Large Language Models for Data Science: A Survey

Anonymous ARR submission

Abstract

Data science is an interdisciplinary field that focuses on extracting knowledge from raw data using statistical analysis and machine learning techniques. However, as data continues to grow in scale and complexity, data scientists face increasing challenges in handling unstructured data, automating workflows, and scaling analytical processes. The advancements of large language models (LLMs) present an unprecedented opportunity to enhance and streamline data science tasks by enabling automation and augmentation of key processes in the data science pipeline. This survey contributes to four core aspects: the role of LLMs in the data science cycle, specialized domain applications, challenges and limitations, and social impact and future directions. Furthermore, we introduce a structured framework defining how LLMs contribute to each stage of data science, provide an in-depth discussion on their applications in key domains such as healthcare and finance, analyze key obstacles such as data quality and model interpretability, and explore ethical concerns and future research opportunities. Serving as a comprehensive resource, this survey aims to assist researchers and practitioners in understanding and utilizing LLMs to advance modern data science methodologies.

1 Introduction

The field of data science has experienced significant advancements over the past decade, driven by the increasing availability of large-scale data and the development of sophisticated machine learning techniques (Wu, 2024; Lyko et al., 2016a; Hong et al., 2024). Data science primarily focuses on extracting knowledge from raw data, with the ultimate goal of deriving data-driven actionable insights across diverse industries, including healthcare, finance, and engineering (Cady, 2024; Sarker, 2021). For example, in healthcare, data science is revolutionizing personalized medicine and predictive diagnostics through data-driven methodologies (Goetz and Schork, 2018). Generally, traditional methods applied in data science rely on a combination of statistical analysis and machine learning techniques (Li et al., 2017; Qin et al., 2020). However, as data complexity continues to grow, researchers in data science face numerous challenges in handling unstructured data, automating workflows, and scaling analytical processes. These challenges necessitate more intelligent, flexible, and scalable solutions to support modern data science solutions.

Recently, the emergence of large language models (LLMs) has introduced a significant shift in how data science tasks are performed. LLMs, such as GPT-4 (OpenAI, 2024a) and LLaMA (Dubey et al., 2024), possess powerful capabilities in question answering (Talmor et al., 2018; Kwiatkowski et al., 2019), code generation (Ni et al., 2023; Nascimento et al., 2024), and contextual reasoning (Liu and colleagues, 2024; Hao et al., 2023). Such capabilities of LLMs have motivated researchers to explore their assistance in automating and enhancing various stages of the data science workflow. For example, LLMs can assist in data extraction (Katz et al., 2024), feature engineering (Li et al., 2024d), statistical analysis (Brugere et al., 2024), and visualization (Ko et al., 2024), as illustrated in Fig. 1. Additionally, LLMs also improve accessibility by allowing interactions based on natural language with complex data systems.

Nevertheless, despite their growing adoption in data science, LLMs also present several challenges. For example, they struggle with numerical computations in data science applications (Fang et al., 2024), limiting their reliability for tasks requiring precise quantitative reasoning. Additionally, societal biases embedded in LLMs may pose ethical concerns, potentially affecting fairness in decision-making (Wang et al., 2024). Given these limitations, there is an urgent need for a systematic review of current LLM applications in data science, as well as an in-depth exploration to advance LLMs for data science.

In this survey, we investigate the role of LLMs in data science based on our proposed taxonomy and analyze their impact across workflow stages. The primary contributions of this survey are as follows:

- **Taxonomy based on the Data Science Workflow:** We systematically examine how LLMs contribute to various stages of the data science workflow, including data acquisition, preparation, analysis, and presentation. Through the taxonomy, we aim to provide a comprehensive overview of LLM-driven automation and augmentation across these critical phases.

- **Limitations of LLMs in Data Science:** We analyze key challenges associated with integrating LLMs into data science workflows, including issues related to inconsistency (e.g., sensitivity to data formats), inefficiency (e.g., difficulty in handling large data), and insecurity (e.g., bias and privacy risks).

Additionally, we highlight potential future research directions and introduce the applications of LLMs for data science in specific domains such as finance and healthcare. In concrete, this survey aims to serve as a guide for researchers and practitioners seeking to leverage LLMs for various data-driven tasks, including data processing, analysis, and communication.

Differences between This Survey and Others. Several surveys have explored various aspects of LLMs in data science (Zeng et al., 2024a, 2023; Hua et al., 2024; Zhao et al., 2023; Hadi et al., 2023). However, these surveys often lack a holistic view of LLMs' roles across the entire data science workflow. Particularly, they tend to focus on specific stages, such as data acquisition (Liu et al., 2024d; Wang et al., 2023h; Albalak et al., 2024), or specific data types, such as tabular data (Sui et al., 2024; Ruan et al., 2024; Fang et al., 2024). Other related surveys concentrate on the application of LLMs in specific domains, including health (He et al., 2025; Goedde et al., 2023; Qiu et al., 2023; Zhou et al., 2023a; Yuan et al., 2024b), finance (Li et al., 2023d; Zhao et al., 2024c; Lee et al., 2024a; Nie et al., 2024), and education (Al-Smadi, 2023; Wang et al., 2024e; Hosseini et al., 2023; Li et al., 2023b). In contrast, this survey offers a comprehensive framework for the utilization of LLMs in four interconnected stages of the data science workflow.

2 Data Science Workflow

In this work, we define the data science workflow that typically involves four interconnected stages: data acquisition, data preparation, data analysis, and data interpretation. Traditional approaches to data science often involve labor-intensive tasks that can limit scalability and adaptability (Khatri and Brown, 2010a). However, with the increasing complexity and volume of data, modern data science requires automated, scalable, and context-aware solutions to effectively manage diverse data sources. In this regard, LLMs are emerging as transformative tools in data science, offering enhancements across all phases of the workflow:

- **Data Acquisition:** This stage involves collecting raw data from various sources, including structured databases, unstructured text, and sensor streams. In this stage, LLMs can help automate the collection, annotation, and synthesis of both structured and unstructured data (Gur and colleagues, 2022).

- **Data Preparation**: This stage encompasses feature engineering, data cleaning, and storage management to ensure data quality and usability. LLMs can help refine features through semantic analysis and dynamic transformations (Choi et al., 2024).

- **Data Analysis**: In this stage, both textual (e.g., summarization and reasoning) and numerical (e.g., statistical modeling) analyses are performed to extract meaningful insights from data. LLMs can improve both quantitative and qualitative analyses by providing contextual insights and complex reasoning (Liu and colleagues, 2024).

- **Data Interpretation**: This stage involves communicating findings and insights through visualization and decision-making systems. LLMs in this stage can help create visualizations and summaries that enhance data accessibility and usability (Chen et al., 2023c).

Together, these stages form an iterative and structured pipeline that enables efficient and scalable data-driven insights. By integrating LLMs across these phases, data scientists can streamline the workflow, minimize manual effort, and improve the reliability of analytical results.

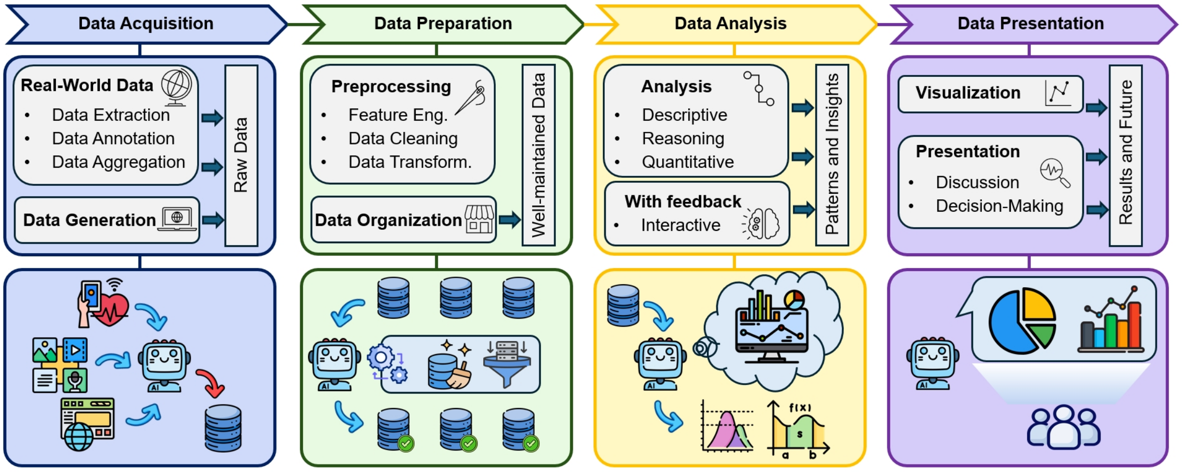

Figure 1: The proposed taxonomy of research works on LLM for data science.

3 LLMs for Data Acquisition

Data acquisition serves as a foundational stage of the data science workflow, encompassing the collection of data from diverse sources (Lyko et al., 2016b). Traditional approaches often face challenges related to data heterogeneity and quality, limiting their efficiency in handling large and complex datasets. LLMs offer promising solutions by automating data extraction, annotation, aggregation, and even generating structured data from raw inputs. In the following sections, we explore existing research on leveraging LLMs to enhance these four key aspects of data acquisition.

3.1 Data Extraction

Extracting valuable data from raw sources is often a time-consuming and domain-specific process, as it relies on rule-based or specialized machine learning approaches. LLMs can streamline this by automatically parsing data across websites, scholarly archives, and online platforms, producing summaries and structured outputs (Gur et al., 2022; Zheng et al., 2024; Dolphin et al., 2024; Jiang and Ferrara, 2023; Katz et al., 2024; Li et al., 2024o). LLMs also excel at entity recognition and metadata generation (Sancheti et al., 2024; Kim and Lee, 2024; Harrod et al., 2024; Zhang and Soh, 2024; Gillani et al., 2024), facilitate multilingual tasks (Whitehouse et al., 2023; Li et al., 2024q; Thellmann et al., 2024; Miah et al., 2024; Kmainasi et al., 2024), and handle multimodal data (Wysocki et al., 2024; Wang et al., 2024c; Ge et al., 2024). In IoT applications, LLMs assist in detecting sensor anomalies and identifying inter-device correlations for proactive maintenance (Cui et al., 2024; Kok et al., 2024; An et al., 2024; Shirali et al., 2024; Abshari et al., 2024; Worae et al., 2024).

Additionally, domain-specific prompting can guide LLMs to focus on the most relevant data, which is particularly effective in medical applications (Xiao et al., 2024a; Xia et al., 2024; Carta et al., 2023; Hu et al., 2024b), leading to more accurate results. Moreover, retrieval-augmented generation (RAG) boosts factual precision and reduces errors (Ding et al., 2024b; Arslan and Cruz, 2024; Li et al., 2024f,h). By streamlining extraction, these techniques ensure that only the most relevant data progresses to subsequent steps of further refinement, such as annotation and aggregation.

3.2 Data Annotation

After data extraction, refining the data through annotation is typically required for effective model training and evaluation. However, this process is often labor-intensive, particularly for large-scale or multimodal datasets (Sorokin and Forsyth, 2008). LLMs reduce this overhead by automatically generating labels for data in tasks such as sentiment analysis, object detection, and transcription (Tan et al., 2024; Wang et al., 2024b; Shvetsova et al., 2025), demonstrating strong performance in zero- and few-shot settings (Sapkota et al., 2024; Zhang et al., 2023a; Sainz et al., 2023; Chen et al., 2024h). Techniques such as advanced prompting (e.g., self-correction, multi-agent debates) and domain-specific fine-tuning further improve label quality (Kamoi et al., 2024; Chen and Si, 2024; Li et al., 2024c; Alizadeh et al., 2025; Bansal and Sharma, 2023; Tang et al., 2023). Additionally, human-in-the-loop and active learning techniques help refine or verify labels, ensuring reliability for downstream training (Li et al., 2023a; Tang et al., 2024; Kim et al., 2024b; Pangakis and Wolken, 2024; Ming et al., 2024).

3.3 Data Aggregation

After annotation, data is generally aggregated into a unified format for efficient analysis and decision-making. Traditional methods often rely on rigid schemas or structured databases, which can falter with semi-structured or unstructured content (Orcutt et al., 1968; Jesus et al., 2014; Bloodgood, 2025). On the other hand, LLMs offer greater flexibility by processing both structured data (e.g., tables) and unstructured (e.g., text, images), enabling seamless integration (Das et al., 2024b; Su et al., 2024; Le et al., 2024). Additionally, LLMs can generate concise summaries of trends and correlations (Kurisinkel and Chen, 2023; Kuang et al., 2024; Li et al., 2024a), producing reports or dashboards that guide decision-making (John et al., 2023; Weng et al., 2024b,a).

