.9776**	0.9788
Prompted LLMs with zero/few shot learning	zero-shot ChatGPT	0.9277	0.6584	0.8880	0.5206
	few-shot ChatGPT	0.9498	0.5831	0.9099	0.5354
	zero-shot GPT-3	0.8762	0.8742	0.8243	0.6448
	few-shot GPT-3	0.9215	0.8320	0.9054	0.6371
	zero-shot Vicuna-7B	0.6987	0.2130	0.7261	0.3739
	few-shot Vicuna-7B	0.8071	<u>0.0785</u>	0.8166	<u>0.2844</u>
	zero-shot BioMed LM	<u>0.6679</u>	0.3485	<u>0.6273</u>	0.3726
	few-shot BioMed LM	0.7905	0.6804	0.7638	0.6804

Table 4: Evaluation (accuracy) over two categories of PLMs on abnormality identification and ambiguity identification tasks (sentence-level NLU). **Bold**: the highest performance. <u>Underlined</u>: the lowest. Results are sourced from (He et al., 2023c).

4.4 Medical Report Generation

Medical report generation (Yan et al., 2021) aims to build models that take medical imaging studies (e.g., X-rays) as input and automatically generate informative medical reports. Unlike conventional image captioning benchmarks (e.g. MS-COCO (Lin et al., 2014)) where referenced captions are usually short, radiology reports are much longer with multiple sentences, which pose higher requirements for information selection, relation extraction, and content ordering. To generate informative text from a radiology image study, a caption model is required to understand the content, identify abnormal positions in an image and organize the wording to describe findings in images.

Evaluation of this task involve two aspects: (1) Automatic metrics for natural language generation: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Denkowski & Lavie, 2011). (2) Clinical Efficiency: CheXpert labeler (Irvin et al., 2019) is used to evaluate the clinical accuracy of the abnormal findings reported by each model, which is a state-of-the-art rule-based chest X-ray report labeling system (Irvin et al., 2019). Given sentences of abnormal findings, CheXpert will give a positive and negative label for 14 diseases. We can then calculate the Precision, Recall and Accuracy for each disease based on the labels obtained from each model's output and from the ground-truth reports.

We report performance of representative clinical models and recent LLMs in Table 5. We consider the following baslines: *ST* (Xu et al., 2015), M^2 *Trans* (Miura et al., 2020), *R2Gen* (Chen et al., 2020), *WCL* (Yan et al., 2021), as well as recent work that uses LLMs: XrayGPT (Thawkar et al., 2023), RaDialog (Pellegrini et al., 2023), Rad-MiniGPT-4 (Liu et al., 2024a). We observe similar trend as the sentence classification tasks, even though LLMs are good at generating fluent text and achieving high NLG scores, domain-specific models can still outperform LLMs in terms of clinical efficacy.

Model	NLG metrics				CE metrics		
	BLEU-1	BLEU-4	METEOR	ROUGE-L	Precision	Recall	F-1
ST	29.9	8.4	12.4	26.3	24.9	20.3	20.4
M^2 Trans	-	11.4	-	-	**50.3**	**65.1**	**56.7**
R2Gen	35.3	10.3	14.2	27.7	33.3	27.3	27.6
WCL	37.3	10.7	14.4	27.4	38.5	27.4	29.4
XrayGPT	-	-	-	20.0	-	-	-
RaDialog	34.6	9.5	14.0	27.1	-	-	39.4
Rad-MiniGPT-4	**40.2**	**12.8**	**17.5**	**29.1**	46.5	48.2	47.3

Table 5: Performance comparison on the test set of MIMIC-CXR with respect to natural language generation (NLG) and clinical efficacy (CE) metrics. Results are reported in percentage (%).

	iCliniq					MedInstruct				
	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG	Text-davinci-003	GPT-3.5-turbo	GPT-4	Claude-2	AVG
Alpaca	38.8	30.4	12.8	15.6	24.4	25.0	20.6	21.5	15.6	22.5
ChatDoctor	25.4	16.7	6.5	9.3	14.5	35.6	18.3	20.4	13.4	18.2
Medalpaca	35.6	24.3	10.1	13.2	20.8	45.1	33.5	34.0	29.2	28.1
PMC	8.3	7.2	6.5	0.2	5.5	5.1	4.5	4.6	0.2	4.6
Baize-H	41.8	36.3	19.2	20.6	29.5	35.1	22.2	22.2	15.6	26.6
AlpaCare	**66.6**	**50.6**	**47.4**	**49.7**	**67.6**	**53.6**	**49.8**	**48.1**	**48.4**	**53.5**

Table 6: **Performance comparison of medical LLMs on medical free-form instruction evaluation.** GPT-3.5-turbo acts as a judge for pairwise auto-evaluation. Each instruction-tuned model is compared with 4 distinct reference models: Text-davinci-003, GPT-3.5-turbo, GPT-4, and Claude-2. 'AVG' denotes the average performance score across all referenced models in each test set. The table is sourced from (Zhang et al., 2023j).

4.5 Medical Free-form Instruction Evaluation

Free-form instruction evaluations assess the practical medical value of language models from a user-centric perspective. This task involves inputting a medical query in a free-text format into the model, which then generates a corresponding response. For instance, if a user inputs, 'Discuss the four major types of leukocytes and their roles in the human immune system in bullet point format,' the model will produce an informed answer based on its internal medical knowledge. This task serves to measure both the medical knowledge capacity and the instruction-following capability of the model. iCliniq (Li et al., 2023i; Chen et al., 2024) contains 10k real online conversations between patients and doctors to evaluate models' medical instruction-following ability in the dialog scenario. MedInstruct-test (Zhang et al., 2023j) contains 217 clinical craft free-form instructions to evaluate the medical capacity and instruction-following ability of models across different medical settings such as treatment recommendation, medical education, disease classification, etc. Evaluating the instruction-following capacity of LLMs is complex due to the wide range of valid responses to a single instruction and the difficulty of replicating human evaluations. Recently, automated evaluation (Zheng et al., 2023; Dubois et al., 2023; Zhang et al., 2023i; Lu et al., 2022) has offered greater scalability and explainability compared to human studies. A strong LLM is used as a judge to compare the outputs of the evaluated model with reference answers and then calculate the winning rate of the evaluated model against the reference answer is used as the evaluation metric. Table 6 shows the current open-source medical LLM performance on medical free-form instruction evaluation with GPT-3.5-turbo acts as a judge and Text-davinci-003, GPT-3.5-turbo, GPT-4, and Claude-2 as reference models, respectively.

4.6 Medical-Imaging Classification Via Natural Language

Medical imaging classification with deep learning has long been studied in the computer vision and clinical community (Li et al., 2014). The task asks a model to take medical imaging (e.g., CT scans) as input, and

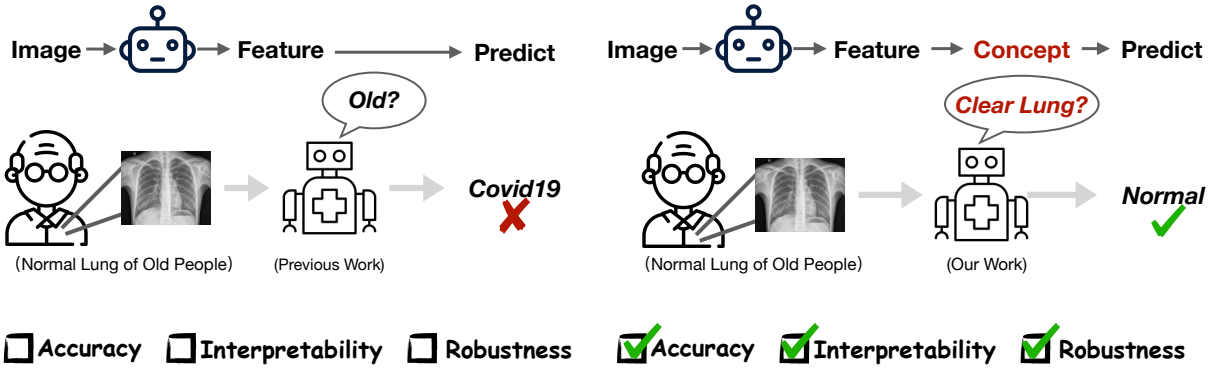

Figure 4: High-level illustration of concept bottleneck models (Yan et al., 2023c). It uses concepts for medical image classification to achieve interpretability and robustness while maintaining accuracy. **Left**: Classification with a classical neural encoder; **Right**: Classification with natural language concepts. A Chest X-ray from a healthy old individual may be classified as Covid-19 due to the patient's age, while introducing language can mitigate the effect of these confounding factors.

assign diagnostic labels to them. However, predicting medical symptoms with "black-box" deep neural models could raise safety and trust issues, as it is hard for human to understand model behaviors and trust model decisions at ease. Clinicians often need to understand the underlying reasoning of the models to carefully make their decisions. Interpretable models allow for better error analysis, bias detection, ensuring patients safety, and trust building. Most recently, the idea of concept bottleneck models (CBMs) (Koh et al., 2020) has been introduced to medical imaging classification, where one can build an intermediate layer by projecting latent image features into a concept space to bring interpretability in the form of natural language.

A follow-up work (Yan et al., 2023c) further shows that classification with concepts not only bring interpretability, but also offers robustness, with the help of pretrained multi-modal LMs (an illustration is presented in Figure 4). This is especially important to medical applications, as confounding factors broadly exist and labeled data are often limited (De Bruijne, 2016). Take the classification of patient X-rays between Covid-19 and normal for instance, certain factors such as the hospitals where the X-rays are performed and the age of the patient strongly correlate with the target disease classification. Yan et al. (2023c) created four diagnostic benchmarks with different confounding factors: age, gender, hospital system. Here we present results on these datasets with comparison for (1) state-of-the-art robust machine learning methods: ERM and Fish (Shi et al., 2021), Lisa (Yao et al., 2022b); (2) linear probing on image features; (3) CBMs with different vision-language models as the backbone: CLIP (Radford et al., 2021b), MedCLIP (Wang et al., 2022), and BioViL (Bannur et al., 2023), shown in Table 7. We find that BioVIL shows promising results when evaluating on challenging datasets with various confounding factors, while another medical VLM, MedCLIP performs similar to general domain CLIP model. To enable useful concept bottleneck models, a strong and robust domain-specific vision-language model is needed.

4.7 Future Prospects

Improving the Capacity of Open-sourced Medical LLMs. Open-source, domain-specific medical LLMs aim to narrow the performance gap between powerful closed-source LLMs and enhance the capability of smaller models to follow various medical instructions and align with user intentions by conducting continuous pretraining and instructional fine-tuning. To further improve the capacity of these models, several future directions can be considered:

- *Data Diversity and Quality:* Although machine-generated datasets accelerate data generation for LLM training, their diversity still lags behind that of real-world collected datasets, which highly impacts model performance (Chiang et al., 2023). Expanding the training datasets to include a broader range of real-world medical texts, such as clinical trial reports, medical journals, patient

Models	NIH-gender	NIH-age	NIH-agemix	Covid-mix	Interpretability
ERM	21.70	3.30	13.80	51.73	✗
Fish	21.70	6.00	17.00	52.16	✗
LISA	23.00	2.30	14.20	51.30	✗
BioViL Image Features	71.60	9.40	13.70	51.08	✗
BioViL Image Features (dropouts)	70.20	19.00	28.60	49.57	✗
CBM w/ CLIP	35.80	15.00	19.10	**64.93**	✓
CBM w/ MedCLIP	42.00	20.10	21.10	51.95	✓
CBM w/ BioViL	**79.60**	**50.70**	**53.40**	62.36	✓

Table 7: Performance comparison for robust medical imaging classification. Results are reported in accuracy (%) and sourced from (Yan et al., 2023c).

records, and health forums, can improve the model's understanding of diverse medical contexts and terminologies. Additionally, ensuring the quality and reliability of the training data is crucial for maintaining the accuracy and trustworthiness of the model's outputs.

- *Retrieval-Augmented Generation:* Integrating retrieval-augmented generation (RAG) techniques can enhance the model's ability to access and incorporate relevant medical knowledge from extensive sources such as large medical knowledge bases, private hospital records, and databases. This approach can improve the model's responses by providing more accurate and contextually appropriate information, particularly in complex medical scenarios in inference time.

- *Addressing Privacy Concerns:* Medical data usage has strong restrictions compared to the general domain and other small domains. Developing methodologies address privacy issues in utilizing LLM APIs and building local LLMs are both important. This can include implementing secure data transmission protocols, ensuring data anonymization, and adopting privacy-preserving techniques such as differential privacy.

Learning in a Data Sparsity Setting. A critical challenge in medical domain for training large-scale models is the restriction of data usage. Data sparsity is a persistent issue due to privacy and confidentiality concerns, the cost of data Acquisition and annotation, as well as ethical considerations. For many practical tasks, e.g., medical report generation, clinical chatbots, medical image classification, the data sparsity issue will be a remaining challenge. Additionally, as mentioned earlier in the performance comparison sections for different applications, task-specific models that trained with in-distribution data and specific architectural design can still outperform foundation models. Based on the empirical finding of training general domain large language models (Brown et al., 2020; Kaplan et al., 2020; OpenAI, 2023), scaling up data is of vital importance for model performance. We discuss some of the potential future directions to address this issue:

- Transfer Learning and Domain Adaptation: For medical LLMs, it is worth exploring how scaling up general domain, publicly-available data can help with in-domain medical tasks. We can explore data selection strategies in pretraining stage to improve the transfer learning performance from general-domain models to medical-specific tasks.

- Synthetic Data Generation: To mitigate the challenges posed by data scarcity, another approach is the generation of synthetic medical data. Leveraging advanced LLMs could enable the creation of diverse synthetic datasets to augment the learning process.

- Few-shot and Zero-shot Learning: Few-shot learning and in-context learning Wei et al. (2022b) methods should be explored more deeply, which has the potential to let medical LLMs adapting to new tasks or domains with minimal training data.

- Privacy-preserving Techniques: Techniques such as differential privacy and federated learning (Rieke et al., 2020) could allow the utilization of patient data for training purposes while ensuring that individual privacy is maintained.

Active Learning: Implementing active learning strategies where the model identifies the most informative data points for labeling can optimize the training process. This method ensures efficient use of scarce data resources and improves learning outcomes in highly specialized medical contexts.

