# A Survey on Tabular Data Understanding and Language Model Applications

www.surveyx.cn

## Abstract

This survey explores the interdisciplinary domain of tabular data understanding and language model applications, emphasizing the integration of structured data processing with advanced language model capabilities. Key frameworks such as SynLM and D ATA T ALES illustrate the synergy between language models and tabular data generation, while challenges in tabular reasoning highlight the need for large language models (LLMs) to process extensive tabular data effectively. The survey investigates methods like UniPredict and benchmarks such as AnaMeta, emphasizing the importance of understanding field semantics and the role of pretrained language models in embedding relational tables. The exploration of multimodal AutoML and simulation of tabular datasets using LLMs underscores the potential for innovation in this field. The survey also addresses the significance of prompt engineering in enhancing model performance and strategies for crafting effective prompts. It evaluates datasets and benchmarks such as WikiTableQuestions and FeTaQA, essential for advancing tabular data analysis. Applications of language models in data visualization, question answering, and predictive modeling are examined, highlighting frameworks like SemTUI and H-STAR. The survey concludes by discussing automated data analysis and feature engineering, showcasing AI's transformative impact on decision-making processes. Challenges and future directions include improving LLM interpretability, optimizing table preprocessing, and enhancing dataset fidelity. By addressing these challenges, the survey aims to propel the field forward, facilitating more efficient and insightful data-driven decision-making across various domains.

## 1 Introduction

### 1.1 Interdisciplinary Domain Overview

The integration of tabular data understanding with language model applications represents a significant advancement in artificial intelligence, merging structured data processing with sophisticated language capabilities. Notable frameworks, such as SynLM, exemplify this synergy by generating synthetic data from tabular datasets, while the D ATA T ALES benchmark assesses language models' proficiency in creating narratives from complex tabular data, including financial reports, effectively bridging data analysis and narrative generation [1, 2].

Challenges in tabular reasoning are pronounced, particularly regarding large language models (LLMs) and their ability to process extensive tabular data within input length constraints [3]. This underscores the necessity of integrating language understanding with structured data analysis for tasks like SQL generation and table reasoning [4]. The AnaMeta benchmark highlights the critical need for understanding field semantics in tabular data analysis, reflecting interdisciplinary challenges [5].

Pretrained language models have garnered attention for embedding relational tables, crucial for various applications, emphasizing the cross-disciplinary nature of embedding techniques in tabular data understanding and language model applications [6]. However, challenges persist in pretraining