3.4 Data Generation

Although LLMs excel at extracting, annotating, and aggregating data, they still face challenges when raw data is scarce in specific scenarios, such as disease modeling and low-resource languages (Yuan et al., 2023a; Whitehouse et al., 2023; Song et al., 2024). This data scarcity can severely hinder model development (Zhou et al., 2024e; Ding et al., 2024a). Nevertheless, LLMs help mitigate these limitations by generating synthetic datasets across various domains. For example, LLMs can simulate diverse IoT sensor conditions (Zhou et al., 2024c,d; An et al., 2024) and impute missing values across numerical, categorical, and textual fields (Little and Rubin, 2019; He et al., 2024; Chen et al., 2024g; Wang et al., 2025). Furthermore, by adhering to domain-specific distributions, LLMs generate specialized datasets for applications in finance, healthcare, and transportation (Nie et al., 2024; Belyaeva et al., 2023; Zhang et al., 2024h,l). These capabilities enable LLMs to produce high-fidelity synthetic data, ensuring dataset availability and enhancing model robustness even in data-scarce environments.

4 LLMs for Data Preparation

Effective data preparation is crucial for ensuring the quality, accessibility, and usability of data for downstream steps (Khatri and Brown, 2010b). Data preparation generally aims to transform extracted data into a structured and analyzable format. Traditional approaches often struggle with labor-intensive workflows, making automation a pressing need. More recently, LLMs offer transformative solutions by automating feature extraction and streamlining data cleaning, thereby enhancing the efficiency and scalability of data preparation.

4.1 Feature Engineering

In the data preparation stage, feature engineering focuses on selecting, transforming, and generating data features to facilitate data analysis or improve model training (Li et al., 2017; Kusiak, 2001). LLMs automate and optimize feature engineering by analyzing dataset descriptions (e.g., column names, sample values) and adaptively selecting relevant features based on semantic or statistical cues (Choi et al.; Jeong et al., 2024; Lee et al., 2024b; Jia et al., 2024b; Toufiq et al., 2023; Li et al., 2024d; Luo et al., 2024; Yang et al., 2024a). Beyond selection, LLMs generate new features through arithmetic operations (Han et al.; Lee et al., 2024b; Zhang et al., 2024g; Küken et al., 2024; Wang et al., 2024d) and leverage advanced prompting techniques, such as evolutionary algorithms (Gong et al., 2024), code generation (Hollmann et al., 2024), and Monte Carlo Tree Search (Zhang et al., 2024f), to create more complex and dynamic feature transformations. By automating these processes, LLMs reduce the reliance on manual intervention and enhance the efficiency and adaptability of feature engineering.

4.2 Data Cleaning

Data cleaning ensures that raw inputs are accurate and consistent, addressing issues such as missing and noisy values, duplicate entries, and inconsistencies across data sources. LLMs can help automate data cleaning tasks, including repairing and imputing missing values in text (Hassan et al., 2023; Bolding et al., 2023; Zhang et al., 2024c; Choi et al., 2024), structured (Yan et al., 2024b; Li et al., 2024j; Zhu et al., 2024a; Ding et al., 2024a; Luo et al., 2024), and semi-structured data (Jain et al., 2023; Mondal et al., 2024; Huang et al., 2024d; Biester et al., 2024; Ni et al., 2024). Their effectiveness is further enhanced by retrieval augmentation (Ahmad et al., 2023), code-driven methods (Huynh and Lin), and efficient tuning approaches (Zhang et al., 2024j). Beyond fixing errors, LLMs help filter out irrelevant or low-quality samples, improving dataset integrity and ensuring cleaner training data for downstream applications. They play a crucial role in maintaining the quality of synthetic data pipelines by automatically detecting and removing noisy or unreliable samples (Li et al., 2024c; Wang et al., 2024b; Yasunaga et al., 2024; Tong et al., 2024; Liang et al., 2024).

4.3 Data Transformation

Data transformation converts raw data into structured, analyzable formats, making it more suitable for modeling and analysis. LLMs facilitate data transformation by converting unstructured data into structured formats such as graphs (Li et al., 2024e; Zhou et al., 2024a), trees (Yuan et al., 2024a), and executable code (Rajkumar et al., 2022a; Sharma et al., 2023; Mayer et al., 2024). They also assist in handling cross-modal data transformations, such as image captioning (Hu et al., 2022) and text-to-image generation (Brade et al., 2023). By automating these transformations, LLMs reduce the manual effort required to preprocess diverse data types, enhancing efficiency and scalability in data pipelines.

4.4 Data Organization

Data organization ensures structured storage, management, and retrieval of information. LLMs enhance various aspects of data organization, including database tuning (Giannankouris and Trummer, 2024; Fan et al., 2024; Huang et al., 2024e; Li et al., 2024n; Lao et al.), query optimization (Akioyamen et al., 2024; Li et al., 2024g; Sun et al., 2024b), system diagnosis (Zhou et al., 2023b; Chen et al., 2024d; Singh et al., 2024; Giannakouris and Trummer, 2024), and pipeline orchestration (Hoseini et al., 2024; Shetty et al., 2024), leveraging their broad generalization capabilities (Fernandez et al., 2023; Weng et al., 2024a; Junior et al., 2024; Kim and Ailamaki, 2024; Li et al., 2024p). Recently, as LLMs continue to scale in parameter size, researchers have been exploring efficient memory management and storage strategies to optimize data organization (Bang, 2023; Li et al., 2024b; Xu et al., 2024a; Wang et al., 2024a; Yuan et al., 2024c; Kim et al., 2024c; Lee et al., 2024c). Additionally, efforts are also focused on improving vector storage for retrieval-augmented generation (RAG) applications, ensuring scalable and efficient retrieval (Zhang et al., 2023d; Jing et al., 2024; Tareaf et al., 2024; Rochan et al., 2024).

5 LLMs for Data Analysis

Data analysis is a crucial stage in the data science workflow, bridging data preparation and data interpretation to extract meaningful patterns and insights. Traditional data analysis methods depend on predefined statistical techniques and limited prior knowledge. In contrast, LLMs enhance edata analysis by leveraging their advanced reasoning and contextual awareness. LLMs can systematically interpret structured and unstructured data, uncover hidden relationships, and provide deeper insights beyond conventional approaches.

5.1 Descriptive Analysis

Descriptive analysis focuses on summarizing and exploring data patterns to extract meaningful patterns and insights. This step is fundamental to data analysis, serving as a foundation for more advanced inferential and predictive techniques. Traditional approaches, often relying on smaller models (e.g., BERT), struggle to capture complex semantics and contextual nuances (Jin et al., 2024a). More recently, LLMs have demonstrated exceptional capabilities in extracting valuable insights across diverse data types. Their ability to process large and unstructured datasets makes them particularly effective in uncovering patterns that might be difficult to detect using conventional methods.

Summarization. Recent research highlights the significant potential of LMs in extracting key insights from textual data. Direct extraction staregies aim to identify and retrieve key phrases or sentences directly from raw text (Viswanathan et al., 2023; Zhang et al., 2023e). LLMs have also demonstrated proficiency in abstractive summarization, such as generating TL;DR for academic papers (Zhang et al., 2024d), summarizing key sentences from documents and news articles (Zhao et al., 2023; Bražinskas et al., 2022), and obtaining descriptions for visual data (Yu et al., 2021) and tabular data (Zeng et al., 2024b). Notably, the effectiveness of these methods often relies heavily on prompt design. Recent advancements, such as automatic prompt discovery (Narayan et al., 2021), chain-of-thought (Wang et al., 2023e), and agent-based approaches (Xiao et al., 2023), have further enhanced LLMs' ability to generate structured and contextually rich summaries.

Exploratory Text Mining. Beyond summarization, LLMs enhance exploratory text mining for the discovery of latent patterns and insights from large text corpora. LLMs aid in topic modeling (Pham et al., 2023; Yang et al., 2024b; Mu et al., 2024) and sentiment analysis (Sun et al., 2023b; Xing, 2024), tackling challenges like incomplete topics and hallucinations. Furthermore, cross-lingual (Miah et al., 2024) and multimodal frameworks (Wang et al., 2024a; Yu et al., 2022) extend LLMs' scope beyond text, integrating diverse data modalities and multiple languages.

5.2 Analytical Reasoning

Analytical reasoning in data analysis involves using deductive techniques to extract meaningful insights from raw data. LLMs enhance this process by integrating pre-trained knowledge (e.g., commonsense and logical reasoning) with new data, enabling them to uncover high-level patterns and relationships across diverse modalities. For example, LLMs can effectively interpret patterns in complex textual data by understanding the contextual nuances (Chowdhery et al., 2023; Wyatte et al., 2024; Chae and Davidson, 2023) For images and videos, LLMs combined with vision encoders (e.g., CLIP (Radford et al., 2021), GPT-4V (Yang et al., 2023a)) exhibit strong reasoning capabilities in identifying objects, scenes, and abstract concepts (Wu et al., 2025; Cooper et al., 2024; Naeem et al., 2023). In audio analysis, LLMs enhance emotion recognition, speaker identification, and music genre detection when combined with audio embeddings (Zhang and Poellabauer, 2024; Dhingra et al., 2024; Meguenani et al., 2024; Li et al., 2021). Their ability to process graph-based data also extends to structural pattern recognition, transforming complex relationships into interpretable insights on graphs (Liu et al., a; Ye et al., 2023a; Guo et al., 2023; Srinivas and Runkana, 2024).

5.3 Interactive Analysis

Interactive analysis enables dynamic and user-driven data exploration by integrating iterative feedback, query-driven interpretation, and adaptive learning. LLMs enhance this process by engaging with data and user queries interactively, refining insights based on external feedback, and autonomously guiding analytical workflows. While traditional reinforcement learning can acquire experience from environmental rewards, it often lacks prior knowledge and exhibits poor generalization.

Feedback-Driven Analysis. LLMs leverage external annotations and user feedback to iteratively analyze data, improving accuracy and adaptability over time. LLMs have demonstrated the ability to reflect on external feedback for raw data (Hong et al., 2024; Ji et al., 2023b). Through memory retention and retrieval, LLMs store valuable analytical experiences, enabling more effective analysis of future data (Kim et al., 2024a; Shinn et al., 2024; Huang et al., 2022; Zhao et al., 2024a). These reflective capabilities extend beyond text, improving image generation(Yang et al., 2023b; Goswami et al., 2025) and personalized recommendation systems, where user feedback optimizes predictions and interactions over time(Xi et al., 2024; Wang et al., 2023f; Zhu et al., 2025).

Query-Driven Analysis. LLMs further support interactive analysis through query-answering interpretation, allowing users to extract insights from data intuitively. LLMs can answer questions on data across charts (Li et al., 2024r; Cheng et al., 2023; Han et al., 2023; Zhang et al., 2024a; Masry et al., 2023), tables (Li et al., 2024l; Zhou et al., 2024b; Zhu et al., 2024b), diagrams (Hu et al., 2024a; Wang et al., 2024d), and graphs (Xu et al., 2024b; Guo et al., 2023; Zhang et al., 2024i). For higher-level and more abstract questions, LLMs can dynamically select analytical techniques (Ma et al., 2023; Guo et al., 2024b; Zhu et al., 2024c; Liu et al., b; Zhang et al., 2023b) and autonomously generate exploration goals, queries, and interpretative answers without explicit prompts (Manatkar et al., 2024; Dibia, 2023).

5.4 Quantitative Analysis

Quantitative analysis focuses on extracting numerical insights from structured data using various techniques, including statistical methods, predictive modeling, and causal inference. LLMs generally enhance this process by automating calculations and improving model selection.

Statistical Analysis. LLMs assist in descriptive tasks, such as computing means and variances, across charts (Masry et al., 2023; Huang et al., 2024b; Liu et al., 2023a; Masry et al., 2024; Meng et al., 2024a; Do et al., 2023), tabular data (Brugere et al., 2024; Liu et al., 2023b), and time series (Jin et al., 2024b). Agent-based LLMs decompose multi-step calculations (Wang et al., 2023c; Huang et al., 2024c; Ye et al., 2023b; Wang et al., 2024h; Guo et al., 2024a) and invoke external tools and code for statistical modeling (Choe et al., 2024; Xia et al., 2023b; Wang et al., 2024f; Yuan et al., 2023b; Hong et al., 2024; Zhang et al., 2023b; Nascimento et al., 2024; Sun et al., 2024a). Moreover, LLMs facilitate hypothesis testing and correlation analysis (Qiu et al., 2024; Paruchuri et al., 2024; Zhu et al., 2024d; Liu et al., 2024c), offering interpretable approaches to statistical analysis.