Evaluating Real-world Medical Application Capacity. Although various benchmarks have been proposed in the medical domain, most of them focus on evaluating models from the perspective of medical knowledge (Pal et al., 2022; Jin et al., 2020; 2019), rather than from a user-oriented perspective. To bridge this gap, free-form instruction evaluation datasets utilize medical dialogues and machine-generated text for medical questions. However, these test sets are still limited in quantity and task diversity. For example, Med-Instruct proposes an evaluation dataset with diverse topics, but it is limited to only two hundred tests. On the other hand, iCliniq contains 10,000 instances, but its scope is restricted to doctor-patient conversations. Therefore, there is a need for a large-scale, diverse, and expert-verified dataset for evaluating the medical capacity of LLMs in real-world medical user applications. To evaluate the response quality of LLM in medical free-form instruction evaluation, Zhang et al. (2023j) utilizes the LLM APIs as judges (Zheng et al., 2023; Dubois et al., 2023; Zhang et al., 2023i). However, calling LLM APIs for evaluation is costly. Therefore, training a smaller LLM with strong medical capacity for response evaluation and comparison could be a more efficient alternative. Future work could focus on techniques such as knowledge distillation or model pruning to create such a medical specialized evaluator, potentially leading to faster and more cost-effective evaluation processes for medical LLM applications.

5 Law

NLP is pivotal in the legal domain, providing sophisticated tools for managing the extensive and intricate textual data inherent in legal documentation and proceedings (Harvard Law School Library, 2023; United States Congress, 2023; Fang et al., 2023). The advent of LLMs has further catalyzed innovation at the frontier of legal applications. This section explores the profound influence of LLMs across a range of legal tasks. These technological advancements have strengthened significant enhancements in areas such as legal judgment prediction, legal event detection, legal text classification, and legal document summarization. The purpose of this chapter is to outline the trajectory of LLMs in revolutionizing legal NLP and to shed light on both the challenges faced and the potential future developments. This chapter is organized as follows: In §5.1, we introduce the current NLP tasks in the legal domain, detailing task formulations and relevant datasets. In section §5.2, we explore various PLMs and LLMs developed specifically for legal applications. In section §5.3, we examine the evaluations and performance analysis of LLMs in legal contexts. In section §5.4, we discuss various LLM-based methodologies developed for tackling legal tasks and challenges. Finally, we summarize insights, draw conclusions, and discuss potential future directions in section §5.5.

5.1 Tasks and Datasets in Legal NLP

In this section, we present an array of legal tasks and corresponding datasets that have been investigated through the LLM methodologies. The domains covered include legal question answering (LQA), legal judgment prediction (LJP), legal event detection (LED), legal text classification (LTC), legal document summarization (LDS), and other NLP tasks. Figure 5 provides an overview of these established legal NLP tasks and related datasets.

Legal Question Answering (LQA). LQA is the process of providing answers to legal questions and promotes the development of systems proficient in handling complex inquiries related to laws, regulations, case precedents, and theoretical syntheses. The LQA dataset comprises a wide array of question-and-answer pairs that serve to evaluate a system's capability in legal reasoning. CRJC (Duan et al., 2019), akin to the SQUAD 2.0 (Rajpurkar et al., 2018) format, includes challenges such as span extraction, yes/no questions, and unanswerable questions. Furthermore, professional qualification examinations like the bar exam require specialized legal knowledge and skills, making datasets such as the MBE (Wyner et al., 2016) from the US,

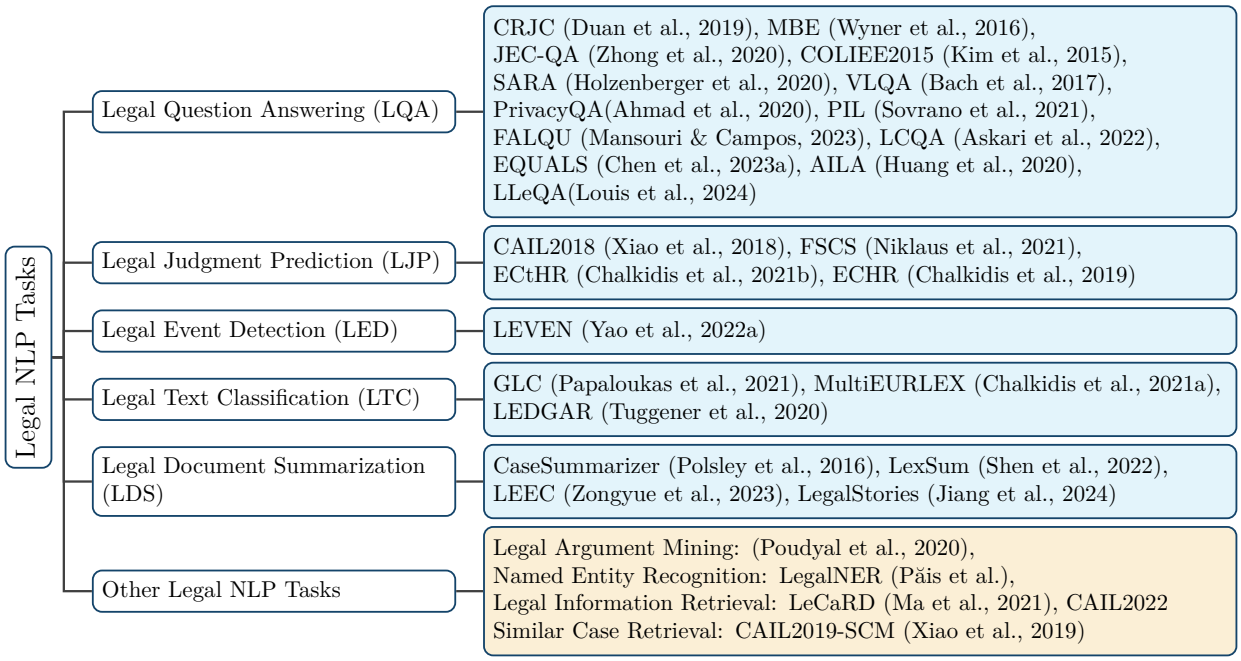

Figure 5: A summarization of existing legal NLP tasks and datasets. The yellow field shows other legal tasks.

JEC-QA (Zhong et al., 2020) from China, and COLIEE2015 (Kim et al., 2015) from Japan particularly demanding. Specific legal domains also have dedicated datasets. For instance, SARA (Holzenberger et al., 2020) focuses on US tax law and includes test cases, while VLQA (Bach et al., 2017) addresses Vietnamese transportation law. In the area of privacy law, PrivacyQA (Ahmad et al., 2020) and PIL (Sovrano et al., 2021) test the system's ability to navigate complex language and regulations regarding data privacy. For the community-oriented legal education, FALQU (Mansouri & Campos, 2023) and LCQA (Askari et al., 2022) are obtained from Law Stack Exchange (law). Several databases employ specific techniques to improve the datasets' quality, for example, EQUALS (Chen et al., 2023a) filters out unqualified legal questions from the raw data. AILA (Huang et al., 2020) integrates domain knowledge from a legal knowledge graph to comprehend and rank question-answer pairs effectively. LLeQA (Louis et al., 2024) provides long-form answers to statutory law questions using a retrieve-then-read pipeline.

Legal Judgment Prediction (LJP). LJP focuses on analyzing legal texts such as case law, statutes, and trial transcripts to predict the outcomes of legal cases. This can assist judges, lawyers, and legal scholars in understanding potential case outcomes based on historical data. The task is generally treated as a classification problem where the input is a legal document and the target is a legal decision (e.g., conviction, acquittal, liability). Researchers have developed several datasets tailored to different legal systems across the globe. For instance, CAIL2018 (Xiao et al., 2018) is a comprehensive Chinese criminal judgment prediction dataset comprising over 2.68 million legal documents published by the Chinese government. Similarly, in Europe, datasets such as FSCS (Niklaus et al., 2021) offer insights into Swiss court judgments with 85,000 cases across two outcomes, reflecting the multilingual nature of the Swiss legal environment. The ECtHR (Chalkidis et al., 2021b) and ECHR (Chalkidis et al., 2019) datasets focus on European Union court judgments, each containing around 11,000 cases but offering a broader scope with 11 potential outcomes.

Legal Event Detection (LED). LED in legal documents involves identifying significant legal proceedings or decisions, such as rulings, motions, or amendments. This task is crucial for enabling legal professionals to monitor pivotal developments within cases efficiently. While Shen et al. (2020) propose hierarchical event features to distinguish similar events in legal texts, and Li et al. (2020b) implement event extraction technologies specifically for the description segments of Chinese legal texts, these studies are constrained by their datasets, which contain only thousands of event mentions. Such limited annotations fail to provide robust training signals or reliable evaluation benchmarks. Addressing this gap, LEVEN (Yao et al., 2022a), a

comprehensive and high-quality dataset, is designed to enhance the capabilities of legal information extraction and LED.

Legal Text Classification (LTC). LTC involves categorizing structured sections within legal documents to enhance their accessibility and comprehensibility. For instance, most legal documents contain sections like "Facts of the case," "Arguments presented by the parties," and "Decisions of the current court," whose identification is crucial for understanding the legal outcomes of cases. These documents can thus be categorized into classes such as facts, argument, and statute, making LTC a multi-class classification task. Key datasets that have propelled advancements in LTC include the following: Greek Legal Code(GLC) (Papaloukas et al., 2021) focuses on categorizing a wide array of Greek legal documents; MultiEURLEX (Chalkidis et al., 2021a) provides a broad collection of EU legislation for classification across multiple languages and jurisdictions; LEDGAR (Tuggener et al., 2020) datasets include large collections of contracts, offering detailed classification based on contract elements and terms.

Legal Document Summarization (LDS). LDS aims at condensing legal documents into succinct summaries while preserving key legal arguments and outcomes. The CaseSummarizer (Polsley et al., 2016) dataset focuses on summarizing case judgments, providing a concise overview of case facts, legal arguments, and judgments. Another dataset, LexSum (Shen et al., 2022), targets the summarization of legislative texts, aiming to extract essential elements and implications for easier comprehension. LEEC (Zongyue et al., 2023) is a comprehensive, large-scale criminal element extraction dataset with 15,831 judicial documents and 159 labels to address the limitations of existing datasets in legal knowledge extraction. LegalStories (Jiang et al., 2024) has 295 complex legal doctrines, each paired with a story and multiple-choice questions generated by LLMs.

Other Legal NLP Tasks. In recent developments, a couple of other tasks have emerged. Among them, legal argument mining (Poudyal et al., 2020) aims to detect and classify arguments within legal texts. Information extraction in the legal domain involves identifying and categorizing key legal entities, such as party names, locations, legal citations, and case facts. LegalNER (Păis et al.) is a dataset for extracting named entities from legal decisions. LeCaRD (Ma et al., 2021) and CAIL2022 datasets (Competition, 2022) enhance criminal case retrieval in Chinese law by linking fact paragraphs to full cases. Another emerging task is similar case retrieval, which aims to identify legal precedents and analogous cases to aid in legal decision-making. The CAIL2019-SCM dataset (Xiao et al., 2019), containing 8,964 triplets of cases published by the Supreme People's Court of China, underscores this task by focusing on the detection of similar cases. These tasks collectively enrich the technological landscape and hold promise for significant enhancements in the efficiency, accessibility, and fairness of legal services.

5.2 Legal LLMs

Since the development of BERT (Devlin et al., 2019), there have been continuous efforts to build PLMs and LLMs specialized for the legal domain. Following the evolving paradigms of general PLMs and LLMs, early legal PLMs adopted the pre-training followed by downstream task fine-tuning paradigm and initially trained relatively small language models. Recent works have scaled up model sizes and introduced instruction fine-tuning, with evaluations covering a broader set of legal tasks. Most existing legal LLMs are text-based, with a focus on Chinese, English, or multi-language support. Table 8 summarizes the PLMs and LLMs for the legal domain.

Pre-Trained and Fine-Tuning PLMs. LegalBERT (Chalkidis et al., 2020) is an early attempt to build a legal PLM targeting tasks like LTC. The model is further pre-trained on a corpus of legal documents and then fine-tuned using task-specific data. Lawformer (Xiao et al., 2021), a transformer-based model, is pre-trained specifically for handling lengthy legal texts, aiding in tasks such as LJP, LRC, and LQA.

Pre-Trained and Fine-Tuning LLMs. Pre-trained and fine-tuning LLMs involve LLMs specifically trained and fine-tuned for legal tasks or datasets. These legal-specific LLMs often integrate external knowledge bases and process extensive initial training to handle a wide range of legal data. Recent developments have led to models like LexiLaw (Haitao, 2024), a fine-tuned Chinese legal model based on the ChatGLM-6B (Group, 2023), meanwhile Fuzi.mingcha (SDU, 2023) is also based on ChatGLM-6B (Group, 2023), which is fine-tuned on CAIL2018 (Xiao et al., 2018) and LaWGPT (Xiao-Song, 2024). Furthermore, WisdomInterrogatory

Model Name	Model Architecture	Main Evaluation Tasks	Languages	Size	Year
Pre-trained and Downstream task Fine-tuning PLMs					
LEGAL-BERT-SMALL (Chalkidis et al., 2020)	BERT (Devlin et al., 2019)	LTC, ST, NER	English	35M	2020
LEGAL-BERT-BASE (SC) (Chalkidis et al., 2020)	BERT (Devlin et al., 2019)	LTC, ST, NER	English	110M	2020
LEGAL-BERT-FP (Chalkidis et al., 2020)	BERT (Devlin et al., 2019)	LTC, ST, NER	English	110M	2020
Lawformer (Xiao et al., 2021)	Longformer (Beltagy et al., 2020)	LJP, LRC, LQA	Chinese	479M	2021
Pre-trained Fine-Tuning LLMs					
JURU (Junior et al., 2024)	Sabiá-2 (Sales Almeida et al., 2024)	LQA	Portuguese	1.9B	2024
LexiLaw (Haitao, 2024)	ChatGLM-6B (Group, 2023)	LRC, LQA	Chinese	6B	2023
Fuzi-Mingcha (SDU, 2023)	ChatGLM-6B (Group, 2023)	LJP, LRC, LQA	Chinese	6B	2023
WisdomInterrogatory (LLM, 2023)	Baichuan-7B (Inc., 2023)	LJP, LRC, LQA	Chinese	7B	2023
LawGPT-7B-beta1.0 (Xiao-Song, 2024)	Chinese-LLaMA-7B (Cui & et al., 2023)	LRC, LQA	Chinese	7B	2023
SaulLM-7B (Colombo et al., 2024)	Mistral-7B (Jiang et al., 2023)	LQA	English	7B	2024
Lawyer-LLaMA (Zhe, 2024)	Chinese-LLaMA-13B (Cui & et al., 2023)	LJP, LRC, LQA	Chinese	13B	2023
ChatLaw-13B (Cui et al., 2023a)	Ziya-LLaMA-13B-v1 (IDEA-CCNL, 2023)	LJP, LRC, LQA	Chinese	13B	2023
ChatLaw-33B (Cui et al., 2023a)	Anima-33B (Ogavinee & et al., 2022)	LJP, LRC, LQA	Chinese	33B	2023

Table 8: Summary of legal PLMs and LLMs. For evaluation tasks, we have **LTC** for Legal Text Classification, **ST** for Sequence Tagging, **NER** for Named Entity Recognition, **LJP** for Legal Judgment Prediction, **SCR** for Similar Case Retrieval, **LRC** for Legal Reading Comprehension, and **LQA** for Legal Question Answering.