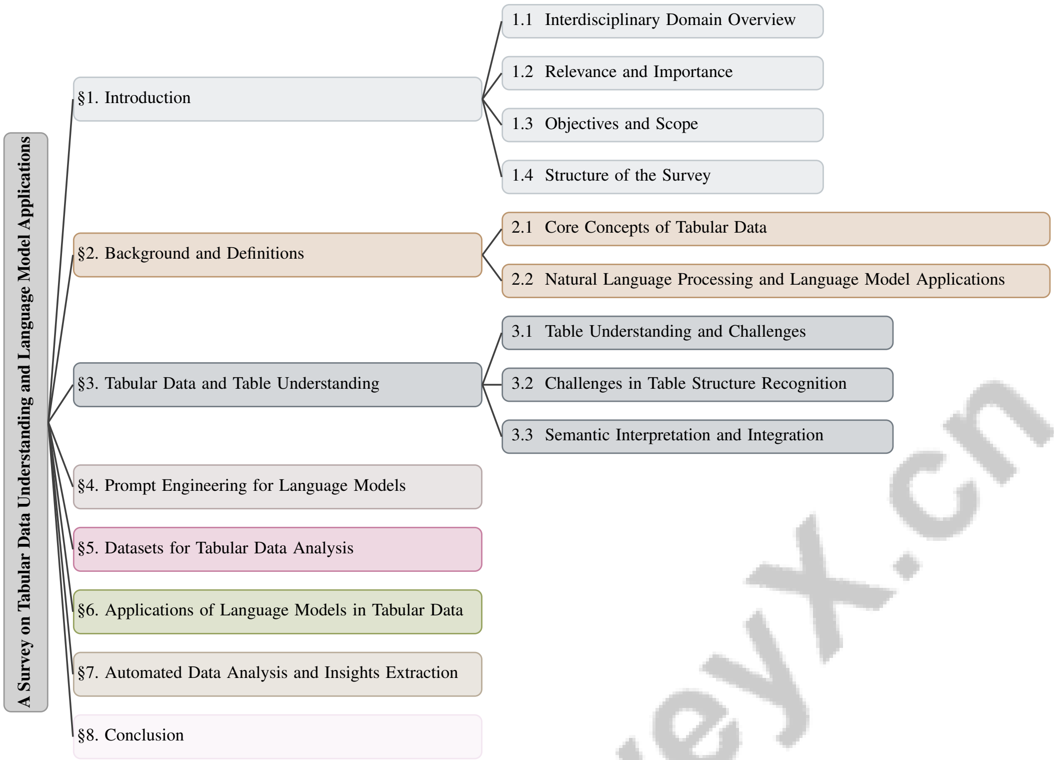

A Survey on Tabular Data Understanding and Language Model Applications
§1. Introduction
1.1 Interdisciplinary Domain Overview
1.2 Relevance and Importance
1.3 Objectives and Scope
1.4 Structure of the Survey
§2. Background and Definitions
2.1 Core Concepts of Tabular Data
2.2 Natural Language Processing and Language Model Applications
§3. Tabular Data and Table Understanding
3.1 Table Understanding and Challenges
3.2 Challenges in Table Structure Recognition
3.3 Semantic Interpretation and Integration
§4. Prompt Engineering for Language Models
§5. Datasets for Tabular Data Analysis
§6. Applications of Language Models in Tabular Data
§7. Automated Data Analysis and Insights Extraction
§8. Conclusion

Figure 1: chapter structure

on tabular data, especially with fixed-schema or single-table scenarios, necessitating more flexible approaches [7].

Furthermore, the benchmarking of multimodal AutoML for tabular data, which incorporates text fields alongside numeric and categorical features, illustrates the interdisciplinary dynamics of this field [8]. The simulation of tabular datasets using LLMs to explore hypotheses about real-world entities showcases the innovative potential at the intersection of tabular data and language model applications [9]. Collectively, these developments underscore the interdisciplinary challenges and opportunities that drive innovation in addressing complex data-centric issues across multiple domains.

## 1.2 Relevance and Importance

Tabular data understanding's significance in AI is underscored by the pivotal role of LLMs in prediction, generation, and understanding tasks within tabular data analysis, impacting current and future technologies [10]. The complexity of this domain is further exemplified by the challenges machine learning faces in rapidly translating raw data into accurate predictions, particularly with multimodal data [8]. The need for benchmarks capturing the intricacies of data narration, especially in financial contexts, is crucial for enhancing business decision-making processes [2].

Maintaining differential privacy during synthetic data generation is essential, emphasizing privacy's importance in AI and data analysis [1]. Despite advancements in embedding techniques, a lack of comprehensive understanding regarding their strengths and limitations leads to inefficiencies in application [6]. LLMs often struggle with reasoning over large tables that exceed their token limits, resulting in truncation or hallucination of outputs, highlighting the need for improved reasoning capabilities [3].

The demand for breakthroughs in cross-table pretraining is evident, as it is vital for leveraging the vast amounts of tabular data available online [7]. Simulating tabular datasets using LLMs can significantly reduce human effort in data collection and hypothesis testing, underlining the importance of this area [9]. The AnaMeta benchmark addresses the critical need for accurate understanding of field semantics to effectively operate on table fields, marking a significant advancement in AI and data analysis [5].

AI-generated, for reference only.

The integration of advanced language models and innovative methodologies for tabular data understanding has transformed AI and data analysis technologies. Recent research highlights enhancements in the accuracy of complex table queries through context enrichment in Retrieval-Augmented Generation (RAG) systems, schema-driven information extraction techniques that convert heterogeneous tables into structured records, and comprehensive data lake designs that manage textual and tabular data jointly. Breakthroughs in LLMs have also improved predictive performance and robustness in tabular data analysis, enabling diverse applications across various domains. These advancements pave the way for more efficient and effective solutions to longstanding challenges in information retrieval and data processing [11, 12, 13, 14, 15].

### 1.3 Objectives and Scope

This survey delineates the convergence of tabular data understanding and language model applications, emphasizing the automation and optimization of intricate tasks such as data transformation, feature generation, and synthesis using LLMs. A primary objective is to investigate frameworks that enhance feature generation through LLMs, streamlining data processing and analysis [16]. The exploration of methods like UniPredict, which employs generative modeling for scalable tabular data prediction without dataset-specific models, broadens LLM applicability across diverse data environments [17].