Predictive Analysis. LLMs contribute to predictive modeling by assisting in the selection, development, and evaluation of models like linear regression and random forests (Nascimento et al., 2023; Junior et al., 2024; Hong et al., 2024). LLMs can also generate code for advanced tasks like time series forecasting (Morales-García et al., 2024; Ye et al., 2024), and adapt their reasoning across diverse data modalities, including text (Xiao et al., 2024b; Jacobs et al., 2024), time series (Jin et al., 2023; Chang et al., 2023; Jia et al., 2024a; Yu et al., 2023), charts (Masry et al., 2024; Wang et al., 2023g), tables (Yang et al., 2024c; Hamman et al., 2024; Wang et al., 2023d), and graphs (Wang et al.; Lin et al., 2024). Since LLMs are not intrinsically optimized for structured data, many frameworks combine LLMs with specialized deep learning techniques to enhance predictive performance (Zhang et al., 2024b; Bogahawatte et al., 2024; Moghadas et al., 2024; Liu et al., 2024a; Nam et al., 2024).

Causal Analysis. LLMs can uncover cause-and-effect relationships by refining Bayesian network structures and collaborating with causal discovery algorithms (Cohrs et al., 2024; Long et al., 2023; Ban et al., 2023b,a; Li et al., 2024p,k; Hu et al., 2024c; Liu et al., 2024c). Tool-augmented LLMs also invoke specialized causal discovery packages to improve inference (Shen et al., 2024).

6 Data Presentation

Data presentation is the final stage in the data science workflow, where processed and analyzed data is transformed into interpretable insights. This stage ensures that findings are effectively conveyed to users, enabling better comprehension. While traditional data presentation relies on manual scripting and specialized visualization tools, LLMs revolutionize the process by allowing users to generate and refine visual representations through natural language interaction, making data-driven insights more accessible and adaptable (Qin et al., 2020; Alvarez et al., 2021).

6.1 Human-Centric Visualization Generation

LLMs can output code, structural specifications, and chart queries based on user input to generate visualizations (Maddigan and Susnjak, 2023; Chen et al., 2024c; Ko et al., 2024; Li et al., 2024m; Wang et al., 2023b). Users provide dataset details and analysis goals, iteratively refining visual outputs when necessary (Tian et al., 2024; Dibia, 2023). Users can also interactively modify LLM-generated visualizations through adaptive interfaces (Chen et al., 2022).

6.2 Interactive Presentation

Discussion. Beyond single-instance visualizations, multi-agent systems extend LLM-driven visualization by enabling collaborative generation and refinement (Chugh et al., 2023; Guan et al., 2024). For instance, one agent may write or debug visualization scripts (ALMutairi et al., 2024), while another creates graphs or serves as a task planner (Islam et al., 2024; Li et al., 2024i; Xue et al., 2024). LLM can also convert natural language instructions into structured commands for visualization agents (Huang et al., 2024a).

Human-Driven Decision-Making. LLMs assist by offering expert-like suggestions or collaborating with human users (Ma et al., 2024; Cao et al., 2024a). In healthcare, they can improve diagnostic accuracy while aligning with professional guidelines (Umerenkov et al., 2023; Benary et al., 2023; Goh et al., 2023; Eigner and Händler, 2024).

7 Social Impact and Future Work

Despite the remarkable capabilities, LLMs also exhibit crucial shortcomings that can hinder data science projects. In this section, we introduce the ethics concerns, limitations, and future work of LLMs in data science.

7.1 Ethics Consideration

Bias & Fairness. LLMs can embed implicit biases that skew results and violate fairness principles (Dai et al., 2024a; Gallegos et al., 2024; Li et al., 2023c). For instance, LLM-generated content may be over-ranked by certain IR or RAG systems (Dai et al., 2024b; Chen et al., 2024e; Bao et al., 2023), while inaccuracies, irrelevancies, or instruction deviations go unchecked (Liu et al., 2024b; Lee et al., 2022; Min et al., 2023; Durmus et al., 2020; Maynez et al., 2020). Moreover, LLM-based evaluators have displayed favoritism toward specific tokens or group attributes (Hou et al., 2024; Chen et al., 2024b; Wang et al., 2024c; Liu et al., 2023c; Zhang et al., 2024k), necessitating careful mitigation strategies.

Privacy. LLMs pose privacy risks by potentially exposing sensitive user data in prompts or inferred attributes (Yan et al., 2024a; Neel and Chang, 2024; Das et al., 2024a). Cloud-hosted services may log personally identifiable information (PII)(Iqbal et al., 2024), and the models themselves can accurately infer unobserved traits such as occupation or location(Staab et al., 2024). Recent surveys highlight the roles of model size, data duplication, and prompt length in such leaks (Neel and Chang, 2024), prompting research into differential privacy (Dwork, 2006), federated learning (McMahan et al., 2017), multi-level privacy evaluations (Shao et al., 2024), and privacy-preserving synthetic data generation (Wang et al., 2024e; Ramesh et al., 2024) to mitigate these threats.

7.2 Limitations and Future Work

Sensitivity to Dataset Formats. LLMs often exhibit inconsistent performance across different serialization formats, making it hard to automate tasks on tabular or semi-structured data. Some studies show superior fact-searching accuracy when data is presented as DFLoader or JSON, yet better table-specific tasks (e.g., partitioning, cell lookup) emerge in HTML or XML formats—especially notable for GPT-4 (Fang et al., 2024; Singha et al., 2023; Sui et al., 2024; OpenAI, 2024a). These findings reveal that minor changes in data format alone can lead to sizable variations in LLM performance, complicating development pipelines.

Reasoning over Numerical Operations. LLMs are also known to have reasoning limitations, especially over numerics. Competitive LLMs such as GPT-4o (OpenAI, 2024b) can easily output incorrect number comparison $9.11 > 9.9$ if the generation order between thought and conclusion is reversed (Xie, 2024), while other evidences show LLMs easily making mistakes on symbolic operations when processing tabular data (Chen, 2023). These LLMs can also frequently struggle with primitive operations such as letter counting within a word, whose error rate strongly correlates with the total (and individual) number of tokens in a word (Fu et al., 2024) and tokenization design (Zhang et al., 2024e).

Test-time Overthinking. Models specialized with extended test time compute or long reasoning such as OpenAI O1 (OpenAI, 2024c), QWQ (Qwen, 2024) and DeepSeek R1 (DeepSeek, 2024) are known to overthink on simpler mathematic questions, causing unnecessarily long reasoning chain and incorrect results that cannot be addressed by length-adjusted preference optimization (Meng et al., 2024b) and only partially mitigated by solution filtering in conjunction with reflection (Chen et al., 2024f).

Token Inefficiency. When raw tables become large (even as small as 30 rows), LLMs often fail to parse the data effectively, leading to inflated token usage and increased computational costs (Chen, 2023; Fang et al., 2024). This token inefficiency often forces external execution strategies, such as Python scripts (Chen et al., 2023b; Gao et al., 2023), SQL queries (Rajkumar et al., 2022b), verifier programs (Ni et al., 2023), or dynamic decomposition (Ye et al., 2023c; Wang et al., 2024g), to reduce context size and limit hallucinations (Ji et al., 2023a).

Tokenizer-Free Representation. Conventional tokenization methods (Schuster and Nakajima, 2012; Sennrich et al., 2016; Kudo and Richardson, 2018) often split numbers into multiple fragments (e.g., "digit chunking"), encouraging memorization rather than true algorithmic processing (Spathis and Kawsar, 2024). Systems like Byte Pair Encoding (BPE) (Sennrich et al., 2016) can produce inconsistently tokenized numerals, whereas LLaMA retains numbers intact (Fang et al., 2024). Recently, Byte Latent Transformers (BLT) (Pagnoni et al., 2024) propose a more flexible approach by treating tokenization as an inference-time byte-grouping problem driven by lightweight encoding-decoding models and byte-level entropies, rather than a fixed, pre-trained vocabulary. This shift promises greater adaptability and performance comparable to conventional tokenizers.

8 Conclusion

In this survey, we examined the role of large language models (LLMs) in enhancing data science workflows, focusing on their applications across various stages, from data acquisition to analysis and presentation. While LLMs offer significant potential to automate and streamline tasks, challenges such as model reliability, data quality, and ethical concerns like bias and privacy risks remain. Future research should aim to address these limitations by improving model robustness, interpretability, and integration with traditional methods. Ultimately, LLMs can transform data science practices, offering more efficient, accessible, and automated solutions for a wide range of industries and domains.

Limitation

While this survey provides a broad overview of the role of LLMs in data science, there are several limitations to consider. First, the rapid pace of developments in LLMs means that some of the discussed techniques, applications, and challenges may quickly become outdated. Additionally, the survey primarily focuses on high-level applications and concepts, which may not capture the full technical depth or domain-specific nuances of LLM usage in data science. Furthermore, given the vast scope of data science, the survey may not address every potential application of LLMs in all subfields, and certain interdisciplinary applications might have been underrepresented. Finally, due to the complexity of LLMs, certain challenges such as model biases and ethical implications may require more focused, in-depth exploration than what this survey can provide.

Ethics Statement

This survey acknowledges the importance of ethical considerations in the use of LLMs within data science. The applications discussed in this work are subject to potential ethical concerns, including but not limited to bias in model predictions, privacy risks, and fairness in decision-making. LLMs, being trained on vast and often uncurated datasets, may unintentionally perpetuate societal biases, which can influence data science workflows and lead to inequitable outcomes. Furthermore, the use of LLMs in sensitive domains, such as healthcare and finance, requires strict adherence to ethical guidelines to safeguard user privacy and ensure transparency in automated decision-making processes. Researchers and practitioners are encouraged to prioritize the ethical implications of deploying LLMs in data science applications and to strive for solutions that mitigate bias, enhance model interpretability, and uphold privacy and fairness standards. This survey aims to raise awareness of these challenges and advocates for continued research into the responsible use of LLMs in data science.

References

Danial Abshari, Chenglong Fu, and Meera Sridhar. 2024. Llm-assisted physical invariant extraction for cyber-physical systems anomaly detection. *arXiv preprint arXiv:2411.10918*.

Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Eltabakh, Mourad Ouzzani, and Nan Tang. 2023. Retclean: Retrieval-based data cleaning using foundation models and data lakes. *arXiv preprint arXiv:2303.16909*.

Peter Akioyamen, Zixuan Yi, and Ryan Marcus. 2024. The unreasonable effectiveness of llms for query optimization. *arXiv preprint arXiv:2411.02862*.

Mohammad Al-Smadi. 2023. Chatgpt and beyond: The generative ai revolution in education. *arXiv preprint arXiv:2311.15198*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Bashar Alhafni, Sowmya Vajjala, Stefano Banno, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. Llms in education: Novel perspectives, challenges, and opportunities. *ArXiv*, abs/2409.11917.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):1–25.

Mariam ALMutairi, Lulwah AlKulaib, Shengkun Wang, Zhiqian Chen, Youssif ALMutairi, Thamer M Alenazi, Kurt Luther, and Chang-Tien Lu. 2024. Fhirviz: Multi-agent platform for fhir visualization to advance healthcare analytics. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–7.

Pavel Anselmo Alvarez, Alessio Ishizaka, and Luis Martínez. 2021. Multiple-criteria decision-making sorting methods: A survey. *Expert Systems with Applications*, 183:115368.

Tuo An, Yunjiao Zhou, Han Zou, and Jianfei Yang. 2024. Iot-llm: Enhancing real-world iot task reasoning with large language models. *arXiv preprint arXiv:2410.02429*.

Muhammad Arslan and Christophe Cruz. 2024. Business-rag: Information extraction for business insights. *ICSBT 2024*, page 88.

Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, and Huanhuan Chen. 2023a. Causal structure learning supervised by large language model. *arXiv preprint arXiv:2311.11689*.

Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023b. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902*.

Fu Bang. 2023. Gptcache: An open-source semantic cache for llm applications enabling faster answers and cost savings. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 212–218.

Yair Bannett, Fatma Gunturkun, Malvika Pillai, Jessica E Herrmann, Ingrid Luo, Lynne C. Huffman, and Heidi M. Feldman. 2024. Applying large language models to assess quality of care: Monitoring adhd medication side effects. *Pediatrics*, 155 1.

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *arXiv preprint arXiv:2306.15766*.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer.

Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, et al. 2023. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689–e2343689.