Model	JEC-QA	LEVEN	LawGPT	CAIL2018
	Accuracy	F-1	Rouge-L	F-1
Fuzi-Mingcha (SDU, 2023) (zero-shot)	0.08	0.17	0.22	0.25
Fuzi-Mingcha (SDU, 2023) (few-shot)	0.13	0.21	0.33	0.04
ChatLaw-13B (Cui et al., 2023a) (zero-shot)	0.28	0.32	0.31	0.33
ChatLaw-13B (Cui et al., 2023a) (few-shot)	0.29	0.40	0.34	0.26
Wisdom-Interrogatory (LLM, 2023) (zero-shot)	0.15	0.16	0.32	0.33
Wisdom-Interrogatory (LLM, 2023) (few-shot)	0.15	0.16	0.23	0.20
GPT-3.5 (zero-shot)	0.36	0.66	0.34	0.29
GPT-3.5 (few-shot)	0.37	0.68	0.52	0.31
GPT-4 (zero-shot)	0.55	0.79	0.33	0.52
GPT-4 (few-shot)	0.55	0.77	0.57	0.53

Table 9: Performance comparisons for LQA tasks (JEC-QA dataset (Zhong et al., 2020)), LED task (LEVEN dataset (Yao et al., 2022a)), and LJP task (LawGPT dataset (Xiao-Song, 2024) and CAIL2018 (Xiao et al., 2018)). We focus more on LJP tasks based on fact-based articles for the CAIL2018 dataset (Xiao et al., 2018) while scene-based articles for the LawGPT dataset (Xiao-Song, 2024). The few-shot setting is one shot for all datasets.

(LLM, 2023) is a pre-trained and fine-tuning model built upon Baichuan-7B (Inc., 2023). More 7B LLMs like LawGPT-7B-beta1.0 (Nguyen, 2023) are pre-trained on 500k Chinese judgment documents upon Chinese-LLaMA-7B (Cui & et al., 2023), and HanFei (He et al., 2023b) is a fully pre-trained and fine-tuned LLM with 7B parameters. There are more explorations on large-scale LLMs, LaywerLLaM (Zhe, 2023) is based on Chinese-LLaMA-13B (Cui & et al., 2023), fine-tuned with general and legal instructions, additionally, ChatLaw-13B (Cui et al., 2023a) is fine-tuned based on Ziya-LLaMA-13B-v1 (IDEA-CCNL, 2023), and ChatLaw-33B (Cui et al., 2023a) is fine-tuned based on Anima-33B (Ogavinee & et al., 2022). It is worth noting that LLMs based on other languages have also recently emerged, such as SaulLM-7B (Colombo et al., 2024) based on Mistral-7B (Jiang et al., 2023) and JURU (Junior et al., 2024), which is the first LLM pre-trained for the Brazilian legal domain. These legal-specific LLMs, often following an initial pre-training phase, are tailored to specific legal datasets and tasks, enhancing both the precision and applicability of legal NLP technologies in practice.

5.3 Evaluation and Analysis of LLMs

The evaluation and analysis of LLMs' performance is crucial for understanding their effectiveness and capabilities, particularly in legal-specific contexts. This section introduces the evaluation benchmarks in

assessing legal capabilities before the rise of LLMs. Subsequently, we explore specialized legal benchmarks designed explicitly for evaluating the performance of LLMs, and summarize their main findings. These works provide a focused and rigorous assessment of LLMs' abilities in handling legal tasks, offering insights into their efficacy and potential for legal applications.

Before the emergence of LLMs, there were benchmarks used to evaluate NLP models' legal performance. To evaluate model performance uniformly across diverse legal natural language understanding (NLU) tasks, LexGLUE benchmarks (Chalkidis et al., 2021c) are introduced. These benchmarks include datasets like ECtHR (Chalkidis et al., 2021b), SCOTUS (Spaeth et al., 2017), EUR-LEX (Chalkidis et al., 2021a), LEDGAR (Tuggener et al., 2020), UNFAIR-ToS (Lippi et al., 2019), and CaseHOLD (Zheng et al., 2021). They provide a standardized framework for assessing the performance of the language models, allowing for systematic comparison and analysis of different models' capabilities across various legal NLP tasks.

Recently, specialized legal benchmarks designed explicitly for evaluating the performance of LLMs include datasets and tasks that specifically target legal language understanding and reasoning, providing a more nuanced and comprehensive assessment of LLMs' capabilities in legal contexts. LawBench (Fei et al., 2023) is a comprehensive benchmark for evaluating LLMs in the legal domain, assessing their abilities in legal knowledge memorization, understanding, and application across 20 diverse tasks. Extensive evaluations of 51 LLMs, including multilingual, Chinese-oriented, and legal-specific models, reveal GPT-4 as the top performer, indicating the need for further development to achieve more reliable legal-specific LLMs for related tasks. Table 9 summarizes the performance of various methods on JEC-QA dataset (Zhong et al., 2020), LEVEN dataset (Yao et al., 2022a), LawGPT dataset (Xiao-Song, 2024), and CAIL2018 dataset (Xiao et al., 2018). LEGALBENCH (Guha et al., 2023), another legal reasoning benchmark with 162 tasks across six legal reasoning types, created collaboratively by legal professionals. LEGALBENCH (Guha et al., 2023) aims to assess the legal reasoning capabilities of LLMs and facilitate cross-disciplinary dialogue by aligning LEGALBENCH (Guha et al., 2023) tasks with popular legal frameworks. An empirical evaluation of 20 LLMs is presented, showcasing LEGALBENCH (Guha et al., 2023)'s utility in guiding research on LLMs in the legal field. Complementing these, LAiW (Dai et al., 2023) focuses on the logic of legal practice, structuring its evaluation around the process of syllogism in legal logic. LAiW (Dai et al., 2023) divides LLMs capabilities into basic information retrieval, legal foundation inference, and complex legal application, across 14 tasks. The findings from LAiW (Dai et al., 2023) suggest that LLMs show proficiency in generating text for complex legal scenarios, but their performance in basic tasks is still unsatisfying. Additionally, while LLMs may exhibit robust performance, there remains a need to reinforce their ability of legal reasoning and logic.

5.4 LLM-based Methodologies for Legal Tasks and Challenges

This section discusses LLM-based approaches aimed at addressing significant challenges in Legal NLP. These challenges cover multiple aspects, including societal legal problems, legal prediction, document analysis, legal hallucinations, legal exams, and the need for robust LLM Agents.

Societal Legal Challenges. LLMs have emerged as powerful tools with the potential to address various societal challenges in daily life. In the area of legal applications, LLMs are being explored for their capabilities in areas such as tax preparation, online disputes, cryptocurrency cases, and copyright violations. For instance, the use of few-shot in-context learning could improve the performance of LLMs in tax-related tasks (Srinivas et al., 2023; Nay et al., 2024). Moreover, Llmediator (Westermann et al., 2023) highlights the role of LLMs in facilitating online dispute resolution, especially for individuals representing themselves in court, it generates dispute suggestions by detecting the inflammatory message and reformulating polite messages. Additionally, the exploration of LLMs in cryptocurrency security cases (Trozze et al., 2024) (Zhang et al., 2023k) demonstrates their utility in navigating intricate legal landscapes. Addressing copyright violations is another area where LLMs are making an impact (Karamolegkou et al., 2023).

LLM Legal Prediction. Legal prediction judgment is a crucial task of leveraging LLMs in the legal domain. Among the various techniques, Legal Prompt Engineering (LPE) stands out as a commonly used method for enhancing legal predictions. LPE (Trautmann et al., 2022) is a technique that enhances legal responses using key strategies like zero-shot learning, few-shot learning, the chain of reference (CoR), and retrieval-augmented generation. Trautmann et al. (2022) show that zero-shot LPE is better compared to the baselines, but it still

falls short compared to state-of-the-art supervised approaches. Kuppa et al. (2023) propose CoR, where legal questions are pre-prompted with legal frameworks to simplify the tasks into manageable steps, leading to a significant improvement in zero-shot performance by up to 12% in LLMs like GPT-3. Jiang & Yang (2023) introduce legal syllogism prompting (LoT), a simple method to teach LLMs for LJP, focusing on the basic components of legal syllogism: the major premise as law, the minor premise as fact, and the conclusion as judgment.

LLM Document Analysis. LLMs could also assist in legal document analysis, and be applied to case files and legal memos for content extraction. Contract management can be enhanced through automated drafting, review, and risk assessment. LLMs facilitate the mining and analysis of legal cases in case and precedent studies. Steenhuis et al. (2023) outline three methods for automating court form completion: using GPT-3 in a generative AI approach to iteratively prompt user responses, employing a template-driven method with GPT-4-turbo for drafting questions for human review, and a hybrid approach. Choi (2023) discuss using LLMs for legal document analysis, assessing best practices, and exploring the advantages and limits of LLMs in empirical legal research. In a study comparing Supreme Court opinion classifications, GPT-4 matched human coder performance and outperformed older NLP classifiers, without gains from training or specialized prompting.

Legal Hallucination Challenges. With the advent of GPT-4, a surge in research has leveraged this advance to assist in legal decision-making, aiming to offer strategic legal advice and support to lawyers. However, this approach is not without its skeptics. A notable concern is the phenomenon of hallucinations (Zhang et al., 2023l), where the LLMs, despite uncertainties, may suggest decisions in cases when it has less confidence. This highlights a crucial area for further scrutiny, balancing the innovative potential of GPT-4 with the need for reliability and accuracy in sensitive legal contexts. Dahl et al. (2024) investigate the phenomenon of legal hallucinations in LLMs and explore the development of a typology for such hallucinations, the high prevalence of inaccuracies in responses from popular LLMs like GPT-3.5 and Llama 2, the models' inability to correct false legal assumptions, and their lack of awareness when generating incorrect legal information.

LLM Agent Challenges. Developing LLM agents is incredibly challenging due to their specialized design for various legal tasks like providing advice and drafting documents. Their role is crucial in improving legal workflows and efficiency, highlighting the difficulty in their development. For instance, Cheong et al. (2024) examine the implications of using LLMs as public-facing chatbots for providing professional advice, highlighting the ethical, legal, and practical challenges, particularly in the legal domain, and it suggests a case-based expert analysis approach to inform responsible AI design and usage in professional settings. Iu & Wong (2023) sense ChatGPT's potential as a substitute for litigation lawyers, focusing on its drafting abilities for legal documents such as demand letters and pleadings, noting its proficient legal drafting capabilities.

Legal Exam Challenges. Numerous attempts have been made to pass various judicial examinations using LLMs (Choi et al., 2021; Bommarito II & Katz, 2022; Martínez, 2024), GPT-4 passes the bar exam (Katz et al., 2024) but has a long way to go for LexGLUE (Chalkidis et al., 2021c) benchmark (Chalkidis, 2023). Yu et al. (2023a) further their application in legal reasoning by conducting experiments with the COLIEE2019 entailment task (Kano et al., 2019), which is based on the Japanese bar exam.

5.5 Future Prospects

Building High-Quality Legal Datasets. Considering the legal domain's intricate semantics and its requirements for precise statutes, obtaining high-quality legal datasets is often a particularly challenging task. More specifically, most existing legal datasets collected from the natural world are incomplete, sparse, and complicated. Its complexity and scholarly nature make it difficult for regular machine learning approaches to provide annotation, while manual annotation in the legal domain requires much higher demands and costs (e.g., legal training and expertise) than in general domains. For example, CUAD (Hendrycks et al., 2021b) was created with dozens of legal experts from the Atticus project (Contributors, 2024) and consists of over 13,000 annotations. In the future, building high-quality legal datasets may cover the following interesting directions:

- **Multi-Source Legal Data Integration for LLMs.** Real-world legal events often involve data from a multitude of different information sources such as court records, evidence documentation, and multimedia materials (Matoesian & Gilbert, 2018). These pieces of information often exhibit significant diversity, ranging from precise and accurate legal texts to trivial and irrelevant details, and even intentionally obfuscated or ambiguously confused testimonies. Integrating information from diverse sources requires advanced data integration techniques. This requires not only general data processing skills such as multi-modal data fusion but also an understanding of domain-specific nuances such as legal terminology and organizational structures. Additionally, real-world legal case handling often requires global information, especially for long-text legal data. LLeQA (Louis et al., 2024) has made a promising start in providing long-form answers to statutory law questions, paving the way for further research on handling long-text data. This enhanced capability in processing long-form text will be important for addressing complex cases and offering comprehensive legal support. Future research focusing on the recognition of key patterns or legal symbols in long-form text may enhance the model's understanding of lengthy documents.

- **Legal Dataset Collection and Augmentation with LLMs.** Firstly, leveraging the capabilities of LLMs offers promising solutions to simplify the data collection process in the legal domain. More specifically, by harnessing the language processing abilities of LLMs, researchers can automate tasks traditionally requiring extensive knowledge and manual effort, such as legal document annotation and classification. Moreover, LLMs can bridge the knowledge gap between the NLP community and legal experts, enabling efficient extraction of relevant legal information from vast repositories of plain text. This can not only simplify data collection but also empower researchers to navigate complex legal documents with ease, facilitating the generation of high-quality datasets with minimal human work. Furthermore, another intriguing direction involves legal data augmentation in terms of varied formats, the dataset should not only comprise structured court documents but also leverage unstructured text from social media, news, and other sources for enrichment. For the few-shot low-frequency datasets within the LJP tasks, methods integrating data augmentation and feature augmentation are crucial (Wang et al., 2021). These augmentation methods effectively boost dataset diversity, thereby enhancing model performance and robustness.