The survey encompasses deep learning methodologies for tabular data, including data transformations, specialized architectures, and regularization models critical for enhancing LLM performance in table-centric tasks [18]. It also addresses challenges such as sub-cell named entity recognition in industrial tables, proposing novel approaches to mitigate the scarcity of labeled training data [19]. Evaluating benchmarks designed to assess LLMs' capabilities in generating textual insights from tabular data is another focal point, facilitating efficient information retrieval and decision-making processes [20].

A significant component of the survey is examining self-supervised pretraining methods like TABBIE, specifically tailored for tabular data to enhance performance on table-centric tasks [21]. It also investigates frameworks integrating tabular operations into reasoning processes, improving the understanding and manipulation of tabular data [22]. Moreover, the role of explainability in LLMs is scrutinized, focusing on how individual words in prompts influence model outputs, which is vital for enhancing AI systems' transparency and interpretability [23].

The survey further addresses challenges of class imbalance and data bias in tabular datasets, providing realistic test beds for evaluating machine learning methods [24]. By encompassing these objectives and scope, the survey aims to advance the field of tabular data understanding and language model applications, highlighting methodologies that enhance data transformation, augmentation, synthesis, and user interaction across various real-world scenarios. This includes optimizing prompt generation for tabular datasets to improve column selection and few-shot example retrieval [25], and exploring the historical development of data analysis methods, focusing on tabular and heterogeneous data analysis [26]. The survey also considers the potential of no-code platforms like JarviX, which empower users to perform data analytics through automated guides utilizing LLMs [27], and frameworks like the P-Transformer for effectively utilizing both structured and unstructured data in medical prediction tasks [28].

Additionally, the survey aims to address the critical need for generating synthetic data from tabular datasets while preserving differential privacy, as highlighted by the limitations of existing methods [1]. The lack of effective methods for pretraining on heterogeneous tabular data across multiple tables, which hinders knowledge generalization from one table to another, is another focal issue [7]. Furthermore, it examines the potential of utilizing LLMs to generate approximate datasets for hypothesis testing, enabling researchers to iterate more rapidly [9]. The AnaMeta benchmark's objectives are also considered, focusing on evaluating models' abilities to infer field metadata from tabular data and improving the understanding of field semantics in analysis tasks [5].

### 1.4 Structure of the Survey

The survey is systematically organized into distinct sections, each addressing a critical aspect of tabular data understanding and language model applications. The introductory section sets the stage by discussing the interdisciplinary nature of the field, its relevance, and the scope of the survey. Following this, the background section provides essential definitions and historical context, elucidating core concepts such as tabular data, prompt engineering, and natural language processing.

Subsequent sections delve into specific topics, beginning with an exploration of tabular data and table understanding, examining methods and challenges associated with table structure recognition and semantic interpretation. This is followed by an analysis of prompt engineering for language models, highlighting its role in enhancing model performance and strategies for crafting effective prompts.

The survey comprehensively evaluates various datasets utilized for tabular data analysis, detailing the different types of datasets, the inherent challenges in their compilation—such as handling missing values, dataset imbalance, and diverse column types—and highlighting significant benchmarks, including the SciTabQA and SEM-TAB-FACTS datasets, that have notably impacted advancements in the field [29, 30, 31, 32, 26]. Applications of language models in tabular data are explored next, focusing on use cases like data visualization, question answering, and predictive modeling.

The penultimate section addresses automated data analysis and insights extraction, examining the role of AI in automating analytical processes and improving decision-making. The survey concludes with a summary of key findings, implications for future research, and potential areas for further investigation. Each section is meticulously referenced to ensure a comprehensive understanding of the current landscape and future directions in this rapidly evolving domain.The following sections are organized as shown in Figure 1.

## 2 Background and Definitions

### 2.1 Core Concepts of Tabular Data

Tabular data, characterized by its row and column structure, is pivotal in sectors like finance, healthcare, and research, supporting tasks such as data integration, cleaning, and feature engineering [8]. The heterogeneity of tabular data, however, complicates novel class discovery due to structural inconsistencies [33]. Effective representation of numerical and categorical features is crucial for AI applications in this context [34].

Large language models (LLMs) face challenges with tabular data in prediction and synthesis tasks due to dataset complexity and size, often constrained by token limits, which can reduce accuracy [10, 3]. Tabular data distillation is essential for optimizing analysis by minimizing data volume while preserving key information [35].