Yifan Bian, Dennis Küster, Hui Liu, and Eva G. Krumhuber. 2023. Understanding naturalistic facial expressions with deep learning and multimodal large language models. *Sensors (Basel, Switzerland)*, 24.

Fabian Biester, Mohamed Abdelaal, and Daniel Del Gaudio. 2024. Llmclean: Context-aware tabular data cleaning via llm-generated ofds. In *European Conference on Advances in Databases and Information Systems*, pages 68–78. Springer.

Elizabeth Bloodgood. 2025. *Data Aggregation: An Overview of Opportunities*. Global Press.

Jayanie Bogahawatte, Sachith Seneviratne, Maneesha Perera, and Saman Halgamuge. 2024. Rethinking time series forecasting with llms via nearest neighbor contrastive learning. *arXiv preprint arXiv:2412.04806*.

Quinten Bolding, Baohao Liao, Brandon Denis, Jun Luo, and Christof Monz. 2023. Ask language model to clean your noisy translation data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3215–3236.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14.

Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525 – 535.

Arthur Bražinskas, Ramesh Nallapati, Mohit Bansal, and Markus Dreyer. 2022. Efficient few-shot fine-tuning for opinion summarization. *arXiv preprint arXiv:2205.02170*.

Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. 2023. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Ivan Brugere, Shubham Sharma, Sanjay Kariyappa, Anh Totti Nguyen, Freddy Lecue, et al. 2024. Interpretable llm-based table question answering. *arXiv preprint arXiv:2412.12386*.

Field Cady. 2024. *The data science handbook*. John Wiley & Sons.

Charles Cao, Jie Zhuang, and Qiang He. 2024a. Llm-assisted modeling and simulations for public sector decision-making: Bridging climate data and policy insights. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *ArXiv*, abs/2311.16208.

Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, KP Subbalakshmi, and Papa Momar Ndiaye. 2024b. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*.

Yupeng Cao, Zhi Chen, Qingyun Pei, Prashant Kumar, K.P. Subbalakshmi, and Papa Momar Ndiaye. 2024c. Ecc analyzer: Extract trading signal from earnings conference calls using large language model for stock performance prediction. *ArXiv*, abs/2404.18470.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

Youngjin Chae and Thomas Davidson. 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

Joymallya Chakraborty, Wei Xia, Anirban Majumder, Dan Ma, Walid Chaabene, and Naveed Janvekar. 2024. Detoxbench: Benchmarking large language models for multitask fraud & abuse detection. *arXiv preprint arXiv:2409.06072*.

Ching Chang, Wen-Chih Peng, and Tien-Fu Chen. 2023. Llm4ts: Two-stage fine-tuning for time-series forecasting with pre-trained llms. *arXiv preprint arXiv:2308.08469*.

Jia-Wei Chang, Neil Yen, and Jason Chi-Shun Hung. 2022. Design of a nlp-empowered finance fraud awareness model: the anti-fraud chatbot for fraud detection and fraud classification as an instance. *Journal of Ambient Intelligence and Humanized Computing*, 13:4663 – 4679.

Chaoran Chen, Daodao Zhou, Yanfang Ye, Toby Li, and Yaxing Yao. 2024a. Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. *ArXiv*, abs/2410.13387.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. 2024c. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*.

Shengchao Chen, Guodong Long, Jing Jiang, Dikai Liu, and Chengqi Zhang. 2023a. Foundation models for weather and climate data understanding: A comprehensive survey. *arXiv preprint arXiv:2312.03014*.

Sibei Chen, Ju Fan, Bin Wu, Nan Tang, Chao Deng, Pengyi Wang, Ye Li, Jian Tan, Feifei Li, Jingren Zhou, et al. 2024d. Automatic database configuration debugging using retrieval-augmented language models. *arXiv preprint arXiv:2412.07548*.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Xiaoyang Chen, Ben He, Hongyu Lin, Xianpei Han, Tianshu Wang, Boxi Cao, Le Sun, and Yingfei Sun. 2024e. Spiral of silence: How is large language model killing information retrieval?—A case study on open domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14930–14951, Bangkok, Thailand. Association for Computational Linguistics.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. 2024f. Do not think that much for 2+3=? on the overthinking of o1-like llms.

Yi Chen, Hanming Fang, Yi Zhao, and Zibo Zhao. 2024g. Recovering overlooked information in categorical variables with llms: An application to labor market mismatch. Technical report, National Bureau of Economic Research.

Yiru Chen, Ryan Li, Austin Mac, Tianbao Xie, Tao Yu, and Eugene Wu. 2022. Nl2interface: Interactive visualization interface generation from natural language queries. *arXiv preprint arXiv:2209.08834*.

Yiru Chen et al. 2023c. Interactive visualization from natural language. *Transactions on Visualization*.

Yu Chen and Mei Si. 2024. Llm-based reflections and enhancements. *International Journal of Computational Linguistics*.

Zihan Chen, Song Wang, Cong Shen, and Jundong Li. 2024h. Fastgas: Fast graph-based annotation selection for in-context learning. In *ACL*.

Liying Cheng, Xingxuan Li, and Lidong Bing. 2023. Is gpt-4 a good data analyst? *arXiv preprint arXiv:2305.15038*.

Kiroong Choe, Chaerin Lee, Soohyun Lee, Jiwon Song, Aeri Cho, Nam Wook Kim, and Jinwook Seo. 2024. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics*.

Juhwan Choi et al. 2024. Llm-driven feature engineering. *Machine Learning Journal*, 35.

Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. Lmpriors: Pre-trained language models as task-specific priors. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Tushar Chugh, Kanishka Tyagi, Rolly Seth, and Pranesh Srinivasan. 2023. Intelligent agents driven data analytics using large language models. In *2023 International Conference on Artificial Intelligence, Blockchain, Cloud Computing, and Data Analytics (ICoABCD)*, pages 152–157. IEEE.

Kai-Hendrik Cohrs, Gherardo Varando, Emiliano Diaz, Vasileios Sitokonstantinou, and Gustau Camps-Valls. 2024. Large language models for constrained-based causal discovery. *arXiv preprint arXiv:2406.07378*.

Avi Cooper, Keizo Kato, Chia-Hsien Shih, Hiroaki Yamane, Kasper Vinken, Kentaro Takemoto, Taro Sunagawa, Hao-Wei Yeh, Jin Yamanaka, Ian Mason, et al. 2024. Rethinking vlms and llms for image classification. *arXiv preprint arXiv:2410.14690*.

Hongwei Cui, Yuyang Du, Qun Yang, Yulin Shao, and Soung Chang Liew. 2024. Llmind: Orchestrating ai and iot with llm for complex task execution. *IEEE Communications Magazine*.

Sunhao Dai and colleagues. 2024. Bias in large language models. *KDD Proceedings*.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024a. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. 2024b. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 526–537, New York, NY, USA. Association for Computing Machinery.

Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2024a. Security and privacy challenges of large language models: A survey.

Debarati Das, Karin De Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024b. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

DeepSeek. 2024. Deepseek-r1-lite-preview is now live: Unleashing supercharged reasoning power. `https://api-docs.deepseek.com/news/news1120`. Accessed: 2025-01-05.

Priyanshu Dhingra, Satyam Agrawal, Chandra Sekar Veerappan, Eng Siong Chng, and Rong Tong. 2024. Enhancing speech de-identification with llm-based data augmentation. In *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–5. IEEE.

Victor Dibia. 2023. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. 2024a. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024b. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.06211*.

Xuan Long Do, Mohammad Hassanpour, Ahmed Masry, Parsa Kavehzadeh, Enamul Hoque, and Shafiq Joty. 2023. Do llms work on charts? designing few-shot prompts for chart question answering and summarization. *arXiv preprint arXiv:2312.10610*.

Rian Dolphin, Joe Dursun, Jonathan Chow, Jarrett Blankenship, Katie Adams, and Quinton Pike. 2024. Extracting structured insights from financial news: An augmented llm driven approach. *arXiv preprint arXiv:2407.15788*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shiran Dudy, Ibrahim Said Ahmad, Ryoko Kitajima, and Agata Lapedriza. 2024. Analyzing cultural representations of emotions in llms through mixed emotion survey. *ArXiv*, abs/2408.02143.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, page 1–12, Berlin, Heidelberg. Springer-Verlag.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Chongjiong Fan, Zhicheng Pan, Wenwen Sun, Chengcheng Yang, and Wei-Neng Chen. 2024. Latuner: An llm-enhanced database tuning system based on adaptive surrogate model. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 372–388. Springer.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models

(LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

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

Raul Castro Fernandez, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan. 2023. How large language models will disrupt data management. *Proc. VLDB Endow.*, 16:3302–3309.

Tairan Fu, Raquel Ferrando, Javier Conde, Carlos Arriaga, and Pedro Reviriego. 2024. Why do large language models (llms) struggle to count letters?

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. In *2023 IEEE international conference on big data (BigData)*, pages 4776–4785. IEEE.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. Worldgpt: Empowering llm as multimodal world model. *arXiv preprint arXiv:2404.18202*.

Victor Giannakouris and Immanuel Trummer. 2024. Dbg-pt: A large language model assisted query performance regression debugger. *Proc. VLDB Endow.*, 17:4337–4340.

Victor Giannankouris and Immanuel Trummer. 2024. {\lambda}-tune: Harnessing large language models for automated database system tuning. *arXiv preprint arXiv:2411.03500*.

Khasa Gillani, Erik Novak, Klemen Kenda, and Dunja Mladenić. 2024. Knowledge graph extraction from textual data using llm.

Daniel Goedde, Sophia Noehl, Carina Wolf, Yannick Rupert, Lukas Rimkus, Jan Ehlers, Frank Breuckmann, and Timur Sellmann. 2023. A swot (strengths, weaknesses, opportunities, and threats) analysis of chatgpt in the medical literature: concise review. *Journal of Medical Internet Research*, 25:e49368.

Laura H Goetz and Nicholas J Schork. 2018. Personalized medicine: motivation, challenges, and progress. *Fertility and sterility*, 109(6):952–963.

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

Nanxu Gong, Chandan K Reddy, Wangyang Ying, and Yanjie Fu. 2024. Evolutionary large language model for automated feature transformation. *arXiv preprint arXiv:2405.16203*.

Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. Plotgen: Multi-agent llm-based scientific data visualization via multimodal feedback. *arXiv preprint arXiv:2502.00988*.

Gadzinski Gregory and Liuzzi Vito. 2024. Chatgpt: a canary in the coal mine or a parrot in the echo chamber? detecting fraud with llm: the case of ftx. *Finance Research Letters*.

Qinghua Guan, Jinhui Ouyang, Di Wu, and Weiren Yu. 2024. Citygpt: Towards urban iot learning, analysis and interaction with multi-agent system. *arXiv preprint arXiv:2405.14691*.

Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023. Gpt4graph: Can large language models understand graph structured data? an empirical evaluation and benchmarking. *arXiv preprint arXiv:2305.15066*.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024a. Ds-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*.

Yi Guo, Danqing Shi, Mingjuan Guo, Yanqiu Wu, Nan Cao, and Qing Chen. 2024b. Talk2data: A natural language interface for exploratory visual analysis via question decomposition. *ACM Transactions on Interactive Intelligent Systems*, 14(2):1–24.

Anil Gur and colleagues. 2022. Data integration using llms. *Data Journal*, 28.

Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2022. Understanding html with large language models. *arXiv preprint arXiv:2210.03945*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Faisal Hamman, Pasan Dissanayake, Saumitra Mishra, Freddy Lecue, and Sanghamitra Dutta. 2024. Quantifying prediction consistency under model multiplicity in tabular llms. *arXiv preprint arXiv:2407.04173*.

Sungwon Han, Jinsung Yoon, Sercan O Arik, and Tomas Pfister. Large language models can automatically engineer features for few-shot tabular learning. In *Forty-first International Conference on Machine Learning*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173.

Jakub Harasta, Tereza Novotn’a, and Jaromír avelka. 2024. It cannot be right if it was written by ai: On lawyers’ preferences of documents perceived as authored by an llm vs a human. *ArXiv*, abs/2407.06798.

Karlyn Harrod, Prabin Bhandari, and Antonios Anastasopoulos. 2024. From text to maps: Llm-driven extraction and geotagging of epidemiological data. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 258–270.

Md Mahadi Hassan, Alex Knipper, and Shubhra Kanti Karmaker Santu. 2023. Chatgpt as your personal data scientist. *arXiv preprint arXiv:2305.13657*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2025. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963.

Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss B Cook, and Jingrui He. 2024. Llm-forest for health tabular data imputation. *arXiv preprint arXiv:2410.21520*.

Noah Hollmann, Samuel Müller, and Frank Hutter. 2024. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems*, 36.