Developing A Comprehensive LLM-based Legal Assistance System. As aforementioned, while LLMs have made progress in addressing several important legal tasks, their coverage over the legal domain is still far from comprehensive. Looking forward, our long-term objective can be developing practical and systematic legal assistance systems that benefit human life and bring positive social impact. Here, we focus on several specific scenarios where such systems can make a significant impact, including:

- **LLM-based Legal Advice.** As an ongoing and significant area of focus, providing legal advice presents a formidable challenge due to its reliance on legal domain knowledge, cultural context, and intricate logical reasoning. However, the prospects for leveraging LLMs in this domain are promising. LLMs have been trained on vast amounts of data, enabling them to embed cultural backgrounds and common sense into their understanding. Moreover, their capacity for comparing cases across jurisdictions and incorporating human knowledge into inference processes holds the potential for advancing legal reasoning capabilities. Despite the complexities involved, the integration of LLMs in legal reasoning stands to enhance efficiency and accuracy in legal decision-making processes. In this direction, some promising research topics would include: (1) Integrating advanced tools like knowledge graphs (Hogan et al., 2021; Huang et al., 2020) to enhance LLMs' legal domain knowledge and logical reasoning capabilities; (2) Considering the fact of data scarcity in existing legal data, incorporating user feedbacks regarding few-shot learning into LLM-based legal advice is vital. These feedbacks may be collected from users with varying levels of legal expertise, and they can also be used to prevent the dissemination of unethical or inaccurate advice. (3) As LQA in legal scenarios requires complex logical reasoning, it naturally leads to future research directions of enhancing the parameters scale for legal-specific LLMs. On the other hand, to improve the efficiency of real-time LQA, research on legal LLM compression would be important in real-world applications.

- **LLM-based Legal Explanation and Analysis.** Existing LLM-based methods often operate like black boxes, thus legal case explanation and analysis represent another critical task in the legal domain. This includes both the examination of real-world court cases and the decisions made by LLMs. In terms of LLM-generated decisions, an intriguing path of research involves developing mechanisms for self-explanation akin to chain-of-thought (Wei et al., 2022b) approaches. For real-world legal cases, LLMs can offer context-specific explanations tied to real-world scenarios when analyzing them. Moreover, LLMs have the potential to provide various types of human-understandable explanations, spanning general legal regulations, example-specific discussions, and comparisons between analogous cases—potentially transcending state, time, and country boundaries. Such multifaceted explanations can enhance the trustworthiness and transparency in the legal domain, mitigating unfairness and ethical issues like gender bias in legal systems (Sevim et al., 2023), reducing legal hallucination significantly (Dahl et al., 2024).

- **Social Impact of LLM-based Legal System.** Investigating the social impact of LLMs on the legal domain also includes many interesting directions: (1) Applying LLMs to democratize legal education and advice, benefiting individuals who have difficulties in visiting a human lawyer due to the lack of professional knowledge or economic resources. This democratization can empower marginalized communities by granting them access to crucial legal information and guidance; (2) The development of LLMs will also accelerate the evolution of privatized and personalized legal LLMs, leading to increasing competition in the legal domain and the creation of more satisfying products for customers (Cui et al., 2023a); (3) Leveraging LLMs to drive future developments in law through enhanced legal analysis. By facilitating deeper insights into legal texts and precedents, LLMs can contribute to more informed law updates and academic research endeavors; (4) Addressing ethical issues with LLMs in the legal system. Through rigorous analysis and scrutiny, LLMs can help identify and rectify instances of injustice and discrimination against certain demographic groups in legal decision-making processes.

6 Ethics

Despite recent breakthroughs in LLMs, concerns regarding their ethics and trust have been raised for their real-world use (Kaddour et al., 2023; Ray, 2023). Especially, when applied in high-stakes domains such as finance, healthcare, and law, these ethical concerns become particularly critical. In a broader sense, the ethics of AI technologies in different domains have been widely discussed in the last few decades (Jobin et al., 2019; Leslie, 2019). Despite these large number of existing discussions, numerous ethical concepts are proposed from diverse disciplines and perspectives with complicated objectives, leading to challenges to constructing a consistent and well-organized ethical framework. Fortunately, these various ethical considerations often originate from similar high-level principles. In this section, we first introduce several general ethical principles and related considerations for LLM applications, and also showcase examples of these ethics in domain-specific contexts. We will describe the basic definitions of these ethical issues and summarize the existing investigations for testing or addressing them in section 6.1, then we discuss future directions in section 6.2.

6.1 Ethical Principles and Considerations

In the discourse on AI ethics, there are several general principles that find broad adoption. Guided by these principles, a multitude of nuanced definitions are elaborated across various domains. For example, an investigation (Jobin et al., 2019) on 84 AI ethical documents summarizes 11 frequent ethical principles and guidelines, including transparency, justice, responsibility, non-maleficence, privacy, beneficence, autonomy, trust, sustainability, dignity, and solidarity. Here, combining the most urgent ethical concerns regarding LLMs and the varying focuses across finance, healthcare, and law, we mainly highlight three **ethical principles** (*Transparency, Justice, Non-maleficence*) and several most prevalent **ethical considerations** (*Explainability, Bias&Fairness, Robustness, Hallucination*) that are associated with these principles. Figure 6 shows the connection between these ethical principles and considerations, as well as some example tasks that prioritize these ethical principles in different domains.

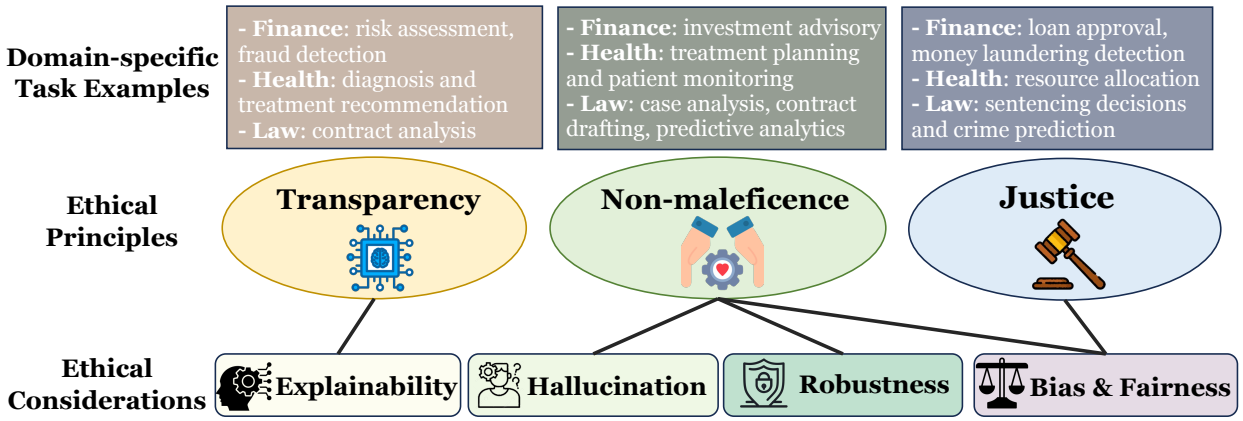

Figure 6: Important ethical principles, considerations, and domain-specific examples in finance, healthcare, and law.

6.1.1 Ethical Principles

Transparency. Transparency refers to "explaining and understanding" the systems, including different stages such as data usage and model behavior. Transparency is the most frequently mentioned AI ethical principle based on the investigation in Jobin et al. (2019). Many concepts are related to transparency, such as explainability, interpretability, communication, and accountability. Transparency is particularly critical when LLM assists in complicated, expertise-intensive, and high-risk applications. In finance, institutions have started to utilize LLMs for tasks such as risk assessment, fraud detection, and automated trading strategies; In healthcare, LLMs are increasingly employed in clinical decision support, such as disease diagnosis and treatment recommendation; In law, LLMs have been used for contract review and analysis. In these examples of applications, transparency is crucial to promote understanding of how LLMs make final decisions and thereby assess their potential risks or issues.

Justice. Justice encompasses a spectrum of meanings, commonly associated with “fairness, equity, inclusion, diversity, non-bias, and non-discrimination". Hence, justice holds particular significance when LLMs are utilized in contexts involving individuals from diverse demographic or societal backgrounds. In studies of law, justice is often widely referenced as a core legal principle. For instance, when using LLMs to assist in sentencing decisions or crime prediction, justice (e.g., anti-discrimination against race, economic/political status, and crime history) is crucial. In finance, applications such as loan approval, money laundering detection, and consumer rights protection have high demands for fairness and equity. In healthcare, fair resource allocation and treatment recommendations without inequalities and discrimination are of great importance when LLMs are applied.

Non-maleficence. Non-maleficence generally means "do no harm". Here, harms can exist in a variety of forms such as incorrect, toxic, outdated, biased, and privacy-violating information. As LLMs are often trained on vast corpora with unknown quality, the removal of such harmful information is imperative in practical applications. In finance, non-maleficence is important because it emphasizes the responsibility of financial institutions, professionals, and regulators to prevent harm to investors, consumers, and the broader financial system, avoiding potential financial losses or harm to individuals or society. In healthcare, non-maleficence plays a key role in maintaining patient safety, trust, and confidentiality, especially in tasks like treatment planning and patient monitoring. In law, non-maleficence is crucial to prevent wrong actions, negligence, and violations of legal rights stemming from reliance on obsolete or inaccurate legal provisions. Especially, due to the dynamic nature of legal systems, updated law may inadvertently leave behind outdated and harmful information, necessitating careful approaches to ensure that legal practices and interpretations remain aligned with current statutes and regulations.

6.1.2 Ethical Considerations

Explainability. Explainability means the capacity to elucidate the model behavior in a human-understandable way (e.g., showing the importance of input data or model component for model output, and estimating the model behavior in interventional or counterfactual cases). AI explainability has been a longstanding concern (Saeed & Omlin, 2023; Došilović et al., 2018), as many AI models inherently function as black boxes, lacking transparency and interoperability. Especially, explanation in LLMs is often even more challenging than most traditional AI techniques due to the extensive scale of training data and the large size of the model. Despite the challenges, from another aspect, the unique ability of LLMs to comprehend and generate natural language empowers them to elucidate their own decision-making processes. Recent investigations (Zhao et al., 2024; Singh et al., 2024) have summarized existing explanation approaches for LLMs in both traditional fine-tuning paradigm (with approaches such as feature-based explanation (Ribeiro et al., 2016; Lundberg & Lee, 2017) or example-based explanation (Koh & Liang, 2017; Verma et al., 2020)), and recent prompting-based paradigm (with approaches such as in-context learning explanation (Li et al., 2023j) and chain-of-thought (CoT) prompting explanation (Wu et al., 2023e)).

Bias & Fairness. Bias and fairness broadly include various ethical terms, such as social stereotypes or discrimination against certain demographic groups related to sensitive features (e.g., race, gender, or disability) (Gallegos et al., 2023; Ghosh & Caliskan, 2023) and monolingual bias (Talat et al., 2022) for the languages different from training data. Uncensored natural language often contains numerous biases, and the culture, language, and demographic information included in training corpora are often highly imbalanced, which are the main causes for unfair language models. Besides, improper model selection and learning paradigms can also lead to biased outcomes. Existing work (Gallegos et al., 2023; Kotek et al., 2023; Zhuo et al., 2023; Ghosh & Caliskan, 2023; McGee, 2023; Motoki et al., 2023) has discussed and evaluated LLMs in terms of bias in different cases, showing that LLMs have a certain ability to resist social discrimination in open-ended dialogue, but still often exhibit varying forms of bias. With such issues, many efforts have been made to mitigate bias in LLMs (Li et al., 2023e; Ferrara, 2023; Gallegos et al., 2023), covering both perspectives from data-related bias or model-related bias. Current debiasing methods mainly include (a) *Pre-processing* approaches which mitigate bias in LLM by changing the model training data, such as data augmentation and generation (Xie & Lukasiewicz, 2023; Stahl et al., 2022), as well as data calibration (Ngo et al., 2021; Thakur et al., 2023; Amrhein et al., 2023) and reweighting (Han et al., 2021; Orgad & Belinkov, 2023); (b) *In-processing* approaches that debiase by changing LLM models. Multiple technologies such as contrastive learning (He et al., 2022; Li et al., 2023f), model retraining (Qian et al., 2022), and alignment (Guo et al., 2022; Ahn & Oh, 2021) have been adopted in these studies; (c) *Post-processing* methods that mitigate bias from model outputs (Liang et al., 2020; Lauscher et al., 2021; Dhingra et al., 2023).

Robustness. Although its definition varies in different contexts, robustness generally denotes a model's capacity to sustain its performance even for input that deviates from the training data. The deviation can be triggered by different factors such as cross-domain distributions and adversarial attacks. Models lacking robustness often result in a cascade of adverse consequences, e.g., privacy leakage (Carlini et al., 2021), model vulnerability (Michel et al., 2022), and generalization issues (Yuan et al., 2024). Various attacks (Zou et al., 2023; Lapid et al., 2023; Liu et al., 2023e; Wei et al., 2024; Shen et al., 2023b; Zhuo et al., 2023; Shi et al., 2024) for LLMs are emerging continuously. Some of them inherit commonly used attacking strategies in traditional domains such as computer vision (Szegedy et al., 2013; Biggio et al., 2013). Others explore "jailbreaks" (Liu et al., 2023e; Wei et al., 2024; Deng et al., 2023a), aiming to strategically craft prompts (with human effort or automatic generation) to result in outputs deviating from the purpose of aligned LLMs. Further studies also focus on the universality and transferability of attacks (Zou et al., 2023; Lapid et al., 2023). These studies found that numerous vulnerabilities and deficiencies in LLMs persist, prompting severe societal and ethical issues. Simultaneously, there has been a surge in literature (Yuan et al., 2024; Altinisik et al., 2022; Stolfo et al., 2022; Moradi & Samwald, 2021; Shi et al., 2024; Ye et al., 2023b; Mozes et al., 2023; Wang et al., 2023b; Schwinn et al., 2023; Jain et al., 2023; Kumar et al., 2023) dedicated to researching and evaluating the LLM robustness. Previous work (Wang et al., 2023e) categorizes existing methods against jailbreak attacks on LLMs into two directions: internal safety training (Ganguli et al., 2022; Touvron et al., 2023b) (i.e., further train the LLM model with adversarial examples to better identify attacks) and external safeguards (Jain et al., 2023; Markov et al., 2023) (i.e., incorporate external models or filters to replace

harmful queries with predefined warnings). SELF-GUARD (Wang et al., 2023e) combines these two types of safety methods. It is also worth mentioning that self-evaluation for LLM outputs (Helbling et al., 2023; Li et al., 2023h) has become an emerging trend in defense strategies.