In financial contexts, the DATA TALES benchmark underscores the importance of generating clear narratives from tabular data for insight extraction [2]. Synthetic data generation plays a critical role in creating privacy-preserving datasets that replicate original data distributions, facilitating secure analysis and model training [1]. Understanding field semantics is paramount for effective operations across domains [5].

Advanced methodologies are needed for embedding relational tables in tasks like question answering and data integration, capturing essential tabular properties [6]. The CM2 pretraining framework enables cross-table learning and heterogeneous table encoding, predicting masked features using retained features and schema prompts [7]. Natural language query interpretation requires both logical and mathematical reasoning, highlighting the complexities of integrating tabular data with language models [4].

Tabular data often exists in unstructured formats, complicating parsing and content extraction [36]. Exploring qualitative hypotheses about real-world entities demands extensive manual dataset curation from these unstructured sources [9]. Despite these challenges, tabular data remains foundational in AI systems, necessitating ongoing research to address complexities and leverage its potential in data-driven decision-making.

### 2.2 Natural Language Processing and Language Model Applications

Natural language processing (NLP) techniques integrated with language models have advanced the interpretation of tabular data, enabling insights across domains. Large language models (LLMs) have been crucial in generating structured tabular data by interpreting complex datasets [37]. Models like TABERT enhance understanding of both natural language and structured data through joint representations [38].

AI-generated, for reference only.

Challenges persist in handling complex queries involving ambiguous or incomplete tabular data, limiting LLM effectiveness and necessitating model design innovations [39]. High dimensionality and intricate feature interactions in tabular data pose hurdles, often leading to overfitting with small sample sizes [40].

Benchmarks like TALENT provide frameworks for evaluating language models on tabular tasks, facilitating fair comparison across models [41]. Other benchmarks focus on transforming complex (table, query) inputs into executable code outputs, highlighting the role of language models in generating actionable insights [42].

Evaluations of deep learning models on tabular data compare deep neural networks and tree-based methods, revealing strengths and limitations [43]. Transfer learning enhances deep tabular model effectiveness, especially with limited downstream data, compared to traditional methods like gradient boosting decision trees [44].

The performance of small language models on classification and regression tasks using raw tabular data underscores their potential despite challenges [45]. Tools like C2, a column-to-concept mapper, use maximum likelihood estimation to predict concepts from structured data [46]. A data-centric evaluation framework includes dataset-specific preprocessing and performance measures, advancing NLP and language model applications in tabular contexts [47].

Exploring numerical reasoning in language models is facilitated by diverse numerical probes from tabular data, allowing comprehensive evaluations [48]. Challenges such as natural language complexity, accurate semantic parsing, and addressing diverse user queries remain significant in developing natural language interfaces for tabular data [49].

# 3 Tabular Data and Table Understanding

## 3.1 Table Understanding and Challenges

The complexity and variability inherent in tabular data pose significant challenges for understanding table structures and their semantic interpretation. High feature heterogeneity and cardinality necessitate sophisticated embedding techniques for effective comprehension [34]. Additionally, the scarcity of high-quality labeled data complicates accurate table parsing, hindering robust model development [36]. Converting tabular formats from images into machine-readable formats requires bridging visual and language processing, complicated by diverse table layouts [50]. Current benchmarks often fall short in evaluating language models' ability to narrate data effectively, highlighting the need for comprehensive evaluation frameworks [1]. Integrating symbolic and semantic reasoning remains challenging, affecting performance on complex queries that demand deep semantic understanding [4]. The labor-intensive task of structuring data from unstructured sources further underscores the need for efficient data processing methodologies [9].

## 3.2 Challenges in Table Structure Recognition

| Method Name | Complexity Handling | Feature Utilization | Interpretation Challenges |
|---|---|---|---|
| SEM[51] | Spanning Cells | Visual Textual Features | Descriptive Text Interpretation |
| CATT-Net[52] | Complex Table Structures | Position Features Limitations | Understanding Descriptive Text |
| SAFS[53] | Spanning Cells | Positional Feature Challenges | Descriptive Text Difficulties |
| UTN[50] | Complex Table Structures | Positional And Features | Descriptive Text Interpretation |
| SS-CoT[54] | Complex Table Structures | Positional And Other | Descriptive Text Interpretation |

Table 1: Comparison of various methods for table structure recognition, highlighting their approaches to complexity handling, feature utilization, and interpretation challenges. This table provides insights into the diverse strategies employed by different models to address the inherent difficulties in processing complex tables.