Sirui Hong, Yizhang Lin, Bang Liu, Bangbang Liu, Binhao Wu, Ceyao Zhang, Chenxing Wei, Danyang Li, Jiaqi Chen, Jiayi Zhang, et al. 2024. Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679*.

Sayed Hoseini, Maximilian Ibbels, and Christoph Quix. 2024. Enhancing machine learning capabilities in data lakes with automl and llms. In *European Conference on Advances in Databases and Information Systems*, pages 184–198. Springer.

Mohammad Hosseini, Catherine A Gao, David M Liebovitz, Alexandre M Carvalho, Faraz S Ahmad, Yuan Luo, Ngan MacDonald, Kristi L Holmes, and Abel Kho. 2023. An exploratory survey about using chatgpt in education, healthcare, and research. *Plos one*, 18(10):e0292216.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II*, page 364–381, Berlin, Heidelberg. Springer-Verlag.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024a. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6929–6938.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17980–17989.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024b. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

Zhizhang Hu, Yue Zhang, Ryan Rossi, Tong Yu, Sungchul Kim, and Shijia Pan. 2024c. Are large language models capable of causal reasoning for sensing data analysis? In *Proceedings of the Workshop on Edge and Mobile Foundation Models*, pages 24–29.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, et al. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

Huazhen Huang, Xianguo Shi, Hongyang Lei, Fan Hu, and Yunpeng Cai. 2024a. Protchat: An ai multi-agent for automated protein analysis leveraging gpt-4 and protein language model. *Journal of Chemical Information and Modeling*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Muye Huang, Lai Han, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. 2024b. Evochart: A benchmark and a self-training approach towards real-world chart understanding. *arXiv preprint arXiv:2409.01577*.

Muye Huang, Lingling Zhang, Lai Han, Wenjun Wu, Xinyu Zhang, and Jun Liu. 2024c. Vprochart: Answering chart question through visual perception alignment agent and programmatic solution reasoning. *arXiv preprint arXiv:2409.01667*.

Xiaobao Huang, Mihir Surve, Yuhan Liu, Tengfei Luo, Olaf Wiest, Xiangliang Zhang, and Nitesh V Chawla. 2024d. Application of large language models in chemistry reaction data extraction and cleaning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3797–3801.

Xinmei Huang, Haoyang Li, Jing Zhang, Xinxin Zhao, Zhiming Yao, Yiyan Li, Zhuohao Yu, Tieying Zhang, Hong Chen, and Cuiping Li. 2024e. Llmtune: Accelerate database knob tuning with large language models. *arXiv preprint arXiv:2404.11581*.

Nam Huynh and Beiyu Lin. Llms automatically generate code to data clean.

Umar Iqbal, Tadayoshi Kohno, and Franziska Roesner. 2024. Llm platform security: Applying a systematic evaluation framework to openai's chatgpt plugins. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*.

Yusei Ishimizu, Jialong Li, Takuto Yamauchi, Sinan Chen, Jinyu Cai, Takanori Hirano, and Kenji Tei. 2024. Towards efficient discrete controller synthesis: Semantics-aware stepwise policy design via llm. *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pages 1–4.

Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Datanarrative: Automated data-driven storytelling with visualizations and texts. *arXiv preprint arXiv:2408.05346*.

Ryan Jacobs, Maciej P Polak, Lane E Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. 2024. Regression with large language models for materials and molecular property prediction. *arXiv preprint arXiv:2409.06080*.

Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E Gonzalez, Koushik Sen, and Ion Stoica. 2023. Llm-assisted code cleaning for training accurate code generators. *arXiv preprint arXiv:2311.14904*.

Daniel P Jeong, Zachary C Lipton, and Pradeep Ravikumar. 2024. Llm-select: Feature selection with large language models. *arXiv preprint arXiv:2407.02694*.

Paulo Jesus, Carlos Baquero, and Paulo Sérgio Almeida. 2014. A survey of distributed data aggregation algorithms. *IEEE Communications Surveys & Tutorials*, 17(1):381–404.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating llm hallucination via self reflection. In *EMNLP*, pages 1827–1843.

Furong Jia, Kevin Wang, Yixiang Zheng, Defu Cao, and Yan Liu. 2024a. Gpt4mts: Prompt-based large language model for multimodal time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23343–23351.

Pengyue Jia, Zhaocheng Du, Yichao Wang, Xiangyu Zhao, Xiaopeng Li, Yuhao Wang, Qidong Liu, Huifeng Guo, and Ruiming Tang. 2024b. Altfs: Agency-light feature selection with large language models in deep recommender systems. *arXiv preprint arXiv:2412.08516*.

Bohan Jiang, Chengshuai Zhao, Zhen Tan, and Huan Liu. 2024. Catching chameleons: Detecting evolving disinformation generated using large language models. *ArXiv*, abs/2406.17992.

Julie Jiang and Emilio Ferrara. 2023. Social-llm: Modeling user behavior at scale using language models and social network data. *arXiv preprint arXiv:2401.00893*.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024a. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024b. Position: What can large language models tell us about time series analysis. In *Forty-first International Conference on Machine Learning*.

Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. 2024. When large language models meet vector databases: A survey. *arXiv preprint arXiv:2402.01763*.

Rogers Jeffrey Leo John, Dylan Bacon, Junda Chen, Ushmal Ramesh, Jiatong Li, Deepan Das, Robert Claus, Amos Kendall, and Jignesh M Patel. 2023. Datachat: An intuitive and collaborative data analytics platform. In *Companion of the 2023 International Conference on Management of Data*, pages 203–215.

Sylvio Barbon Junior, Paolo Ceravolo, Sven Groppe, Mustafa Jarrar, Samira Maghool, Florence Sèdes, Soror Sahri, and Maurice Van Keulen. 2024. Are large language models the new interface for data pipelines? *arXiv preprint arXiv:2406.06596*.

Rishi Kalra, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Soares Koshiyama, and Philip C. Treleaven. 2024. Hypa-rag: A hybrid parameter adaptive retrieval-augmented generation system for ai legal and policy applications. *ArXiv*, abs/2409.09046.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Karina Kasztelnik and Edo Branch. 2024. Insights from large language model analysis on black women in accounting leadership in the u.s. *Journal of Accounting and Finance*.

Uri Katz, Mosh Levy, and Yoav Goldberg. 2024. Knowledge navigator: Llm-guided browsing framework for exploratory search in scientific literature. *arXiv preprint arXiv:2408.15836*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Vijay Khatri and Carol Brown. 2010a. *Data Management: Best Practices*. Information Systems Press.

Vijay Khatri and Carol V Brown. 2010b. Designing data governance. *Communications of the ACM*, 53(1):148–152.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2024a. Language models can solve computer tasks. In *NeurIPS*.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024b. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

Hyebin Kim and Sugie Lee. 2024. Poi gpt: Extracting poi information from social media text data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:113–118.

Kyoungmin Kim and Anastasia Ailamaki. 2024. Trustworthy and efficient llms meet databases. *arXiv preprint arXiv:2412.18022*.

Kyoungmin Kim, Kijae Hong, Caglar Gulcehre, and Anastasia Ailamaki. 2024c. The effect of scheduling and preemption on the efficiency of llm inference serving. *arXiv preprint arXiv:2411.07447*.

Mohamed Bayan Kmainasi, Ali Ezzat Shahroor, Maram Hasanain, Sahinur Rahman Laskar, Naeemul Hassan, and Firoj Alam. 2024. Llamalens: Specialized multilingual llm for analyzing news and social media content. *arXiv preprint arXiv:2410.15308*.

Hyung-Kwon Ko, Hyeon Jeon, Gwanmo Park, Dae Hyun Kim, Nam Wook Kim, Juho Kim, and Jinwook Seo. 2024. Natural language dataset generation framework for visualizations powered by large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Ibrahim Kok, Orhan Demirci, and Suat Ozdemir. 2024. When iot meet llms: Applications and challenges. *arXiv preprint arXiv:2411.17722*.

Mathias Kraus, Julia Anna Bingler, Markus Leippold, Tobias Schimanski, Chiara Colesanti Senni, Dominik Stammbach, Saeid Ashraf Vaghefi, and Nicolas Webersinke. 2023. Enhancing large language models with climate resources. *arXiv preprint arXiv:2304.00116*.

Jinxi Kuang, Jinyang Liu, Junjie Huang, Renyi Zhong, Jiazhen Gu, Lan Yu, Rui Tan, Zengyin Yang, and Michael R Lyu. 2024. Knowledge-aware alert aggregation in large-scale cloud systems: a hybrid approach. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Practice*, pages 369–380.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jaris Küken, Lennart Purucker, and Frank Hutter. 2024. Large language models engineer too many simple features for tabular data. *arXiv preprint arXiv:2410.17787*.

Litton J Kurisinkel and Nancy F Chen. 2023. Llm based multi-document summarization exploiting main-event biased monotone submodular content extraction. *arXiv preprint arXiv:2310.03414*.

Andrew Kusiak. 2001. Feature transformation methods in data mining. *IEEE Transactions on Electronics packaging manufacturing*, 24(3):214–221.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.

Jiale Lao, Yibo Wang, Yufei Li, Jianping Wang, Yunjia Zhang, Zhiyuan Cheng, Wanghu Chen, Mingjie Tang, and Jianguo Wang. Gptuner: A manual-reading database tuning system via gpt-guided bayesian optimization.

Dexter Le, Aybars Yunusoglu, Karn Tiwari, Murat Isik, and I Dikmen. 2024. Multimodal llm for intelligent transportation systems. *arXiv preprint arXiv:2412.11683*.

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

Joseph Lee, Shu Yang, Jae Young Baik, Xiaoxi Liu, Zhen Tan, Dawei Li, Zixuan Wen, Bojian Hou, Duy Duong-Tran, Tianlong Chen, et al. 2024b. Knowledge-driven feature selection and engineering for genotype data with large language models. *arXiv preprint arXiv:2410.01795*.

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

Wonbeom Lee, Jungi Lee, Junghwan Seo, and Jaewoong Sim. 2024c. {InfiniGen}: Efficient generative inference of large language models with dynamic {KV} cache management. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pages 155–172.

Juho Leinonen, Stephen MacNeil, Paul Denny, and Arto Hellas. 2024. Using large language models for teaching computing. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2*, pages 1901–1901.

Aodong Li, Yunhan Zhao, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. 2024a. Anomaly detection of tabular data using llms. *arXiv preprint arXiv:2406.16308*.

Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024b. Llm inference serving: Survey of recent advances and opportunities. *arXiv preprint arXiv:2407.12391*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024c. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Dawei Li, Zhen Tan, and Huan Liu. 2024d. Exploring large language models for feature selection: A data-centric perspective. *arXiv preprint arXiv:2408.12025*.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, et al. 2024e. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*.

Diya Li, Asim Kadav, Aijing Gao, Rui Li, and Richard Bourgon. 2024f. Automated clinical data extraction with knowledge conditioned llms. *arXiv preprint arXiv:2406.18027*.

Guoliang Li, Xuanhe Zhou, and Xinyang Zhao. 2024g. Llm for data management. *Proceedings of the VLDB Endowment*, 17(12):4213–4216.

Hang Li, Yu Kang, Tianqiao Liu, Wenbiao Ding, and Zitao Liu. 2021. Ctal: Pre-training cross-modal transformer for audio-and-language representations. *arXiv preprint arXiv:2109.00181*.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024h. Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Jinyang Li, Nan Huo, Yan Gao, Jiayi Shi, Yingxiu Zhao, Ge Qu, Yurong Wu, Chenhao Ma, Jian-Guang Lou, and Reynold Cheng. 2024i. Tapilot-crossing: Benchmarking and evolving llms towards interactive data analysis agents. *arXiv preprint arXiv:2403.05307*.

Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. 2017. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45.

Lan Li, Liri Fang, and Vetle I Torvik. 2024j. Autodcworkflow: Llm-based data cleaning workflow auto-generation and benchmark. *arXiv preprint arXiv:2412.06724*.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023a. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Peiwen Li, Xin Wang, Zeyang Zhang, Yuan Meng, Fang Shen, Yue Li, Jialong Wang, Yang Li, and Wenwu Zhu. 2024k. Realtcd: Temporal causal discovery from interventional data with large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4669–4677.

Qianlong Li, Chen Huang, Shuai Li, Yuanxin Xiang, Deng Xiong, and Wenqiang Lei. 2024l. Graphotter: Evolving llm-based graph reasoning for complex table question answering. *arXiv preprint arXiv:2412.01230*.

Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023b. Adapting large language models for education: Foundational capabilities, potentials, and challenges. *ArXiv*, abs/2401.08664.

Shuaimin Li, Xuanang Chen, Yuanfeng Song, Yunze Song, and Chen Zhang. 2024m. Prompt4vis: Prompting large language models with example mining and schema filtering for tabular data visualization. *arXiv preprint arXiv:2402.07909*.

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

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023d. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382.

Yiyan Li, Haoyang Li, Zhao Pu, Jing Zhang, Xinyi Zhang, Tao Ji, Luming Sun, Cuiping Li, and Hong Chen. 2024n. Is large language model good at database knob tuning? a comprehensive experimental evaluation. *arXiv preprint arXiv:2408.02213*.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. 2024o. Chatcite: Llm agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*.

Zhaodonghui Li, Haitao Yuan, Huiming Wang, Gao Cong, and Lidong Bing. 2024p. Llm-r2: A large language model enhanced rule-based rewrite system for boosting query efficiency. *arXiv preprint arXiv:2404.12872*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024q. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, et al. 2024r. Autokaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*.

Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, et al. 2024. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. *arXiv preprint arXiv:2408.08072*.

Tianqianjin Lin, Pengwei Yan, Kaisong Song, Zhuoren Jiang, Yangyang Kang, Jun Lin, Weikang Yuan, Junjie Cao, Changlong Sun, and Xiaozhong Liu. 2024. Langgfm: A large language model alone can be a powerful graph foundation model. *arXiv preprint arXiv:2410.14961*.

Roderick JA Little and Donald B Rubin. 2019. *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Chenxi Liu, Sun Yang, Qianxiong Xu, Zhishuai Li, Cheng Long, Ziyue Li, and Rui Zhao. 2024a. Spatial-temporal large language model for traffic prediction. *arXiv preprint arXiv:2401.10134*.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. a. One for all: Towards training one graph model for all classification tasks. In *The Twelfth International Conference on Learning Representations*.

Haoyang Liu, Shuyu Chen, and Haohan Wang. b. Genoagent: A baseline method for llm-based exploration of gene expression data in alignment with bioinformaticians.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tianyang Liu, Fei Wang, and Muhao Chen. 2023b. Rethinking tabular data understanding with large language models. *arXiv preprint arXiv:2312.16702*.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024c. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv preprint arXiv:2402.17644*.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024d. Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.

Yang Liu and colleagues. 2024. Reasoning capabilities of large models. *AI Review*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. 2023. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*.

Giuliano Lorenzoni, Pedro Elkind Velmovitsky, Paulo Alencar, and Donald Cowan. 2024. Gpt-4 on clinic depression assessment: An llm-based pilot study.

Daqin Luo, Chengjian Feng, Yuxuan Nong, and Yiqing Shen. 2024. Autom3l: An automated multimodal machine learning framework with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8586–8594.

Karsten Lyko et al. 2016a. Data acquisition and its challenges in data science. *Data Science Journal*, 15:1–12.

Klaus Lyko, Marcus Nitzschke, and Axel-Cyrille Ngonga Ngomo. 2016b. Big data acquisition. *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe*, pages 39–61.

Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. Insightpilot: An llm-empowered automated data exploration system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 346–352.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812*.

Paula Maddigan and Teo Susnjak. 2023. Chat2vis: generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *Ieee Access*, 11:45181–45193.

Abhijit Manatkar, Ashlesha Akella, Parthivi Gupta, and Krishnasuri Narayanam. 2024. Quis: Question-guided insights generation for automated exploratory data analysis. *arXiv preprint arXiv:2410.10270*.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2024. Chartinstruct: Instruction tuning for chart comprehension and reasoning. *arXiv preprint arXiv:2403.09028*.

Luis Mayer, Christian Heumann, and Matthias Aßenmacher. 2024. Can opensource beat chatgpt?–a comparative study of large language models for text-to-code generation. *arXiv preprint arXiv:2409.04164*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR.

Grady McPeak, Anja Sautmann, Ohia George, Adham Hallal, Eduardo Arancón Simal, Aaron L. Schwartz, Jason Abaluck, Nirmal Ravi, and Robert Pless. 2024. An llm's medical testing recommendations in a nigerian clinic: Potential and limits of prompt engineering for clinical decision support. *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 586–591.

Mohamed El Amine Meguenani, Alceu de Souza Britto Jr, and Alessandro Lameiras Koerich. 2024. Music genre classification using large language models. *arXiv preprint arXiv:2410.08321*.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024a. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024b. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Md Saef Ullah Miah, Md Mohsin Kabir, Talha Bin Sarwar, Mejdl Safran, Sultan Alfarhood, and MF Mridha. 2024. A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm. *Scientific Reports*, 14(1):9603.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Xuran Ming, Shoubin Li, Mingyang Li, Lvlong He, and Qing Wang. 2024. Autolabel: Automated textual data annotation method based on active learning and large language model. In *International Conference on Knowledge Science, Engineering and Management*, pages 400–411. Springer.

Seyed Mohamad Moghadas, Yangxintong Lyu, Bruno Cornelis, Alexandre Alahi, and Adrian Munteanu. 2024. Strada-llm: Graph llm for traffic prediction. *arXiv preprint arXiv:2410.20856*.

Manuel Mondal, Julien Audiffren, Ljiljana Dolamic, Gérôme Bovet, and Philippe Cudré-Mauroux. 2024. Cleaning semi-structured errors in open data using large language models. In *2024 11th IEEE Swiss Conference on Data Science (SDS)*, pages 258–261. IEEE.

Juan Morales-García, Antonio Llanes, Francisco Arcas-Túnez, and Fernando Terroso-Sáenz. 2024. Developing time series forecasting models with generative large language models. *ACM Transactions on Intelligent Systems and Technology*.

Yida Mu, Peizhen Bai, Kalina Bontcheva, and Xingyi Song. 2024. Addressing topic granularity and hallucination in large language models for topic modelling. *arXiv preprint arXiv:2405.00611*.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. 2023. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15169–15179.

Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. 2024. Optimized feature generation for tabular data via llms with decision tree reasoning. *arXiv preprint arXiv:2406.08527*.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Nathalia Nascimento, Everton Guimaraes, Sai Sanjna Chintakunta, and Santhosh Anitha Boominathan. 2024. Llm4ds: Evaluating large language models for data science code generation. *arXiv preprint arXiv:2411.11908*.

Nathalia Nascimento, Cristina Tavares, Paulo Alencar, and Donald Cowan. 2023. Gpt in data science: A practical exploration of model selection. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4325–4334. IEEE.

Seth Neel and Peter Chang. 2024. Privacy issues in large language models: A survey.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-Tau Yih, Sida Wang, and Xi Victoria Lin. 2023. LEVER: Learning to verify language-to-code generation with execution. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26106–26128. PMLR.

Wei Ni, Kaihang Zhang, Xiaoye Miao, Xiangyu Zhao, Yangyang Wu, and Jianwei Yin. 2024. Iterclean: An iterative data cleaning framework with large language models. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*, pages 100–105.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.

Ruth Y Oliver, Melissa Chapman, Nathan Emery, Lauren Gillespie, Natasha Gownaris, Sophia Leiker, Anna C Nisi, David Ayers, Ian Breckheimer, Hannah Blondin, et al. 2024. Opening a conversation on responsible environmental data science in the age of large language models. *Environmental Data Science*, 3:e14.

OpenAI. 2024a. Gpt-4 technical report.

OpenAI. 2024b. Gpt-4o system card.

OpenAI. 2024c. Openai o1 system card.

Guy H Orcutt, Harold W Watts, and John B Edwards. 1968. Data aggregation and information loss. *The American Economic Review*, pages 773–787.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinivasan Iyer. 2024. Byte latent transformer: Patches scale better than tokens.

Dimitrios P. Panagoulias, Maria K. Virvou, and George A. Tsihrintzis. 2024. Evaluating llm - generated multimodal diagnosis from medical images and symptom analysis. *ArXiv*, abs/2402.01730.

Nicholas Pangakis and Samuel Wolken. 2024. Keeping humans in the loop: Human-centered automated annotation with generative ai. *arXiv preprint arXiv:2409.09467*.

Rahul Pankajakshan, Sumitra Biswal, Yuvaraj Govindarajulu, and Gilad Gressel. 2024. Mapping llm security landscapes: A comprehensive stakeholder risk assessment proposal. *ArXiv*, abs/2403.13309.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*.

Xuedi Qin, Yuyu Luo, Nan Tang, and Guoliang Li. 2020. Making data visualization more efficient and effective: a survey. *The VLDB Journal*, 29(1):93–117.

Jianing Qiu, Lin Li, Jiankai Sun, Jiachuan Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P-W Lo, Bo Xiao, et al. 2023. Large ai models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*.

Linlu Qiu, Fei Sha, Kelsey R Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. 2024. Can language models perform implicit bayesian inference over user preference states? In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*.

Qwen. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022a. Evaluating the text-to-sql capabilities of large language models. *arXiv preprint arXiv:2204.00498*.

Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022b. Evaluating the text-to-sql capabilities of large language models.

Krithika Ramesh, Nupoor Gandhi, Pulkit Madaan, Lisa Bauer, Charith Peris, and Anjalie Field. 2024. Evaluating differentially private synthetic data generation in high-stakes domains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15254–15269, Miami, Florida, USA. Association for Computational Linguistics.

A Rochan, P Sowmya, and D Anand Joseph Daniel. 2024. Large language model based document query solution using vector databases. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–5. IEEE.

Yucheng Ruan, Xiang Lan, Jingying Ma, Yizhi Dong, Kai He, and Mengling Feng. 2024. Language modeling on tabular data: A survey of foundations, techniques and evolution. *arXiv preprint arXiv:2408.10548*.

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

Malik Sallam. 2023. Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. *Healthcare*, 11.

Prateek Sancheti, Kamalakar Karlapalem, and Kavita Vemuri. 2024. Llm driven web profile extraction for identical names. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1616–1625.

Ranjan Sapkota, Achyut Paudel, and Manoj Karkee. 2024. Zero-shot automatic annotation and instance segmentation using llm-generated datasets: Eliminating field imaging and manual annotation for deep learning model development. *arXiv preprint arXiv:2411.11285*.

Iqbal H Sarker. 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2(5):377.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sarang Shaikh, Sher Muhammad Daudpota, Sule Yildirim Yayilgan, and Sindhu Sindhu. 2023. Exploring the potential of large-language models (llms) for student feedback sentiment analysis. In *2023 International Conference on Frontiers of Information Technology (FIT)*, pages 214–219. IEEE.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. 2024. Privacylens: Evaluating privacy norm awareness of language models in action. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ankita Sharma, Xuanmao Li, Hong Guan, Guoxin Sun, Liang Zhang, Lanjun Wang, Kesheng Wu, Lei Cao, Erkang Zhu, Alexander Sim, et al. 2023. Automatic data transformation using large language model-an experimental study on building energy data. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1824–1834. IEEE.

ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. 2024. Exploring multi-modal integration with tool-augmented llm agents for precise causal discovery. *arXiv preprint arXiv:2412.13667*.

Manisha M Shetty, Yinfang Chen, Gagan Somashekar, Ming-Jie Ma, Yogesh L. Simmhan, Xuchao Zhang, Jonathan Mace, Dax Vandevoorde, Pedro Henrique B. Las-Casas, Shachee Mishra Gupta, Suman Nath, Chetan Bansal, and S. Rajmohan. 2024. Building ai agents for autonomous clouds: Challenges and design principles. *ArXiv*, abs/2407.12165.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*.

Mohsen Shirali, Mohammadreza Fani Sani, Zahra Ahmadi, and Estefania Serral. 2024. Llm-based event abstraction and integration for iot-sourced logs. *arXiv preprint arXiv:2409.03478*.

Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du, and Yongfeng Zhang. 2024. Lawllm: Law large language model for the us legal system. In *International Conference on Information and Knowledge Management*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. *arXiv preprint arXiv:2310.11761*.

Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. 2025. Howtocaption: Prompting llms to transform video annotations at scale. In *European Conference on Computer Vision*, pages 1–18. Springer.

Vikramank Singh, Kapil Eknath Vaidya, Vinayshekhar Bannihatti Kumar, Sopan Khosla, Murali Narayanaswamy, Rashmi Gangadharaiah, and Tim Kraska. 2024. Panda: Performance debugging for databases using llm agents.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs. In *NeurIPS 2023 Second Table Representation Learning Workshop*.

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

Alexander Sorokin and David Forsyth. 2008. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE.

Dimitris Spathis and Fahim Kawsar. 2024. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158.