Hallucination. Hallucination has been a ubiquitous issue in NLP, which refers to the generation of incorrect, nonsensical, or misleading information. Especially, hallucination in LLMs faces unique challenges due to its difference from traditional language models (Ji et al., 2023). The authenticity and precision of pursuits in finance, healthcare, and law inevitably underscore the pressing concern of hallucination (Alkaissi & McFarlane, 2023). Mainstream work (Ji et al., 2023; Maynez et al., 2020; Kaddour et al., 2023) categorizes hallucination into two types: (1) *Intrinsic hallucination*: the generated output conflicts with the source content (e.g., the prompt); and (2) *Extrinsic hallucination*: the correctness of the generated output cannot be verified based on the source content. Although extrinsic hallucination is not always incorrect, and sometimes can even provide useful background information (Maynez et al., 2020), we should still handle any unverified information cautiously. Current work (Min et al., 2023b;b; Ren et al., 2023) has identified and evaluated hallucination in different ways, including those based on external verified knowledge such as Wikipedia (e.g., Kola (Yu et al., 2023b), FActScore (Min et al., 2023b), FactualityPrompts (Lee et al., 2022)), as well as those based on probabilistic metrics such as uncertainty of LLM generation (Manakul et al., 2023; Varshney et al., 2023). Many factors can result in hallucination, such as biases inside data, outdated corpora, prompt strategy, and intrinsic model limitations. Correspondingly, as discussed in Ji et al. (2023), existing approaches for LLM hallucination limitation cover different branches including data-centric and model-centric. Data-centric methods eliminate hallucination by improving the data quality in different stages (Zhang et al., 2023f; Penedo et al., 2023; Es et al., 2023). Model-centric approaches focus on the model design and their training or tuning procedure. Representative methods in this line include reinforcement learning with human feedback (RLHF) (Ouyang et al., 2022), model editing (Daheim et al., 2023), and decoding strategies (Dziri et al., 2021; Tian et al., 2019).

6.1.3 Domain-specific Ethics

On top of general ethical principles and considerations, in the specific context of different domains, the definitions of ethics display their distinct focus and subtle differences. Here, we introduce domain-specific investigations of ethics in finance, health, and law, respectively.

Finance. Many ethical guidelines (Attard-Frost et al., 2023; Svetlova, 2022; Kurshan et al., 2021; Farina et al., 2024) for AI practice in the finance sector have been published in recent years. In Attard-Frost et al. (2023), based on the general Fairness, Accountability, Sustainability, and Transparency (FAST) AI ethical principles in the public sector proposed by Leslie (2019), a series of business-oriented ethical themes (e.g., market fairness, bias & diversity in professional practice, and business model transparency) are organized under each principle. When using LLMs for finance, a few studies have begun to discuss the ethics of LLMs such as ChatGPT (Khan & Umer, 2024) and BloombergGPT (Wu et al., 2023d). Exploratory efforts for addressing ethical issues of LLMs in finance, such as hallucination (Kang & Liu, 2024; Roychowdhury et al., 2023) and financial crime (Ji et al., 2024), have laid a promising groundwork for further investigation.

Healthcare. Ethics in healthcare has long garnered significant attention (Pressman et al., 2024; Beauchamp & Childress, 2001) due to potentially severe and irreversible consequences, notably the loss of human life. As a result, a set of widely adopted ethical principles (Autonomy, Beneficence, Non-maleficence, and Justice) (Beauchamp & Childress, 2001) has been established in clinical and medical practice. Apart from the aforementioned non-maleficence and justice, autonomy in health centers on an individual's right to make informed medical decisions, and beneficence in health focuses on "doing good" to promote patient well-being. Recent discussions (Li et al., 2023b; Karabacak & Margetis, 2023; Minssen et al., 2023; Yu et al., 2023c; Thirunavukarasu et al., 2023; Haltaufderheide & Ranisch, 2024; Ullah et al., 2024) for the ethics of LLMs in the health & medicine sector have reached the consensus that existing LLMs still have a substantial gap to bridge in order to meet ideal ethical standards. This situation leads to the development of more nuanced ethical considerations across various healthcare scenarios. For instance, a recent review (Haltaufderheide & Ranisch, 2024) summarized LLM ethics in four key clinical themes, including clinical applications, patient support, health professionals, and public health. Other discussions about ethics in specific healthcare contexts

such as surgery (Pressman et al., 2024) and mental health (Cabrera et al., 2023) also provide valuable insight into LLM applications in real-world health systems.

Law. In the domain of law, numerous deliberations (Cranston, 1995; Yamane, 2020; Wright, 2020; Nunez, 2017) has taken place about legal ethics for AI. The recent progress of LLMs brings new challenges and discussions about ethics in the legal domain, stimulating the refinement of existing legal ethics and the development of more feasible evaluation standards. Among these works, Zhang et al. (2024a) design a multi-level ethical evaluation framework and evaluates mainstream LLMs under the framework. This evaluation framework covers three aspects with increasing level of ethical proficiency: legal instruction following (i.e., the ability of LLMs to address user needs based on given instructions), legal knowledge (i.e., the ability of LLMs to distinguish the legal/nonlegal elements), and legal robustness (i.e., the consistency of LLM responses to identical questions presented in varying formats and contexts). Another recent work (Cheong et al., 2024) collects opinions from 20 legal experts, revealing detailed policy considerations for LLM employment in the professional legal domain. Moreover, a few exploratory works in more specific tasks such as profiling legal hallucinations also start to attract people's attention (Dahl et al., 2024). These studies set the stage for more comprehensive ethics regulations in future "LLM + Law" applications.

6.2 Future Prospects

For future work addressing ethical concerns in LLMs, a multifaceted strategy is imperative. (1) **Dataset censorship:** Meticulous dataset censorship is vital, involving a thorough examination and elimination of improper content (e.g., biased or incorrect information) from the training data. This step ensures that the model is shielded from potentially harmful information, minimizing the risk of encoding unwanted patterns. (2) **Human and domain knowledge:** The integration of humans in the AI loop is essential. Human reviewers contribute nuanced perspectives, provide domain knowledge, identify ethical issues, and guide the model's learning process by refining its responses. Human-in-the-loop systems allow for ongoing monitoring and adjustments to address emerging ethical problems. (3) **Theoretical bounds:** Establishing theoretical bounds on the model's behavior is important. The development of clear theoretical frameworks and ethical guidelines helps delineate the limits of the model's decision-making, preventing it from generating potentially harmful or biased outputs. Through the implementation of these measures, we can elevate the ethical standards of LLMs, fostering responsible AI development. (4) **Explanation and causality:** It is essential to delve into the underlying causes and mechanisms behind the generation of LLM outputs. Understanding the root causes of ethical issues is beneficial for developing effective mitigation strategies.

7 Conclusion

The exploration of LLMs across diverse fields illuminates the vast potential and inherent challenges of integrating advanced AI tools into various real-world applications. This survey focuses on three critical societal domains: finance, healthcare & medicine, and law, underscoring the transformative impact of LLMs in enhancing research methodologies and accelerating the pace of knowledge discovery and decision-making in these domains. Through detailed examination across disciplines, we highlight significant advancements achieved by leveraging LLMs in these domains, foreseeing a promising future full of breakthroughs and opportunities.

However, the integration of LLMs also brings to light challenges and ethical considerations. Concerns such as explainability, bias & fairness, robustness, and hallucination necessitate ongoing scrutiny and development of mitigation strategies. Furthermore, the interdisciplinary nature of LLM applications calls for collaborative efforts among AI researchers, domain experts, and policymakers to navigate the ethical landscape and harness the full potential of LLMs responsibly. As LLMs continue to evolve and find broader utility, it becomes increasingly imperative to address these challenges systematically and proactively.

References

Law stack exchange. URL https://law.stackexchange.com. Accessed on 30 April 2024.

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.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130.

Toyin Aguda, Suchetha Siddagangappa, Elena Kochkina, Simerjot Kaur, Dongsheng Wang, Charese Smiley, and Sameena Shah. Large language models as financial data annotators: A study on effectiveness and efficiency, 2024.

Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. Policyqa: A reading comprehension dataset for privacy policies. *arXiv preprint arXiv:2010.02557*, 2020.

Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. *arXiv preprint arXiv:2109.05704*, 2021.

Raghad Al-Shabandar, Gaye Lightbody, Fiona Browne, Jun Liu, Haiying Wang, and Huiru Zheng. The application of artificial intelligence in financial compliance management. In *Proceedings of the 2019 International Conference on Artificial Intelligence and Advanced Manufacturing*, pp. 1–6, 2019.

Shuhaib Ali, Omer Shahab, Reem Al Shabeeb, Farah Ladak, Jamie O. Yang, Girish Nadkarni, Juan José Solozábal Echavarría, Sumbal Babar, Aasma Shaukat, Ali Soroush, and Bara El Kurdi. General purpose large language models match human performance on gastroenterology board exam self-assessments. In *medRxiv*, 2023. URL https://api.semanticscholar.org/CorpusID:262323763.

Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2), 2023.

Enes Altinisik, Hassan Sajjad, Husrev Taha Sencar, Safa Messaoud, and Sanjay Chawla. Impact of adversarial training on robustness and generalizability of language models. *arXiv preprint arXiv:2211.05523*, 2022.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. Domain adaption of named entity recognition to support credit risk assessment. In Ben Hachey and Kellie Webster (eds.), *Proceedings of the Australasian Language Technology Association Workshop, ALTA 2015, Parramatta, Australia, December 8 - 9, 2015*, pp. 84–90. ACL, 2015. URL https://aclanthology.org/U15-1010/.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. Exploiting biased models to de-bias text: A gender-fair rewriting model. *arXiv preprint arXiv:2305.11140*, 2023.

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.

Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063, 2019. URL http://arxiv.org/abs/1908.10063.

Arian Askari, Suzan Verberne, and Gabriella Pasi. Expert finding in legal community question answering. In *European Conference on Information Retrieval*, pp. 22–30. Springer, 2022.

Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the First ACM International Conference on AI in Finance*, pp. 1–8, 2020.

Blair Attard-Frost, Andrés De los Ríos, and Deneille R Walters. The ethics of ai business practices: a review of 47 ai ethics guidelines. *AI and Ethics*, 3(2):389–406, 2023.

Saqib Aziz and Michael Dowling. *Machine learning and AI for risk management*. Springer International Publishing, 2019.

Ngo Xuan Bach, Tran Ha Ngoc Thien, Tu Minh Phuong, et al. Question analysis for vietnamese legal question answering. In *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 154–159. IEEE, 2017.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Tom L Beauchamp and James F Childress. *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Simon Benninga. *Financial modeling*. MIT press, 2014.

Ankita Bhatia, Arti Chandani, Rizwana Atiq, Mita Mehta, and Rajiv Divekar. Artificial intelligence in financial services: a qualitative research to discover robo-advisory services. *Qualitative Research in Financial Markets*, 13(5):632–654, 2021.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. Fintral: A family of GPT-4 level multimodal financial large language models. *CoRR*, abs/2402.10986, 2024. doi: 10.48550/ARXIV.2402.10986. URL `https://doi.org/10.48550/arXiv.2402.10986`.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pp. 387–402. Springer, 2013.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

Michael Bommarito II and Daniel Martin Katz. Gpt takes the bar exam. *arXiv preprint arXiv:2212.14402*, 2022.

Petros Boulieris, John Pavlopoulos, Alexandros Xenos, and Vasilis Vassalos. Fraud detection with natural language processing. *Machine Learning*, pp. 1–22, 2023.

David G Brauer and Kristi J Ferguson. The integrated curriculum in medical education: Amee guide no. 96. *Medical teacher*, 37(4):312–322, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

Johana Cabrera, M Soledad Loyola, Irene Magaña, and Rodrigo Rojas. Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pp. 313–326. Springer, 2023.

Aaron Calafato, Christian Colombo, and Gordon J Pace. A controlled natural language for tax fraud detection. In *Controlled Natural Language: 5th International Workshop, CNL 2016, Aberdeen, UK, July 25-27, 2016, Proceedings 5*, pp. 1–12. Springer, 2016.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams, 2023.

Longbing Cao. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*, 55(3):1–38, 2022.

Yi Cao and Jia Zhai. A survey of ai in finance. *Journal of Chinese Economic and Business Studies*, 20(2): 125–137, 2022.

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)*, pp. 2633–2650, 2021.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

Ilias Chalkidis. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *arXiv preprint arXiv:2304.12202*, 2023.

Ilias Chalkidis and Dimitrios Kampas. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198, 2019.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*, 2019.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*, 2021a.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. *arXiv preprint arXiv:2103.13084*, 2021b.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*, 2021c.

Ching Chang, Wei-Yao Wang, Wen-Chih Peng, and Tien-Fu Chen. Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters, 2024.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Michael Chary, Saumil Parikh, Alex F Manini, Edward W Boyer, and Michael Radeos. A review of natural language processing in medical education. *Western Journal of Emergency Medicine*, 20(1):78, 2019.

Erwin Chemerinsky. *Constitutional law*. Aspen Publishing, 2023.

Andong Chen, Feng Yao, Xinyan Zhao, Yating Zhang, Changlong Sun, Yun Liu, and Weixing Shen. Equals: A real-world dataset for legal question answering via reading chinese laws. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 71–80, 2023a.

Lulu Chen, Yingzhou Lu, Chiung-Ting Wu, Robert Clarke, Guoqiang Yu, Jennifer E Van Eyk, David M Herrington, and Yue Wang. Data-driven detection of subtype-specific differentially expressed genes. *Scientific reports*, 11(1):332, 2021a.

Tianyi Chen, Nan Hao, Yingzhou Lu, and Capucine Van Rechem. Uncertainty quantification on clinical trial outcome prediction. *arXiv preprint arXiv:2401.03482*, 2024.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *CoRR*, abs/2310.15205, 2023b. doi: 10.48550/ARXIV.2310.15205. URL https://doi.org/10.48550/arXiv.2310.15205.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023c.

Yuh-Jen Chen, Chun-Han Wu, Yuh-Min Chen, Hsin-Ying Li, and Huei-Kuen Chen. Enhancement of fraud detection for narratives in annual reports. *International Journal of Accounting Information Systems*, 26: 32–45, 2017.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023d.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R. Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 3697–3711. Association for Computational Linguistics, 2021b. doi: 10.18653/V1/2021.EMNLP-MAIN.300. URL https://doi.org/10.18653/v1/2021.emnlp-main.300.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 6279–6292. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.421. URL https://doi.org/10.18653/v1/2022.emnlp-main.421.