Recognizing and processing complex table structures present numerous challenges, especially with tables containing spanning cells that carry critical semantic information [51]. Methods often focus on simpler tables, hindering generalization across varying document domains and resolutions due to reliance on position features [52]. The exponential growth of feature combinations complicates subgroup detection and processing [53]. Descriptive text within table cells introduces interpretation ambiguities, reducing recognition accuracy [50]. The lack of standardized evaluation criteria complicates the development of robust table recognition systems [55]. Inaccuracies arise in code generation when LLM prompts lack representative data, compounded by LLMs' limited numerical data handling capabilities [56, 57]. Simplified task approaches without leveraging reasoning processes result in inadequate understanding of complex structures [54].

This is illustrated in Figure 2, which categorizes the primary challenges in table structure recognition into three main areas: issues with complex tables, concerns related to features and evaluation, and difficulties associated with reasoning and interpretation. Table 1 offers a comparative analysis of different methods used in table structure recognition, focusing on their handling of complexity, utilization of features, and the challenges they face in interpretation. Despite progress in managing heterogeneous features and improving interpretability through attention mechanisms [40], substantial challenges remain in enhancing table structure recognition accuracy.

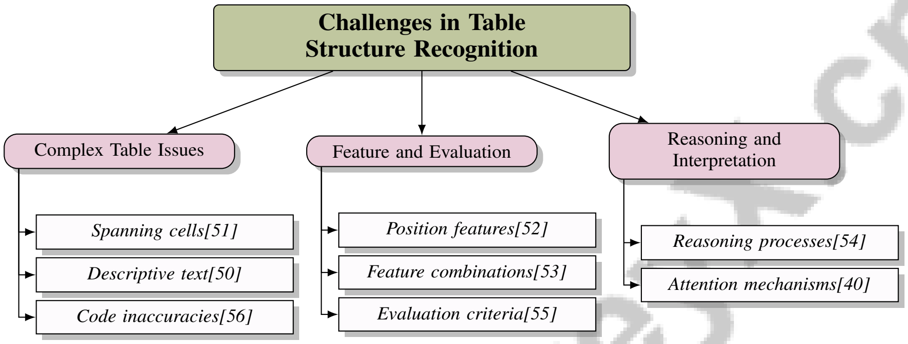


Figure 2: This figure illustrates the primary challenges in table structure recognition, categorized into issues with complex tables, feature and evaluation concerns, and reasoning and interpretation difficulties.

### 3.3 Semantic Interpretation and Integration

| Method Name | Integration Approaches | Reasoning Frameworks | Handling Challenges |
|---|---|---|---|
| TN[58] | Semantic Rules | Iterative Reasoning Feedback | Noise Tolerance |
| CoT[22] | Tabular Operations Integration | Iterative Reasoning Framework | Context Integration Challenges |
| TEMED-LLM[59] | Reasoning Guidelines | Reasoning Correction Feedback | Noise Tolerance |
| SIMON[60] | Character-level Cnns | Iterative Reasoning | Misspellings, Informal Language |
| C2[46] | Ensemble Approach | Feedback Loops | Noise Tolerance |
| WIM[23] | - | - | - |
| DFE[34] | Deep Neural Networks | Iterative Reasoning | Noise Tolerance |
| UTN[50] | Vision Guider | Feedback Loops | Informal Language |

Table 2: Overview of various methodologies for semantic interpretation and integration in table processing, detailing their integration approaches, reasoning frameworks, and methods for handling challenges. This table highlights the strengths and limitations of each method, offering insights into their applicability for enhancing data interpretation and extraction accuracy.

Integrating semantic understanding into table processing is essential for improving data interpretation and extraction accuracy. Advanced methodologies like TableNet unify detection and structure recognition within a deep learning framework, streamlining the extraction process [58]. The CHAIN OF-TABLE framework enables iterative reasoning over tables, fostering comprehensive semantic interpretation [22]. TEMED-LLM incorporates reasoning guidelines and validation feedback loops, enhancing data extraction from medical reports [59]. Character-level CNNs in the SIMON method improve handling of misspellings and informal language, enhancing semantic classification [60]. The C2 framework integrates multiple data sources and applies a noise-tolerant likelihood estimation framework, crucial for semantic annotation [46]. Hackmann's method of systematically masking prompt words enhances LLM explainability by evaluating word impact on outputs [23]. Wu's deep embedding framework captures complex feature relationships, facilitating nuanced data interpretation [34]. The UniTabNet model combines visual and language guidance, improving textual semantic understanding in table processing [50]. Table 2 provides a comprehensive comparison of different methods for semantic interpretation and integration in table processing, emphasizing their integration approaches, reasoning frameworks, and strategies for handling common challenges.