Sakhinana Sagar Srinivas and Venkataramana Runkana. 2024. Cross-modal learning for chemistry property prediction: Large language models meet graph machine learning. *arXiv preprint arXiv:2408.14964*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, et al. 2024. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *The 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*.

Guanglin Sun, Ting Li, Yongfang Ai, and Qinghai Li. 2023a. Digital finance and corporate financial fraud. *International Review of Financial Analysis*.

Maojun Sun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. 2024a. Lambda: A large model based data agent. *arXiv preprint arXiv:2407.17535*.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023b. Sentiment analysis through llm negotiations. *arXiv preprint arXiv:2311.01876*.

Zhaoyan Sun, Xuanhe Zhou, and Guoliang Li. 2024b. R-bot: An llm-based query rewrite system. *arXiv preprint arXiv:2412.01661*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Yi Tang, Chia-Ming Chang, and Xi Yang. 2024. Pdfchatannotator: A human-llm collaborative multimodal data annotation tool for pdf-format catalogs. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 419–430.

Raad Bin Tareaf, Mohammed AbuJarour, Tom Engelman, Philipp Liermann, and Jesse Klotz. 2024. Accelerating contextualization in ai large language models using vector databases. In *2024 International Conference on Information Networking (ICOIN)*, pages 316–321. IEEE.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Rricha Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, et al. 2024. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*.

Yuan Tian, Weiwei Cui, Dazhen Deng, Xinjing Yi, Yurun Yang, Haidong Zhang, and Yingcai Wu. 2024. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*.

Yongqi Tong, Sizhe Wang, Dawei Li, Yifan Wang, Simeng Han, Zi Lin, Chengsong Huang, Jiaxin Huang, and Jingbo Shang. 2024. Optimizing language model's reasoning abilities with weak supervision. *arXiv preprint arXiv:2405.04086*.

Petter Törnberg. 2023. How to use large language models for text analysis. *arXiv preprint arXiv:2307.13106*.

Petter Törnberg. 2024. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, page 08944393241286471.

Mohammed Toufiq, Darawan Rinchai, Eleonore Bettacchioli, Basirudeen Syed Ahamed Kabeer, Taushif Khan, Bishesh Subba, Olivia White, Marina Yurieva, Joshy George, Noemie Jourde-Chiche, et al. 2023. Harnessing large language models (llms) for candidate gene prioritization and selection. *Journal of Translational Medicine*, 21(1):728.

Satvik Tripathi, Kyla Gabriel, Pushpendra Kumar Tripathi, and Edward Kim. 2024. Large language models reshaping molecular biology and drug development. *Chemical Biology & Drug Design*, 103(6):e14568.

Dmitriy Umerenkov, Galina Zubkova, and Aleksandr Nesterov. 2023. Deciphering diagnoses: How large language models explanations influence clinical decision making. *arXiv preprint arXiv:2310.01708*.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large language models enable few-shot clustering. *arXiv preprint arXiv:2307.00524*.

Can Wang, Bolin Zhang, Dianbo Sui, Zhiying Tu, Xiaoyu Liu, and Jiabao Kang. 2024a. A survey on effective invocation methods of massive llm services. *arXiv preprint arXiv:2402.03408*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Jianwei Wang, Kai Wang, Ying Zhang, Wenjie Zhang, Xiwei Xu, and Xuemin Lin. 2025. On llm-enhanced mixed-type data imputation with high-order message passing. *arXiv preprint arXiv:2501.02191*.

Lei Wang, Songheng Zhang, Yun Wang, Ee-Peng Lim, and Yong Wang. 2023b. Llm4vis: Explainable visualization recommendation using chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 675–692.

Peidong Wang, Ming Wang, Zhiming Ma, Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2024b. Language models as continuous self-evolving data engineers. *arXiv preprint arXiv:2412.15151*.

Peifang Wang, Olga Golovneva, Armen Aghajanyan, Xiang Ren, Muhao Chen, Asli Celikyilmaz, and Maryam Fazel-Zarandi. 2023c. Domino: A dual-system for multi-step visual language reasoning. *arXiv preprint arXiv:2310.02804*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024c. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Ruiyu Wang, Zifeng Wang, and Jimeng Sun. 2023d. Unipredict: Large language models are universal tabular predictors. *arXiv preprint arXiv:2310.03266*.

Runze Wang, Mingqi Yang, and Yanming Shen. Graph2token: Make llms understand molecule graphs. In *ICML 2024 Workshop on Efficient and Accessible Foundation Models for Biological Discovery*.

Shaowei Wang, Lingling Zhang, Longji Zhu, Tao Qin, Kim-Hui Yap, Xinyu Zhang, and Jun Liu. 2024d. Cog-dqa: Chain-of-guiding learning with large language models for diagram question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13969–13979.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024e. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2024. Ceb: Compositional evaluation benchmark for fairness in large language models. In *NeurIPS Socially Responsible Research (SoLaR) Workshop*.

Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024a. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2282–2291.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024b. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. 2024c. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023e. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Yining Wang, Jinman Zhao, and Yuri Lawryshyn. 2024d. Gpt-signal: Generative ai for semi-automated feature engineering in the alpha research process. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 42–53.

Yu Wang, Zhiwei Liu, Jianguo Zhang, Weiran Yao, Shelby Heinecke, and Philip S Yu. 2023f. Drdt: Dynamic reflection with divergent thinking for llm-based sequential recommendation. *arXiv preprint arXiv:2312.11336*.

Yuxin Wang, Duanyu Feng, Yongfu Dai, Zhengyu Chen, Jimin Huang, Sophia Ananiadou, Qianqian Xie, and Hao Wang. 2024e. HARMONIC: Harnessing LLMs for tabular data synthesis and privacy protection. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhiruo Wang, Daniel Fried, and Graham Neubig. 2024f. Trove: Inducing verifiable and efficient toolboxes for solving programmatic tasks. *arXiv preprint arXiv:2401.12869*.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. 2023g. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*.

Zige Wang, Wanjun Zhong, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023h. Data management for large language models: A survey. *CoRR*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024g. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*.

Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. 2024h. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*.

Maximilian Weber and Merle Reichardt. 2023. Evaluation is all you need. prompting generative large language models for annotation tasks in the social sciences. a primer using open models. *arXiv preprint arXiv:2401.00284*.

Luoxuan Weng, Yinghao Tang, Yingchaojie Feng, Zhuo Chang, Peng Chen, Ruiqin Chen, Haozhe Feng, Chen Hou, Danqing Huang, Yang Li, et al. 2024a. Datalab: A unifed platform for llm-powered business intelligence. *arXiv preprint arXiv:2412.02205*.

Luoxuan Weng, Xingbo Wang, Junyu Lu, Yingchaojie Feng, Yihan Liu, Haozhe Feng, Danqing Huang, and Wei Chen. 2024b. Insightlens: Augmenting llm-powered data analysis with interactive insight management and navigation. *arXiv preprint arXiv:2404.01644*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance. *arXiv preprint arXiv:2305.14288*.

Daniel Adu Worae, Athar Sheikh, and Spyridon Mastorakis. 2024. A unified framework for context-aware iot management and state-of-the-art iot traffic anomaly detection. *arXiv preprint arXiv:2412.19830*.

Hulin Wu. 2024. Statistics evolution and revolution to meet data science challenges. *Statistics in Biosciences*, pages 1–19.

Jiaxuan Wu, Xushuo Tang, Zhengyi Yang, Kongzhang Hao, Longbin Lai, and Yongfei Liu. 2025. An experimental evaluation of llm on image classification. In *Australasian Database Conference*, pages 506–518. Springer.

Dean Wyatte, Fatemeh Tahmasbi, Ming Li, and Thomas Markovich. 2024. Scaling laws for discriminative classification in large language models. *arXiv preprint arXiv:2405.15765*.

Oskar Wysocki, Magdalena Wysocka, Danilo Carvalho, Alex Teodor Bogatu, Danilo Miranda Gusicuma, Maxime Delmas, Harriet Unsworth, and Andre Freitas. 2024. An llm-based knowledge synthesis and scientific reasoning framework for biomedical discovery. *arXiv preprint arXiv:2406.18626*.

Yunjia Xi, Weiwen Liu, Jianghao Lin, Bo Chen, Ruiming Tang, Weinan Zhang, and Yong Yu. 2024. Memocrs: Memory-enhanced sequential conversational recommender systems with large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2585–2595.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023a. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *International Conference on Learning Representations*.

Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. 2023b. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Yikuan Xia, Jiazun Chen, Xinchi Li, and Jun Gao. 2024. Aprompt4em: Augmented prompt tuning for generalized entity matching. *arXiv preprint arXiv:2405.04820*.

Bin Xiao, Burak Kantarci, Jiawen Kang, Dusit Niyato, and Mohsen Guizani. 2024a. Efficient prompting for llm-based generative internet of things. *arXiv preprint arXiv:2406.10382*.

Wen Xiao, Yujia Xie, Giuseppe Carenini, and Pengcheng He. 2023. Personalized abstractive summarization by tri-agent generation pipeline. *arXiv preprint arXiv:2305.02483*.

Yijia Xiao, Edward Sun, Yiqiao Jin, Qifan Wang, and Wei Wang. 2024b. Proteingpt: Multimodal llm for protein property prediction and structure understanding. *arXiv preprint arXiv:2408.11363*.

Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. 2024c. Tradingagents: Multi-agents llm financial trading framework.

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

Zikai Xie. 2024. Order matters in hallucination: Reasoning order as benchmark and reflexive prompting for large-language-models.

Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Chengqi Xu and Olivier Elemento. 2024. The potential and pitfalls of large language models in molecular biosciences. *The Biochemist*, 46(2):13–17.

Mengwei Xu, Dongqi Cai, Wangsong Yin, Shangguang Wang, Xin Jin, and Xuanzhe Liu. 2024a. Resource-efficient algorithms and systems of foundation models: A survey. *ACM Computing Surveys*.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Kang Liu, and Jun Zhao. 2024b. Generate-on-graph: Treat llm as both agent and kg in incomplete knowledge graph question answering. *arXiv preprint arXiv:2404.14741*.

Siqiao Xue, Danrui Qi, Caigao Jiang, Wenhui Shi, Fangyin Cheng, Keting Chen, Hongjun Yang, Zhiping Zhang, Jianshan He, Hongyang Zhang, et al. 2024. Demonstration of db-gpt: Next generation data interaction system empowered by large language models. *arXiv preprint arXiv:2404.10209*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024a. On protecting the data privacy of large language models (llms): A survey.

Mengyi Yan, Yaoshu Wang, Yue Wang, Xiaoye Miao, and Jianxin Li. 2024b. Gidcl: A graph-enhanced interpretable data cleaning framework with large language models. *Proceedings of the ACM on Management of Data*, 2(6):1–29.

Tianze Yang, Tianyi Yang, Fuyuan Lyu, Shaoshan Liu, et al. 2024a. Ice-search: A language model-driven feature selection approach. *arXiv preprint arXiv:2402.18609*.

Xiaohao Yang, He Zhao, Weijie Xu, Yuanyuan Qi, Jueqing Lu, Dinh Phung, and Lan Du. 2024b. Neural topic modeling with large language models in the loop. *arXiv preprint arXiv:2411.08534*.

Yazheng Yang, Yuqi Wang, Sankalok Sen, Lei Li, and Qi Liu. 2024c. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*.

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

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. Idea2img: Iterative self-refinement with gpt-4v (ision) for automatic image design and generation. *arXiv preprint arXiv:2310.08541*.

Michihiro Yasunaga, Leonid Shamis, Chunting Zhou, Andrew Cohen, Jason Weston, Luke Zettlemoyer, and Marjan Ghazvininejad. 2024. Alma: Alignment with minimal annotation. *arXiv preprint arXiv:2412.04305*.

Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, Yongfeng Zhang, et al. 2023a. Natural language is all a graph needs. *arXiv preprint arXiv:2308.07134*, 4(5):7.

Wen Ye, Yizhou Zhang, Wei Yang, Lumingyuan Tang, Defu Cao, Jie Cai, and Yan Liu. 2024. Beyond forecasting: Compositional time series reasoning for end-to-end task execution. *arXiv preprint arXiv:2410.04047*.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023b. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–184.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023c. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 174–184, New York, NY, USA. Association for Computing Machinery.

Yuwei Yin, Yazheng Yang, Jian Yang, and Qi Liu. 2023. Finpt: Financial risk prediction with profile tuning on pretrained foundation models. *arXiv preprint arXiv:2308.00065*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 189–198.

Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023a. Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability. In *American Medical Informatics Association (AMIA) Annual Symposium*.