Xueqi Cheng, Shenghua Liu, Xiaoqian Sun, Zidong Wang, Houquan Zhou, Yu Shao, and Huawei Shen. Combating emerging financial risks in the big data era: A perspective review. *Fundamental Research*, 1(5): 595–606, 2021.

Inyoung Cheong, King Xia, KJ Feng, Quan Ze Chen, and Amy X Zhang. (a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. *arXiv preprint arXiv:2402.01864*, 2024.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Jonathan H Choi. How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics (Forthcoming)*, 2023.

Jonathan H Choi, Kristin E Hickman, Amy B Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *J. Legal Educ.*, 71:387, 2021.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=r1xMH1BtvB`.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*, 2024.

CAIL AI Legal Competition. Cail ai legal competition. `https://github.com/china-ai-law-challenge/CAIL2022`, 2022.

The Atticus Project Contributors. The Atticus Project: Open-source tools for forensic analysis, 2024. URL `https://github.com/TheAtticusProject`. GitHub repository.

Patricia Craja, Alisa Kim, and Stefan Lessmann. Deep learning for detecting financial statement fraud. *Decision Support Systems*, 139:113421, 2020.

Ross Cranston. Legal ethics and professional responsibility. 1995.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*, 2023a.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*, 2023b.

Yiming Cui and et al. Chinese-llama-alpaca-2: A chinese large language model, 2023. URL `https://github.com/ymcui/Chinese-LLaMA-Alpaca-2/`.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*, 2023.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *arXiv preprint arXiv:2401.01301*, 2024.

Yongfu Dai, Duanyu Feng, Jimin Huang, Haochen Jia, Qianqian Xie, Yifang Zhang, Weiguang Han, Wei Tian, and Hao Wang. Laiw: A chinese legal large language models benchmark (a technical report). *arXiv preprint arXiv:2310.05620*, 2023.

Min-Yuh Day, Jian-Ting Lin, and Yuan-Chih Chen. Artificial intelligence for conversational robo-advisor. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1057–1064. IEEE, 2018.

Marleen De Bruijne. Machine learning approaches in medical image analysis: From detection to diagnosis, 2016.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Jailbreaker: Automated jailbreak across multiple large language model chatbots. *arXiv preprint arXiv:2307.08715*, 2023a.

Xiang Deng, Vasilisa Bashlovkina, Feng Han, Simon Baumgartner, and Michael Bendersky. What do llms know about financial markets? a case study on reddit market sentiment analysis. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pp. 107–110, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9781450394192. doi: 10.1145/3543873.3587324. URL https://doi.org/10.1145/3543873.3587324.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. PACIFIC: towards proactive conversational question answering over tabular and textual data in finance. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 6970–6984. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.469. URL https://doi.org/10.18653/v1/2022.emnlp-main.469.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pp. 85–91, 2011.

Franck Dernoncourt and Ji Young Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019.*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*, 2023.

João Dias, Pedro A Santos, Nuno Cordeiro, Ana Antunes, Bruno Martins, Jorge Baptista, and Carlos Gonçalves. State of the art in artificial intelligence applied to the legal domain. *arXiv preprint arXiv:2204.07047*, 2022.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Using structured events to predict stock price movement: An empirical investigation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1415–1425. ACL, 2014. doi: 10.3115/V1/D14-1148. URL https://doi.org/10.3115/v1/d14-1148.

Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215. IEEE, 2018.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pp. 439–451. Springer, 2019.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

Ronald Dworkin. *Law's empire*. Harvard University Press, 1986.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*, 2021.

William Easterly and Sergio Rebelo. Fiscal policy and economic growth. *Journal of monetary economics*, 32 (3):417–458, 1993.

Jessica Echterhoff, An Yan, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. Driving through the concept gridlock: Unraveling explainability bottlenecks in automated driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7346–7355, 2024.

Jan Egger, Christina Gsaxner, Antonio Pepe, Kelsey L Pomykala, Frederic Jonske, Manuel Kurz, Jianning Li, and Jens Kleesiek. Medical deep learning—a systematic meta-review. *Computer methods and programs in biomedicine*, 221:106874, 2022.

Richard A Epstein and Catherine M Sharkey. *Cases and materials on torts*. Aspen Publishing, 2020.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang. Large language models for software engineering: Survey and open problems. *arXiv preprint arXiv:2310.03533*, 2023.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Super-scotus: A multi-sourced dataset for the supreme court of the us. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pp. 202–214, 2023.

Mirko Farina, Xiao Yu, and Andrea Lavazza. Ethical considerations and policy interventions concerning the impact of generative ai tools in the economy and in society. *AI and Ethics*, pp. 1–9, 2024.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*, 2023.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Ingrid E Fisher, Margaret R Garnsey, and Mark E Hughes. Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3):157–214, 2016.

Philip Hans Franses. *Time series models for business and economic forecasting*. Cambridge university press, 1998.

Martin S Fridson and Fernando Alvarez. *Financial statement analysis: a practitioner's guide*. John Wiley & Sons, 2022.

Lawrence M Friedman. *A history of American law*. Simon and Schuster, 2005.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021a. URL `https://arxiv.org/abs/2101.00027`.

Ruizhuo Gao, Zeqi Zhang, Zhenning Shi, Dan Xu, Weijuan Zhang, and Dewei Zhu. A review of natural language processing for financial technology. In *International Symposium on Artificial Intelligence and Robotics 2021*, volume 11884, pp. 262–277. SPIE, 2021b.

Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, et al. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue. *arXiv preprint arXiv:2306.12174*, 2023a.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023b.

Arunim Garg and Vijay Mago. Role of machine learning in medical research: A survey. *Computer science review*, 40:100370, 2021.

Andrew Gee, R Prager, G Treece, C Cash, and L Berman. Processing and visualizing three-dimensional ultrasound data. *The British journal of radiology*, 77(suppl_2):S186–S193, 2004.

Frank W Geels. The impact of the financial–economic crisis on sustainability transitions: Financial investment, governance and public discourse. *Environmental Innovation and Societal Transitions*, 6:67–95, 2013.

Sourojit Ghosh and Aylin Caliskan. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *arXiv preprint arXiv:2305.10510*, 2023.

Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398, 2023.

Paolo Giudici. Fintech risk management: A research challenge for artificial intelligence in finance. *Frontiers in Artificial Intelligence*, 1:1, 2018.

Sunita Goel and Ozlem Uzuner. Do sentiments matter in fraud detection? estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3):215–239, 2016.

Deepthiman Gowda, Benjamin Blatt, Mary Johanna Fink, Lynn Y Kosowicz, Aileen Baecker, and Ronald C Silvestri. A core physical exam for medical students: results of a national survey. *Academic Medicine*, 89 (3):436–442, 2014.

Tsinghua University Data Mining Group. Chatglm-6b: A large-scale chinese generative language model, 2023. URL `https://github.com/THUDM/ChatGLM-6B/tree/main`.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=md68e8iZK1`.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*, 2023.

Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1012–1023, 2022.

Yue Guo, Zian Xu, and Yi Yang. Is chatgpt a financial expert? evaluating language models on financial natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 815–821. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.58.

Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6:1–25, 2020.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1112–1123, 2023.

C. Haitao. LexiLaw: A Legal Text Processing Toolkit. https://github.com/CSHaitao/LexiLaw, 2024. Accessed: 2024-04-29.

Joschka Haltaufderheide and Robert Ranisch. The ethics of chatgpt in medicine and healthcare: A systematic review on large language models (llms). *arXiv preprint arXiv:2403.14473*, 2024.

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.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*, 2021.

Herbert Lionel Adolphus Hart and Leslie Green. *The concept of law*. oxford university press, 2012.

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*, 2024.

Harvard Law School Library. Caselaw access project. https://case.law/, 2023. Free, public access to over 6.5 million decisions published by state and federal courts throughout U.S. history.

Philipp Harzig, Yan-Ying Chen, Francine Chen, and Rainer Lienhart. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*, 2019.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*, 2023a.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. Mabel: Attenuating gender bias using textual entailment data. *arXiv preprint arXiv:2210.14975*, 2022.

Wanwei He, Jiabao Wen, Lei Zhang, Hao Cheng, Bowen Qin, Yunshui Li, Feng Jiang, Junying Chen, Benyou Wang, and Min Yang. Hanfei-1.0. https://github.com/siat-nlp/HanFei, 2023b.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Y Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. *arXiv preprint arXiv:2310.14088*, 2023c.

Zexue He, An Yan, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. "nothing abnormal": Disambiguating medical reports via contrastive knowledge infusion. *arXiv preprint arXiv:2305.08300*, 2023d.

Alec Helbling, Mansi Phute, Matthew Hull, and Duen Horng Chau. Llm self defense: By self examination, llms know they are being tricked. *arXiv preprint arXiv:2308.07308*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021b.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*, 2020.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gang Hu, Ke Qin, Chenhan Yuan, Min Peng, Alejandro Lopez-Lira, Benyou Wang, Sophia Ananiadou, Wanlong Yu, Jimin Huang, and Qianqian Xie. No language is an island: Unifying chinese and english in financial large language models, instruction data, and benchmarks, 2024.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *CHIL Workshop*, 2019.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Weiyi Huang, Jiahao Jiang, Qiang Qu, and Min Yang. Aila: A question answering system in the legal domain. In *IJCAI*, pp. 5258–5260, 2020.

Zengyi Huang, Chang Che, Haotian Zheng, and Chen Li. Research on generative artificial intelligence for virtual financial robo-advisor. *Academic Journal of Science and Technology*, 10(1):74–80, 2024.

Sewoong Hwang and Jonghyuk Kim. Toward a chatbot for financial sustainability. *Sustainability*, 13(6):3173, 2021.

IDEA-CCNL. Fengshenbang-lm: A large-scale generative language model for chinese, 2023. URL https://github.com/IDEA-CCNL/Fengshenbang-LM.

Baichuan Inc. Baichuan-7b: A large-scale chinese generative language model, 2023. URL https://github.com/baichuan-inc/Baichuan-7B.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *CoRR*, abs/2311.11944, 2023. doi: 10.48550/ARXIV.2311.11944. URL https://doi.org/10.48550/arXiv.2311.11944.

Kwan Yuen Iu and Vanessa Man-Yi Wong. Chatgpt by openai: The end of litigation lawyers? *Available at SSRN 4339839*, 2023.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

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 Computing Surveys*, 55(12):1–38, 2023.

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.

Cong Jiang and Xiaolei Yang. Legal syllogism prompting: Teaching large language models for legal judgment prediction. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pp. 417–421, 2023.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex'Sandy' Pentland, Yoon Kim, Jad Kabbara, et al. Leveraging large language models for learning complex legal concepts through storytelling. *arXiv preprint arXiv:2402.17019*, 2024.

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. *CoRR*, abs/2009.13081, 2020. URL `https://arxiv.org/abs/2009.13081`.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2567–2577. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1259. URL `https://doi.org/10.18653/v1/D19-1259`.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Roseval Malaquias Junior, Ramon Pires, Roseli Romero, and Rodrigo Nogueira. Juru: Legal brazilian large language model from reputable sources. *arXiv preprint arXiv:2403.18140*, 2024.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023.

Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323, 2020.

Haoqiang Kang and Xiao-Yang Liu. Deficiency of large language models in finance: An empirical examination of hallucination. In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*, 2024. URL https://openreview.net/forum?id=SGiQxu8zFL.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *New Frontiers in Artificial Intelligence: JSAI-isAI 2018 Workshops, JURISIN, AI-Biz, SKL, LENLS, IDAA, Yokohama, Japan, November 12–14, 2018, Revised Selected Papers*, pp. 177–192. Springer, 2019.

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.

Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. *arXiv preprint arXiv:2310.13771*, 2023.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*, 2023.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. Refind: Relation extraction financial dataset. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete (eds.), *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pp. 3054–3063. ACM, 2023. doi: 10.1145/3539618.3591911. URL https://doi.org/10.1145/3539618.3591911.

Arif Ali Khan, Sher Badshah, Peng Liang, Muhammad Waseem, Bilal Khan, Aakash Ahmad, Mahdi Fahmideh, Mahmood Niazi, and Muhammad Azeem Akbar. Ethics of ai: A systematic literature review of principles and challenges. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, pp. 383–392, 2022.

Muhammad Salar Khan and Hamza Umer. Chatgpt in finance: Applications, challenges, and solutions. *Heliyon*, 10(2), 2024.

Mi-Young Kim, Randy Goebel, and S Ken. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*, 2015.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. Bizbench: A quantitative reasoning benchmark for business and finance. *CoRR*, abs/2311.06602, 2023. doi: 10.48550/ARXIV.2311.06602. URL https://doi.org/10.48550/arXiv.2311.06602.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24, 2023.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1007. URL https://aclanthology.org/P18-1007.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

B Shravan Kumar and Vadlamani Ravi. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147, 2016.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. Chain of reference prompting helps llm to think like a lawyer. In *Generative AI+ Law Workshop*, 2023.

Eren Kurshan, Jiahao Chen, Victor Storchan, and Hongda Shen. On the current and emerging challenges of developing fair and ethical ai solutions in financial services. In *Proceedings of the second ACM international conference on AI in finance*, pp. 1–8, 2021.

Taeyoon Kwon, Kai Tzu iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Yongsik Sim, Beomseok Sohn, Dongha Lee, and Jinyoung Yeo. Large language models are clinical reasoners: Reasoning-aware diagnosis framework with prompt-generated rationales, 2024.

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. *arXiv preprint arXiv:2402.10373*, 2024.

Raz Lapid, Ron Langberg, and Moshe Sipper. Open sesame! universal black box jailbreaking of large language models. *arXiv preprint arXiv:2309.01446*, 2023.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*, 2021.

Jean Lee, Nicholas Stevens, Soyeon Caren Han, and Minseok Song. A survey of large language models in finance (finllms). *arXiv preprint arXiv:2402.02315*, 2024.

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, 2020.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599, 2022.

David Leslie. Understanding artificial intelligence ethics and safety. *arXiv preprint arXiv:1906.05684*, 2019.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024a.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a.

Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, and Judy W Gichoya. Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, 5(6):e333–e335, 2023b.

Haohang Li, Yangyang Yu, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced LLM trading agent with layered memory and character design. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024b. URL https://openreview.net/forum?id=sstfVOwbiG.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. CFGPT: chinese financial assistant with large language model. *CoRR*, abs/2309.10654, 2023c. doi: 10.48550/ARXIV.2309.10654. URL https://doi.org/10.48550/arXiv.2309.10654.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3063–3070, 2020a.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pretrained language models for text generation: A survey. *arXiv preprint arXiv:2201.05273*, 2022a.

Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, pp. 844–848. IEEE, 2014.

Qingquan Li, Qifan Zhang, Junjie Yao, and Yingjie Zhang. Event extraction for criminal legal text. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 573–580. IEEE, 2020b.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. Explanations from large language models make small reasoners better, 2022b.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 408–422, 2023d.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023e.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. *arXiv preprint arXiv:2307.01595*, 2023f.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pp. 374–382, 2023g.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. Rain: Your language models can align themselves without finetuning. *arXiv preprint arXiv:2309.07124*, 2023h.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15 (6), 2023i.

Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*, 2023j.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*, 2020.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139, 2019.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18635–18643, 2024a.

Chuang Liu, Junzhuo Li, and Deyi Xiong. Tab-cqa: A tabular conversational question answering dataset on financial reports. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D. Williams (eds.), *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 196–207. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-INDUSTRY.20. URL https://doi.org/10.18653/v1/2023.acl-industry.20.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Qianchu Liu, Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Tajdin Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew P. Lungren, Ozan Oktay, and Javier Alvarez-Valle. Exploring the boundaries of gpt-4 in radiology. *ArXiv*, abs/2310.14573, 2023b. URL https://api.semanticscholar.org/CorpusID:264425949.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models. *CoRR*, abs/2307.10485, 2023c. doi: 10.48550/ARXIV.2307.10485. URL https://doi.org/10.48550/arXiv.2307.10485.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*, 2024c.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023d.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*, 2023e.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Zhengliang Liu, Mengshen He, Zuowei Jiang, Zihao Wu, Haixing Dai, Lian Zhang, Siyi Luo, Tianle Han, Xiang Li, Xi Jiang, et al. Survey on natural language processing in medical image analysis. *Zhong nan da xue xue bao. Yi xue ban= Journal of Central South University. Medical Sciences*, 47(8):981–993, 2022.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pp. 4513–4519, 2021.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. Can large language models reason about medical questions?, 2023.

Zhihai LLM. Wisdom interrogatory: A toolkit for legal language understanding, 2023. URL https://github.com/zhihaiLLM/wisdomInterrogatory.

Cai Long, Kayle Lowe, Jessica Zhang, André dos Santos, Alaa Alanazi, Daniel O'Brien, Erin Wright, and David Cote. A novel evaluation model for assessing chatgpt on otolaryngology–head and neck surgery certification examinations: Performance study. *JMIR Medical Education*, 10, 2023. URL https://api.semanticscholar.org/CorpusID:265110947.

Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *CoRR*, abs/2304.07619, 2023. doi: 10.48550/ARXIV.2304.07619. URL https://doi.org/10.48550/arXiv.2304.07619.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22266–22275, 2024.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *CoRR*, abs/2302.09432, 2023. doi: 10.48550/ARXIV.2302.09432. URL https://doi.org/10.48550/arXiv.2302.09432.

Yingzhou Lu, Chiung-Ting Wu, Sarah J Parker, Zuolin Cheng, Georgia Saylor, Jennifer E Van Eyk, Guoqiang Yu, Robert Clarke, David M Herrington, and Yue Wang. COT: an efficient and accurate method for detecting marker genes among many subtypes. *Bioinformatics Advances*, 2(1):vbac037, 2022.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2342–2348, 2021.

Neil MacCormick. *Legal reasoning and legal theory*. Clarendon Press, 1994.

Pekka Malo, Ankur Sinha, Pekka J. Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796, 2014. doi: 10.1002/ASI.23062. URL https://doi.org/10.1002/asi.23062.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Behrooz Mansouri and Ricardo Campos. Falqu: Finding answers to legal questions. *arXiv preprint arXiv:2304.05611*, 2023.

Carsten Maple, Lukasz Szpruch, Gregory Epiphaniou, Kalina Staykova, Simran Singh, William Penwarden, Yisi Wen, Zijian Wang, Jagdish Hariharan, and Pavle Avramovic. The ai revolution: opportunities and challenges for the finance sector. *arXiv preprint arXiv:2308.16538*, 2023.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15009–15018, 2023.

Eric Martínez. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pp. 1–24, 2024.

Akib Mashrur, Wei Luo, Nayyar A Zaidi, and Antonio Robles-Kelly. Machine learning for financial risk management: a survey. *Ieee Access*, 8:203203–203223, 2020.

Gregory Matoesian and Kristin Enola Gilbert. *Multimodal conduct in the law: Language, gesture and materiality in legal interaction*, volume 32. Cambridge University Press, 2018.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020.

Robert W McGee. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)*, 2023.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Donald E Melnick, Gerard F Dillon, and David B Swanson. Medical licensing examinations in the united states. *Journal of dental education*, 66(5):595–599, 2002.

Bertalan Meskó and Eric J Topol. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4741–4749, 2022.

Andy Michel, Sumit Kumar Jha, and Rickard Ewetz. A survey on the vulnerability of deep neural networks against adversarial attacks. *Progress in Artificial Intelligence*, 11(2):131–141, 2022.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023a.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*, 2023b.

Timo Minssen, Effy Vayena, and I Glenn Cohen. The challenges for regulating medical use of chatgpt and other large language models. *Jama*, 2023.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.

Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*, 2019.

Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*, 2021.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*, 2023.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D Griffin. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*, 2023.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 10893–10906. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.748. URL https://doi.org/10.18653/v1/2022.emnlp-main.748.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

John J Nay, David Karamardian, Sarah B Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H Choi, and Jungo Kasai. Large language models as tax attorneys: a case study in legal capabilities emergence. *Philosophical Transactions of the Royal Society A*, 382(2270):20230159, 2024.

James O' Neill, Paul Buitelaar, Cecile Robin, and Leona O' Brien. Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pp. 159–168, 2017.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*, 2021.

Ha-Thanh Nguyen. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*, 2023.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*, 2021.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of gpt-4 on medical challenge problems, 2023a.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023b.

Catherine Nunez. Artificial intelligence and legal ethics: Whether ai lawyers can make ethical decisions. *Tul. J. Tech. & Intell. Prop.*, 20:189, 2017.

Benjamin E. Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. *CoRR*, abs/1806.04185, 2018. URL http://arxiv.org/abs/1806.04185.

Vinayak Yogesh Ogavinee and et al. Anima: A comprehensive toolkit for medical image analysis, 2022. URL https://github.com/lyogavin/Anima.

Joel Oksanen, Abhilash Majumder, Kumar Saunack, Francesca Toni, and Arun Dhondiyal. A graph-based method for unsupervised knowledge discovery from financial texts. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pp. 5412–5417. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.579.

Takuma Okuda and Sanae Shoda. Ai-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal*, 54(2):4–8, 2018.

Jasmine Chiat Ling Ong, Shelley Yin-Hsi Chang, Wasswa William, Atul J Butte, Nigam H Shah, Lita Sui Tjien Chew, Nan Liu, Finale Doshi-Velez, Wei Lu, Julian Savulescu, et al. Ethical and regulatory challenges of large language models in medicine. *The Lancet Digital Health*, 2024.

OpenAI. Introducing chatgpt, 2022. URL https://openai.com/blog/chatgpt. Accessed: 2023-05-11.

OpenAI. Gpt-4 technical report, 2023.

Hadas Orgad and Yonatan Belinkov. Blind: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8801–8821, 2023.

Kun Ouyang, Yi Liu, Shicheng Li, Ruihan Bao, Keiko Harimoto, and Xu Sun. Modal-adaptive knowledge-enhanced graph-based financial prediction from monetary policy conference calls with llm, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Ahmet Murat Ozbayoglu, Mehmet Ugur Gudelek, and Omer Berat Sezer. Deep learning for financial applications: A survey. *Applied soft computing*, 93:106384, 2020.

J P. Collins, RM Harden. Amee medical education guide no. 13: real patients, simulated patients and simulators in clinical examinations. *Medical teacher*, 20(6):508–521, 1998.

Vasile Păis, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Vlad Silviu Coneschi, Andrei Onut, et al. Legalnero: A linked corpus for named entity recognition in the romanian legal domain. *Semantic Web*, (Preprint):1–14.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H. Chen, Tom J. Pollard, Joyce C. Ho, and Tristan Naumann (eds.), *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*, 2018.

Yamini Pandey. Credit card fraud detection using deep learning. *International Journal of Advanced Research in Computer Science*, 8(5), 2017.

Dimitris Papailiopoulos. Gpt-4 "discovered" the same sorting algorithm as alphadev by removing "mov s p"., June 2023. URL https://x.com/DimitrisPapail/status/1666843952824168465?s=20.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*, 2021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Taejin Park. Enhancing anomaly detection in financial markets with an llm-based multi-agent framework, 2024.

Yulong Pei, Amarachi Mbakwe, Akshat Gupta, Salwa Alamir, Hanxuan Lin, Xiaomo Liu, and Sameena Shah. TweetFinSent: A dataset of stock sentiments on Twitter. In Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen (eds.), *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pp. 37–47, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.finnlp-1.5. URL https://aclanthology.org/2022.finnlp-1.5.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Kuashuai Peng and Guofeng Yan. A survey on deep learning for financial risk prediction. *Quantitative Finance and Economics*, 5(4):716–737, 2021.

Shengxin Peng, Deqiang Wang, Yuanhao Liang, Wenshan Xiao, Yixiang Zhang, and Lei Liu. Ai-chatgpt/gpt-4: An booster for the development of physical medicine and rehabilitation in the new era! *Annals of Biomedical Engineering*, 52:462 – 466, 2023. URL https://api.semanticscholar.org/CorpusID:260245954.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pp. 258–262, 2016.

Prakash Poudyal, Jaromír Šavelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. Echr: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pp. 67–75, 2020.

Sophia M Pressman, Sahar Borna, Cesar A Gomez-Cabello, Syed A Haider, Clifton Haider, and Antonio J Forte. Ai and ethics: A systematic review of the ethical considerations of large language model use in surgery research. In *Healthcare*, volume 12, pp. 825. MDPI, 2024.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021a. URL http://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021b.

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.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 2023.

Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, and Chris Tanner. Docfinqa: A long-context financial reasoning dataset. *CoRR*, abs/2401.06915, 2024. doi: 10.48550/ARXIV.2401.06915. URL https://doi.org/10.48550/arXiv.2401.06915.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Harry V Roberts. Stock-market" patterns" and financial analysis: methodological suggestions. *The Journal of Finance*, 14(1):1–10, 1959.

Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*, 2018.

Sohini Roychowdhury, Andres Alvarez, Brian Moore, Marko Krema, Maria Paz Gelpi, Punit Agrawal, Federico Martín Rodríguez, Ángel Rodríguez, José Ramón Cabrejas, Pablo Martínez Serrano, et al. Hallucination-minimized data-to-answer framework for financial decision-makers. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 4693–4702. IEEE, 2023.

Francesco Rundo, Francesca Trenta, Agatino Luigi Di Stallo, and Sebastiano Battiato. Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24):5574, 2019.

Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. Sabiá-2: A new generation of portuguese large language models. *arXiv e-prints*, pp. arXiv–2403, 2024.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff,

Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022. doi: 10.48550/ARXIV.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Leo Schwinn, David Dobre, Stephan Günnemann, and Gauthier Gidel. Adversarial attacks and defenses in large language models: Old and new threats. *arXiv preprint arXiv:2310.19737*, 2023.

IRLab SDU. Fuzi Mingcha: Project for [add project description here]. https://github.com/irlab-sdu/fuzi.mingcha, 2023. Accessed: April 15, 2024.

Prasad Seemakurthi, Shuhao Zhang, and Yibing Qi. Detection of fraudulent financial reports with machine learning techniques. In *2015 Systems and information engineering design symposium*, pp. 358–361. IEEE, 2015.

Isabel Segura-Bedmar, Paloma Martínez Fernández, and María Herrero Zazo. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Association for Computational Linguistics, 2013.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Nurullah Sevim, Furkan Şahinuç, and Aykut Koç. Gender bias in legal corpora and debiasing it. *Natural Language Engineering*, 29(2):449–482, 2023.

Agam Shah, Suvan Paturi, and Sudheer Chava. Trillion dollar words: A new financial dataset, task & market analysis. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6664–6679. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.368. URL https://doi.org/10.18653/v1/2023.acl-long.368.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When FLUE meets FLANG: benchmarks and large pretrained language model for financial domain. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 2322–2335. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.148. URL https://doi.org/10.18653/v1/2022.emnlp-main.148.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H Gandomi. Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

Shirong Shen, Guilin Qi, Zhen Li, Sheng Bi, and Lusheng Wang. Hierarchical chinese legal event extraction via pedal attention mechanism. In *Proceedings of the 28th international conference on computational linguistics*, pp. 100–113, 2020.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*, 2023a.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023b.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *Advances in Neural Information Processing Systems*, 35:13158–13173, 2022.

Yuge Shi, Jeffrey Seely, Philip HS Torr, N Siddharth, Awni Hannun, Nicolas Usunier, and Gabriel Synnaeve. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.

Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. Red teaming language model detectors with language models. *Transactions of the Association for Computational Linguistics*, 12:174–189, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

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. *arXiv preprint arXiv:2212.13138*, 2022.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

Ankur Sinha and Tanmay Khandait. Impact of news on the commodity market: Dataset and results. *CoRR*, abs/2009.04202, 2020. URL `https://arxiv.org/abs/2009.04202`.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 2017.

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. Beyond classification: Financial reasoning in state-of-the-art language models. *CoRR*, abs/2305.01505, 2023. doi: 10.48550/ARXIV.2305.01505. URL `https://doi.org/10.48550/arXiv.2305.01505`.

Yejun Soun, Jaemin Yoo, Minyong Cho, Jihyeong Jeon, and U Kang. Accurate stock movement prediction with self-supervised learning from sparse noisy tweets. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1691–1700. IEEE, 2022.

Francesco Sovrano, Monica Palmirani, Biagio Distefano, Salvatore Sapienza, and Fabio Vitali. A dataset for evaluating legal question answering on private international law. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 230–234, 2021.

Harold J Spaeth, Lee Epstein, Andrew D Martin, Jeffrey A Segal, Theodore J Ruger, and Sara C Benesh. 2017 supreme court database, version 2017 release 01. *URL: http://Supremecourtdatabase. org*, 2017.

Dananjay Srinivas, Rohan Das, Saeid Tizpaz-Niari, Ashutosh Trivedi, and Maria Leonor Pacheco. On the potential and limitations of few-shot in-context learning to generate metamorphic specifications for tax preparation software. *arXiv preprint arXiv:2311.11979*, 2023.