# 4 Prompt Engineering for Language Models

## 4.1 Role and Importance of Prompt Engineering

Prompt engineering is crucial for enhancing language model performance, particularly in complex tabular contexts. Strategic prompt design enables models to excel in tasks from data visualization to numerical reasoning. The SynLM model, utilizing transformer architecture for synthetic data generation, exemplifies prompt engineering's role in model training [1]. TabSQLify further underscores its importance by employing text-to-SQL generation to decompose large tables into contextually relevant sub-tables, enhancing reasoning capabilities [3].

As illustrated in Figure 3, the role of prompt engineering in enhancing language model performance spans various domains. This figure categorizes applications into model training, reasoning enhancement, and feature embedding, thereby highlighting the transformative impact of prompt engineering techniques in tabular data processing. Cross-Table Masked Pretraining (CM2) showcases prompt engineering's effectiveness in pretraining, leveraging NLP and CV techniques to learn from vast online tabular data [7]. H-STAR's integration of symbolic and semantic reasoning further emphasizes prompt engineering's role in enhancing tabular reasoning [4]. The LLM-driven Dataset Simulation (LLMDS) method illustrates how prompt engineering defines hypotheses for precise data generation [9]. TableParser's use of weak supervision through spreadsheet annotations exemplifies the critical role of prompt engineering in domain-adaptive model training [36].

The deep embedding framework enhances feature embeddings for numerical and categorical data, impacting model performance and highlighting prompt engineering's importance [34]. UniTabNet's performance on diverse datasets, particularly in processing tables with descriptive content, demonstrates effective visual and textual integration through prompt engineering [50]. These examples collectively illustrate prompt engineering's transformative impact, enhancing model performance and facilitating accurate tabular data interpretations across domains. Empirical studies show improved data generation quality and efficiency through context-enriched prompts, while innovative techniques like reinforcement learning for column sequencing optimize complex dataset processing [23, 61, 62, 28, 25].

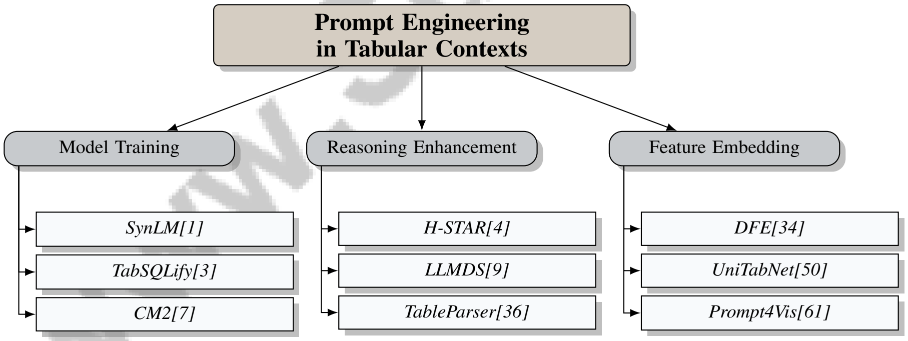


Figure 3: This figure illustrates the role of prompt engineering in enhancing language model performance across different domains. It categorizes the applications into model training, reasoning enhancement, and feature embedding, highlighting the transformative impact of prompt engineering techniques in tabular data processing.

## 4.2 Strategies for Crafting Effective Prompts

Designing effective prompts for language models, especially for tabular data, involves strategic approaches that enhance interpretative and generative capabilities. The TAP4LLM framework exemplifies this by incorporating table sampling, augmentation, and packing to improve interactions between LLMs and tabular data [63]. This framework enhances understanding of tabular data structure and semantics, facilitating more accurate responses.

Multistage inference methods like LRwBins, using logistic regression on binned feature subsets, allow rapid processing of substantial input portions, optimizing prompt-based interactions with tabular data

[64]. By segmenting input into manageable subsets, this method improves focus on relevant features, enhancing output quality.