Lifan Yuan, Yangyi Chen, Xingyao Wang, Yi R Fung, Hao Peng, and Heng Ji. 2023b. Craft: Customizing llms by creating and retrieving from specialized toolsets. *arXiv preprint arXiv:2309.17428*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, et al. 2024a. Advancing llm reasoning generalists with preference trees. *arXiv preprint arXiv:2404.02078*.

Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Quanzheng Li, Yang Chen, Li Zhang, et al. 2024b. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. *Medicine Plus*, page 100030.

Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, et al. 2024c. Llm inference unveiled: Survey and roofline model insights. *arXiv preprint arXiv:2402.16363*.

R Zafarani. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

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

Jingying Zeng, Richard Huang, Waleed Malik, Langxuan Yin, Bojan Babic, Danny Shacham, Xiao Yan, Jaewon Yang, and Qi He. 2024a. Large language models for social networks: Applications, challenges, and solutions. *arXiv preprint arXiv:2401.02575*.

Xianlong Zeng, Yijing Gao, Fanghao Song, and Ang Liu. 2024b. Similar data points identification with llm: A human-in-the-loop strategy using summarization and hidden state insights. *arXiv preprint arXiv:2404.04281*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *arXiv preprint arXiv:2404.03868*.

Enshi Zhang and Christian Poellabauer. 2024. Contextual speech emotion recognition with large language models and asr-based transcriptions. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. 2024a. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*.

Lingzheng Zhang, Lifeng Shen, Yimin Zheng, Shiyuan Piao, Ziyue Li, and Fugee Tsung. 2024b. Lemole: Llm-enhanced mixture of linear experts for time series forecasting. *arXiv preprint arXiv:2412.00053*.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023a. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

Shuo Zhang, Zezhou Huang, and Eugene Wu. 2024c. Data cleaning using large language models. *arXiv preprint arXiv:2410.15547*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024d. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wenhu Zhang. 2024. Llm-driven table reasoning. In *EACL Proceedings*.

Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2023b. Data-copilot: Bridging billions of data and humans with autonomous workflow. *arXiv preprint arXiv:2306.07209*.

Wenxuan Zhang, Yue Deng, Bing-Quan Liu, Sinno Jialin Pan, and Lidong Bing. 2023c. Sentiment analysis in the era of large language models: A reality check. *ArXiv*, abs/2305.15005.

Xiang Zhang, Juntai Cao, and Chenyu You. 2024e. Counting ability of large language models and impact of tokenization.

Xinhao Zhang, Jinghan Zhang, Banafsheh Rekabdar, Yuanchun Zhou, Pengfei Wang, and Kunpeng Liu. 2024f. Dynamic and adaptive feature generation with llm. *arXiv preprint arXiv:2406.03505*.

Yanlin Zhang, Ning Li, Quan Gan, Weinan Zhang, David Wipf, and Minjie Wang. 2024g. Elf-gym: Evaluating large language models generated features for tabular prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5420–5424.

Yi Zhang, Zhongyang Yu, Wanqi Jiang, Yufeng Shen, and Jin Li. 2023d. Long-term memory for large language models through topic-based vector database. In *2023 International Conference on Asian Language Processing (IALP)*, pages 258–264. IEEE.

Yimei Zhang, Xiangjie Kong, Wenfeng Zhou, Jin Liu, Yanjie Fu, and Guojiang Shen. 2024h. A comprehensive survey on traffic missing data imputation. *IEEE Transactions on Intelligent Transportation Systems*.

Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, and Yulia Tsvetkov. 2024i. Can llm graph reasoning generalize beyond pattern memorization? *arXiv preprint arXiv:2406.15992*.

Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023e. Clusterllm: Large language models as a guide for text clustering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zeyu Zhang, Paul Groth, Iacer Calixto, and Sebastian Schelter. 2024j. Directions towards efficient and automated data wrangling with large language models. In *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, pages 301–304. IEEE.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024k. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Zijian Zhang, Yujie Sun, Zepu Wang, Yuqi Nie, Xiaobo Ma, Peng Sun, and Ruolin Li. 2024l. Large language models for mobility in transportation systems: A survey on forecasting tasks. *arXiv preprint arXiv:2405.02357*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024a. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Chengshuai Zhao, Garima Agrawal, Tharindu Kumarage, Zhen Tan, Yuli Deng, Yingying Chen, and Huan Liu. 2024b. Ontology-aware rag for improved question-answering in cybersecurity education.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. 2024c. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

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

Xiaofeng Zhao, Xiaosong Qiao, Kaiwen Ou, Min Zhang, Chang Su, Mengyao Piao, Yuang Li, Yinglu Li, Ming Zhu, and Yilun Liu. 2024d. Hw-tsc at semeval-2024 task 5: Self-eval? a confident llm system for auto prediction and evaluation for the legal argument reasoning task. In *International Workshop on Semantic Evaluation*.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023a. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Jiaming Zhou, Abbas Ghaddar, Ge Zhang, Liheng Ma, Yaochen Hu, Soumyasundar Pal, Mark Coates, Bin Wang, Yingxue Zhang, and Jianye Hao. 2024a. Enhancing logical reasoning in large language models through graph-based synthetic data. *arXiv preprint arXiv:2409.12437*.

Wei Zhou, Mohsen Mesgar, Annemarie Friedrich, and Heike Adel. 2024b. Efficient multi-agent collaboration with tool use for online planning in complex table question answering. *arXiv preprint arXiv:2412.20145*.

Xiaomao Zhou, Yujiao Hu, Qingmin Jia, and Renchao Xie. 2024c. Cross-domain integration for general sensor data synthesis: Leveraging llms and domain-specific generative models in collaborative environments. *IEEE Sensors Journal*.

Xiaomao Zhou, Qingmin Jia, and Yujiao Hu. 2024d. Advancing general sensor data synthesis by integrating llms and domain-specific generative models. *IEEE Sensors Letters*.

Xuanhe Zhou, Guoliang Li, Zhaoyan Sun, Zhiyuan Liu, Weize Chen, Jianming Wu, Jiesi Liu, Ruohang Feng, and Guoyang Zeng. 2023b. D-bot: Database diagnosis system using large language models. *arXiv preprint arXiv:2312.01454*.

Yue Zhou, Chenlu Guo, Xu Wang, Yi Chang, and Yuan Wu. 2024e. A survey on data augmentation in large model era. *arXiv preprint arXiv:2401.15422*.

Yujuan Zhou. 2024. Data augmentation and generation with llms. In *ACL Findings*.

Jingyu Zhu, Xintong Zhao, Yu Sun, Shaoxu Song, and Xiaojie Yuan. 2024a. Relational data cleaning meets artificial intelligence: A survey. *Data Science and Engineering*, pages 1–28.

Jun-Peng Zhu, Peng Cai, Kai Xu, Li Li, Yishen Sun, Shuai Zhou, Haihuang Su, Liu Tang, and Qi Liu. 2024b. Autotqa: Towards autonomous tabular question answering through multi-agent large language models. *Proceedings of the VLDB Endowment*, 17(12):3920–3933.

Jun-Peng Zhu, Boyan Niu, Zheming Ni, Jianwei Wan, Kai Xu, Jiajun Huang, Shengbo Ma, Bing Wang, Xuan Zhou, Guanglei Bao, et al. 2024c. Towards automated cross-domain exploratory data analysis through large language models. *arXiv preprint arXiv:2412.07214*.

Yaochen Zhu, Chao Wan, Harald Steck, Dawen Liang, Yesu Feng, Nathan Kallus, and Jundong Li. 2025. Collaborative retrieval for large language model-based conversational recommender systems. In *The Web Conference 2025*.

Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. 2024d. Are large language models good statisticians? *arXiv preprint arXiv:2406.07815*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Bridging Traditional Data Science with LLM Capabilities

The integration of LLMs into data science represents a paradigm shift in how data is processed and analyzed. Traditional approaches often rely on rigid rules and extensive feature engineering. LLMs, however, can generalize across tasks and domains by leveraging extensive pre-trained knowledge and contextual understanding (Dai and colleagues, 2024). Particularly, LLMs reduce dependency on predefined schemas and enable more adaptive and flexible data workflows. For instance, they can generate synthetic data to address gaps in existing datasets (Zhou, 2024), clean and structure raw data, and support complex analytical tasks using natural language interfaces (Chen and Si, 2024). Moreover, LLMs democratize data science by allowing non-experts to engage with data via conversational interfaces. This capability lowers barriers to data-driven decision-making and encourages collaboration between technical and non-technical stakeholders (Zhang, 2024).

In conclusion, the evolution of data science reflects a continuous progression towards greater automation, scalability, and accessibility. LLMs are instrumental in this transformation, offering unprecedented opportunities to enhance every stage of the data science life cycle.

B Applications

LLMs have transformed specialized fields by automating complex language tasks, streamlining domain-specific workflows, and supporting more informed decision-making (Shu et al., 2017; Li et al., 2017; Zafarani, 2014). Applications extend beyond generic data science processes to include social science, medical research, finance, legal analysis, education, and environmental sciences.

B.1 Social Science

In social science, LLMs analyze large volumes of unstructured data to uncover human behavior, societal trends, and policy impacts.

Sentiment Analysis. By processing texts from platforms like Twitter or Reddit, LLMs detect emerging issues, predict social shifts, and inform policymaking (Törnberg, 2023; Zeng et al., 2024a; Zhang et al., 2023c; Jiang et al., 2024). Their ability to handle slang, dialects, and multilingual inputs makes them invaluable for studying diverse populations (Törnberg, 2024; Broekens et al., 2023; Shaikh et al., 2023; Dudy et al., 2024; Bian et al., 2023).

Policy Making. LLMs streamline policy research by analyzing government documents, academic literature, and public feedback. They identify key themes, summarize large text corpora, and highlight policy outcomes, allowing real-time evaluation of effectiveness and reducing manual effort (Chen et al., 2024a; Ishimizu et al., 2024; Kasztelnik and Branch, 2024; Ziems et al., 2024; Weber and Reichardt, 2023).

B.2 Medical

In healthcare, LLMs enhance clinical decision-making, patient care, and drug discovery, offering structured insights from vast clinical records and scientific literature.

Clinical Diagnostics. LLMs extract critical information from unstructured patient data—such as symptoms and test results—to assist with diagnosis recommendations and risk detection (Wang et al., 2023a; McPeak et al., 2024; Lorenzoni et al., 2024; Xia et al., 2023a). This reduces clinicians' workload while improving diagnostic accuracy (Xie et al., 2024; Bannett et al., 2024; Panagoulias et al., 2024).

Drug Discovery. In pharmaceutical research, LLMs parse scientific articles and experimental data to identify promising compounds, predict molecular interactions, and accelerate drug development timelines (Xu and Elemento, 2024; Tripathi et al., 2024; Sallam, 2023; Bran et al., 2023; Cao et al., 2023).

B.3 Finance

Financial applications of LLMs include risk analysis, fraud detection, and market prediction, all benefiting from their ability to handle large, domain-specific datasets.

Risk Management. LLMs analyze financial statements, regulatory documents, and news feeds to identify market volatility, credit risk, and other threats (Cao et al., 2024b,c; Xiao et al., 2024c; Pankajakshan et al., 2024), enabling proactive strategies in fast-paced markets.

Fraud Prevention. By examining transactional data for anomalies, LLMs help financial institutions mitigate fraud risks and ensure regulatory compliance (Yin et al., 2023; Chakraborty et al., 2024; Gregory and Vito, 2024; Sun et al., 2023a; Chang et al., 2022).

B.4 Other Domains

Beyond social science, medicine, and finance, LLMs also support legal, educational, and environmental sectors by automating core data-related tasks.

Legal. LLMs extract and structure case law, identify critical clauses, and compare precedents to forecast legal outcomes (Shui et al., 2023; Fei et al., 2023; Zhao et al., 2024d; Shu et al., 2024; Kalra et al., 2024; Harasta et al., 2024).

Education. In academic settings, LLMs preprocess student interaction data from learning platforms to reveal engagement patterns, generate personalized learning features, and boost predictive models for performance (Wang et al., 2024e; Leinonen et al., 2024; Gan et al., 2023; Zhao et al., 2024b; Alhafni et al., 2024; Li et al., 2023b).

Environmental Sciences. LLMs ingest climate reports, track environmental indicators, and produce actionable insights, helping policymakers develop strategies for climate change mitigation (Kraus et al., 2023; Thulke et al., 2024; Oliver et al., 2024; Chen et al., 2023a).

By applying LLM-based techniques to domain-specific challenges, practitioners gain deeper insights, streamline workflows, and make data-driven decisions that span multiple fields.

Figure 2: LLMs for Data Science