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, and Dan Roth. Evaluating llms' mathematical reasoning in financial document question answering, 2024.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, pp. 39–51, 2022.

Quinten Steenhuis, David Colarusso, and Bryce Willey. Weaving pathways for justice with gpt: Llm-driven automated drafting of interactive legal applications. *arXiv preprint arXiv:2312.09198*, 2023.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schölkopf, and Mrinmaya Sachan. A causal framework to quantify the robustness of mathematical reasoning with language models. *arXiv preprint arXiv:2210.12023*, 2022.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

Zhongxiang Sun. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*, 2023.

Cass R Sunstein. *Legal reasoning and political conflict*. Oxford University Press, 2018.

Ekaterina Svetlova. Ai ethics and systemic risks in finance. *AI and Ethics*, 2(4):713–725, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 26–41, 2022.

Jinzhe Tan, Hannes Westermann, and Karim Benyekhlef. Chatgpt as an artificial lawyer. *Artificial Intelligence for Access to Justice (AI4AJ 2023)*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. *arXiv preprint arXiv:2306.04597*, 2023.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

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.

Patricia A Thomas, David E Kern, Mark T Hughes, Sean A Tackett, and Belinda Y Chen. *Curriculum development for medical education: a six-step approach*. JHU press, 2022.

Chandra S Throckmorton, William J Mayew, Mohan Venkatachalam, and Leslie M Collins. Financial fraud detection using vocal, linguistic and financial cues. *Decision Support Systems*, 74:78–87, 2015.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*, 2019.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a.

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*, 2023b.

Hieu Tran, Zhichao Yang, Zonghai Yao, and Hong Yu. Bioinstruct: Instruction tuning of large language models for biomedical natural language processing. *ArXiv*, abs/2310.19975, 2023. URL `https://api.semanticscholar.org/CorpusID:264744285`.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*, 2022.

Arianna Trozze, Toby Davies, and Bennett Kleinberg. Large language models in cryptocurrency securities cases: can a gpt model meaningfully assist lawyers? *Artificial Intelligence and Law*, pp. 1–47, 2024.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artières, Axel-Cyrille Ngonga Ngomo, Norman Heino, Éric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.*, 16:138:1–138:28, 2015. doi: 10.1186/S12859-015-0564-6. URL `https://doi.org/10.1186/s12859-015-0564-6`.

Don Tuggener, Pius Von Däniken, Thomas Peetz, and Mark Cieliebak. Ledgar: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1235–1241, 2020.

Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic Pathology*, 19(1):1–9, 2024.

United States Congress. Table of supreme court decisions overruled by subsequent decisions. `https://constitution.congress.gov/resources/decisionsoverruled/`, 2023. A comprehensive table listing the decisions of the U.S. Supreme Court that have been overruled by subsequent decisions.

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.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2, 2020.

David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 960–966. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1092. URL https://doi.org/10.18653/v1/p19-1092.

Dingzirui Wang, Longxu Dou, Wenbin Zhang, Junyu Zeng, and Wanxiang Che. Exploring equation as a better intermediate meaning representation for numerical reasoning of large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19116–19125, Mar. 2024a. doi: 10.1609/aaai.v38i17.29879. URL https://ojs.aaai.org/index.php/AAAI/article/view/29879.

Dingzirui Wang, Longxu Dou, Xuanliang Zhang, Qingfu Zhu, and Wanxiang Che. Enhancing numerical reasoning with the guidance of reliable reasoning processes. *arXiv preprint arXiv:2402.10654*, 2024b.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding, 2023a.

Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, et al. Are large language models really robust to word-level perturbations? *arXiv preprint arXiv:2309.11166*, 2023b.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):1–26, 2024c.

Neng Wang, Hongyang Yang, and Christina Dan Wang. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *CoRR*, abs/2310.04793, 2023c. doi: 10.48550/ARXIV.2310.04793. URL https://doi.org/10.48550/arXiv.2310.04793.

Peipeng Wang, Xiuguo Zhang, and Zhiying Cao. Few-shot charge prediction with data augmentation and feature augmentation. *Applied Sciences*, 11(22):10811, 2021.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023d.

Zezhong Wang, Fangkai Yang, Lu Wang, Pu Zhao, Hongru Wang, Liang Chen, Qingwei Lin, and Kam-Fai Wong. Self-guard: Empower the llm to safeguard itself. *arXiv preprint arXiv:2310.15851*, 2023e.

Zhuo Wang, Rong Hua Li, Bowen Dong, Jie Wang, Xiuxing Li, Ning Liu, Chenhui Mao, Wei Zhang, Liling Dong, Jing Gao, and Jianyong Wang. Can llms like gpt-4 outperform traditional ai tools in dementia diagnosis? maybe, but not today. *ArXiv*, abs/2306.01499, 2023f. URL https://api.semanticscholar.org/CorpusID:259064252.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

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, 2022b.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: a comprehensive review. *Computers & security*, 57:47–66, 2016.

Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. Llmediator: Gpt-4 assisted online dispute resolution. *arXiv preprint arXiv:2307.16732*, 2023.

David A Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, et al. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Medical Imaging with Deep Learning*, pp. 811–826. PMLR, 2020.

Steven A Wright. Ai in the law: Towards assessing ethical risks. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 2160–2169. IEEE, 2020.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023a.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023b.

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*, pp. ocae045, 2024.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. Hybrid deep sequential modeling for social text-driven stock prediction. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang (eds.), *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pp. 1627–1630. ACM, 2018. doi: 10.1145/3269206.3269290. URL `https://doi.org/10.1145/3269206.3269290`.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *arXiv preprint arXiv:2305.19860*, 2023c.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564, 2023d. doi: 10.48550/ARXIV.2303.17564. URL `https://doi.org/10.48550/arXiv.2303.17564`.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*, 2023e.

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.

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. *CoRR*, abs/1609.08144, 2016. URL `http://arxiv.org/abs/1609.08144`.

Hana Demma Wube, Sintayehu Zekarias Esubalew, Firesew Fayiso Weldesellasie, and Taye Girma Debelee. Text-based chatbot in financial sector: A systematic literature review. *Data Sci. Financ. Econ*, 2(3): 232–259, 2022.

Adam Zachary Wyner, Biralatei James Fawei, and Jeff Z Pan. Passing a usa national bar exam: a first corpus for experimentation. In *LREC 2016, Tenth International Conference on Language Resources and Evaluation*. LREC, 2016.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*, 2018.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2019-scm: A dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*, 2019.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84, 2021.

Peng Xiao-Song. LaWGPT: A Legal Writing GPT Model. `https://github.com/pengxiao-song/LaWGPT`, 2024. Accessed: 2024-04-29.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *CoRR*, abs/2306.05443, 2023. doi: 10.48550/ARXIV.2306.05443. URL `https://doi.org/10.48550/arXiv.2306.05443`.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*, 2024a.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*, 2024b.

Zhongbin Xie and Thomas Lukasiewicz. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. *arXiv preprint arXiv:2306.04067*, 2023.

Frank Xing. Designing heterogeneous llm agents for financial sentiment analysis. *arXiv preprint arXiv:2401.05799*, 2024.

Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data, 2023.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057, 2015.

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 1970–1979. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1183. URL https://aclanthology.org/P18-1183/.

Nicole Yamane. Artificial intelligence in the legal field and the indispensable human element legal ethics demands. *Geo. J. Legal Ethics*, 33:877, 2020.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*, 2021.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. Personalized showcases: Generating multi-modal explanations for recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2251–2255, 2023a.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3090–3100, 2023b.

An Yan, Yu Wang, Yiwu Zhong, Zexue He, Petros Karypis, Zihan Wang, Chengyu Dong, Amilcare Gentili, Chun-Nan Hsu, Jingbo Shang, et al. Robust and interpretable medical image classifiers via concept bottleneck models. *arXiv preprint arXiv:2310.03182*, 2023c.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024a.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024b.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. Finbert: A pretrained language model for financial communications. *CoRR*, abs/2006.08097, 2020. URL https://arxiv.org/abs/2006.08097.

Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning. *CoRR*, abs/2309.13064, 2023a. doi: 10.48550/ARXIV.2309.13064. URL https://doi.org/10.48550/arXiv.2309.13064.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023b.

Feng Yao, Chaojun Xiao, Xiaozhi Wang, Zhiyuan Liu, Lei Hou, Cunchao Tu, Juanzi Li, Yun Liu, Weixing Shen, and Maosong Sun. Leven: A large-scale chinese legal event detection dataset. *arXiv preprint arXiv:2203.08556*, 2022a.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25407–25437. PMLR, 2022b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, pp. 100211, 2024.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2841–2858, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.187. URL https://aclanthology.org/2023.findings-emnlp.187.

Wentao Ye, Mingfeng Ou, Tianyi Li, Xuetao Ma, Yifan Yanggong, Sai Wu, Jie Fu, Gang Chen, Junbo Zhao, et al. Assessing hidden risks of llms: an empirical study on robustness, consistency, and credibility. *arXiv preprint arXiv:2305.10235*, 2023b.

Fangyi Yu, Lee Quartey, and Frank Schilder. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*, 2022.

Fangyi Yu, Lee Quartey, and Frank Schilder. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13582–13596, 2023a.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*, 2023b.

Ping Yu, Hua Xu, Xia Hu, and Chao Deng. Leveraging generative ai and large language models: a comprehensive roadmap for healthcare integration. In *Healthcare*, volume 11, pp. 2776. MDPI, 2023c.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators, 2023d.

Xinli Yu, Zheng Chen, and Yanbin Lu. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In Mingxuan Wang and Imed Zitouni (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 739–753, Singapore, December 2023e. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.69. URL https://aclanthology.org/2023.emnlp-industry.69.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36, 2024.

Chongjian Yue, Xinrun Xu, Xiaojun Ma, Lun Du, Hengyu Liu, Zhiming Ding, Yanbing Jiang, Shi Han, and Dongmei Zhang. Enabling and analyzing how to efficiently extract information from hybrid long documents with llms, 2024.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Yongji Wang, and Jian-Guang Lou. Large language models meet nl2code: A survey. *arXiv preprint arXiv:2212.09420*, 2022.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*, 2023.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *CoRR*, abs/2306.12659, 2023a. doi: 10.48550/ARXIV.2306.12659. URL https://doi.org/10.48550/arXiv.2306.12659.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, pp. 349–356, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400702402. doi: 10.1145/3604237.3626866. URL https://doi.org/10.1145/3604237.3626866.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37, 2023c.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*, 2023d.

Ruizhe Zhang, Haitao Li, Yueyue Wu, Qingyao Ai, Yiqun Liu, Min Zhang, and Shaoping Ma. Evaluation ethics of llms in legal domain. *arXiv preprint arXiv:2403.11152*, 2024a.

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*, 2023e.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. Mitigating language model hallucination with interactive question-knowledge alignment. *arXiv preprint arXiv:2305.13669*, 2023f.

Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Sophia Ananiadou, Min Peng, Jimin Huang, and Qianqian Xie. Dólares or dollars? unraveling the bilingual prowess of financial llms between spanish and english. *CoRR*, abs/2402.07405, 2024b. doi: 10.48550/ARXIV.2402.07405. URL https://doi.org/10.48550/arXiv.2402.07405.

Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. Enhancing small medical learners with privacy-preserving contextual prompting. *CoRR*, abs/2305.12723, 2023g. doi: 10.48550/ARXIV.2305.12723. URL https://doi.org/10.48550/arXiv.2305.12723.

Xinlu Zhang, Shiyang Li, Xianjun Yang, Chenxin Tian, Yao Qin, and Linda Ruth Petzold. Enhancing small medical learners with privacy-preserving contextual prompting, 2023h.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks, 2023i.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare:instruction-tuned large language models for medical application, 2023j.

Yuanmin Zhang, Junjun Jiang, and Yanjun Li. Intelligent analysis and application of judicial big data sharing based on blockchain. In *2023 6th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pp. 592–596. IEEE, 2023k.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023l.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Zhuosheng Zhang, Hanqing Zhang, Keming Chen, Yuhang Guo, Jingyun Hua, Yulong Wang, and Ming Zhou. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *CoRR*, abs/2110.06696, 2021. URL https://arxiv.org/abs/2110.06696.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

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*, 2023a.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6588–6600, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.454. URL https://aclanthology.org/2022.acl-long.454.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. Knowledgemath: Knowledge-intensive math word problem solving in finance domains, 2023b.

Andrew Zhe. Lawyer-llama: A legal-specific language model, 2023. URL https://github.com/AndrewZhe/lawyer-llama.

Andrew Zhe. lawyer-llama: A Machine Learning Toolkit for Legal Analysis. https://github.com/AndrewZhe/lawyer-llama, 2024. Accessed: 2024-04-29.

Jiaping Zheng, Wendy W Chapman, Rebecca S Crowley, and Guergana K Savova. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6):1113–1122, 2011.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pp. 159–168, 2021.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. Jec-qa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 9701–9708, 2020.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. A survey of large language models in medicine: Progress, application, and challenge, 2024a.

Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. Are large language models rational investors? *arXiv preprint arXiv:2402.12713*, 2024b.

Zhihan Zhou, Liqian Ma, and Han Liu. Trade the event: Corporate events detection for news-based event-driven trading. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pp. 2114–2124. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-ACL.186. URL https://doi.org/10.18653/v1/2021.findings-acl.186.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.254. URL https://doi.org/10.18653/v1/2021.acl-long.254.

Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat-Seng Chua. Tat-llm: A specialized language model for discrete reasoning over tabular and textual data, 2024.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*, 2022.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*, 2023a.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023b.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867*, 2023.

Xue Zongyue, Liu Huanghai, Hu Yiran, Kong Kangle, Wang Chenlu, Liu Yun, and Shen Weixing. Leec: A legal element extraction dataset with an extensive domain-specific label system. *arXiv preprint arXiv:2310.01271*, 2023.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Jinan Zou, Haiyao Cao, Lingqiao Liu, Yuhao Lin, Ehsan Abbasnejad, and Javen Qinfeng Shi. Astock: A new dataset and automated stock trading based on stock-specific news analyzing model. *CoRR*, abs/2206.06606, 2022. doi: 10.48550/ARXIV.2206.06606. URL https://doi.org/10.48550/arXiv.2206.06606.

Yang Zou, Arto Kiviniemi, and Stephen W Jones. Retrieving similar cases for construction project risk management using natural language processing techniques. *Automation in construction*, 80:66–76, 2017.