Incorporating domain-specific knowledge into prompt design boosts language model performance and output quality by bridging tabular data and textual descriptions, enhancing transparency and reliability [61, 62, 23, 65]. Tailoring prompts with relevant context and examples guides models toward precise interpretations, improving accuracy in handling complex queries and diverse data formats.

### 4.3 Impact on Model Performance

Prompt engineering significantly influences language model performance by shaping interpretation and generation of outputs from tabular data. Context-enriched prompts enhance data generation quality and training efficiency, achieving comparable performance with fewer epochs than baseline methods [62]. The LaTeX serialization framework optimizes memory usage, improving complex data structure processing and overall model performance [66].

Language models as weak learners encode data knowledge into natural language summaries, enhancing classification performance [67]. This is crucial for handling textual and tabular data, as shown by the TTGen table-to-text generator integrated with the K-BERT MRC model, yielding promising results in complex question answering [68].

Despite advancements, a unified evaluation framework for textual data complexity is lacking, necessitating tailored evaluation methods for XAI techniques. Establishing comprehensive guidelines for improving XAI techniques is essential for accurate tabular data interpretation and output generation. Existing XAI methods primarily target image and text data, leaving a gap in applicability to tabular formats. Inadequate evaluation of XAI techniques in tabular contexts leads to potential misinterpretations and reduced transparency. Developing targeted guidelines for tabular data challenges will enhance language models' functionality in tasks like classification, regression, and imputation. Innovative approaches improving contextual understanding of tabular data can significantly enhance precision in complex queries, fostering effective XAI integration in real-world applications [55, 12, 69].

Frameworks like AIGT and DEVTC in table-to-text generation tasks illustrate prompt engineering's impact on model performance. AIGT's use of table metadata enhances contextual understanding and generation quality [70], while DEVTC leverages logic-type supervision for output diversity and factual accuracy [71]. The structured approach of C.L.E.A.R emphasizes evidence extraction and logical reasoning, improving model performance in temporal reasoning tasks [72].

Prompt engineering is vital for optimizing language model performance in processing complex tabular data. Enriching prompts with domain-specific insights and employing innovative construction protocols—such as Expert-guided, LLM-guided, and Novel-Mapping—significantly improve data generation quality and efficiency. Understanding the statistical impact of individual words in prompts enhances transparency and reliability, addressing concerns about models' "black box" nature. Automatic prompt generation systems using reinforcement learning and cell-level similarity methods streamline handling extensive tabular datasets, enhancing performance in tasks like data imputation, error detection, and entity matching [62, 23, 25].

## 5 Datasets for Tabular Data Analysis

### 5.1 Types of Datasets in Tabular Data Analysis

Tabular data analysis relies on a diverse range of datasets that facilitate the development and evaluation of analytical methodologies. These datasets are pivotal for assessing the generalization capabilities of language models across various domains. The AnaMeta dataset, with its 467,000 tables and supervision labels for field metadata, exemplifies datasets that support metadata tasks in tabular data analysis [5]. OpenTabs, comprising around 46 million samples, is instrumental in large-scale pretraining and model performance evaluation [7].

Datasets such as WikiTQ, FeTaQA, TabFact, and WikiSQL are crucial for evaluating methods like TabSQLify, which enhances reasoning capabilities by benchmarking against several baseline methods

[3]. The H-STAR framework demonstrates significant improvements over state-of-the-art methods across multiple datasets, underscoring its efficacy in tabular reasoning [4].

Real-world datasets, including the Zoo Dataset and demographic datasets, are vital for validating LLM-generated data accuracy, emphasizing the role of synthetic data in analysis [9]. PubTabNet, PubTables1M, WTW, and iFLYTAB illustrate the integration of visual and textual information, enhancing understanding of table structures and semantics [50]. The diversity in table structures found in datasets for characterizing relational table embeddings allows for comprehensive evaluation [6].

As depicted in Figure 4, this figure illustrates the hierarchical categorization of datasets used in tabular data analysis, divided into metadata, reasoning, and visual-textual datasets, highlighting key examples from recent research. These datasets collectively advance robust model development capable of addressing tabular data complexities across diverse applications. Their variety enhances analytical methodologies, enabling precise interpretations of structured data and supporting the advancement of techniques like deep learning and synthetic data generation. Addressing challenges such as dataset heterogeneity and complex data distributions, these resources enable researchers to gain deeper insights into data relationships and improve analytical tool performance, leading to impactful outcomes across domains [26, 12, 43, 29].

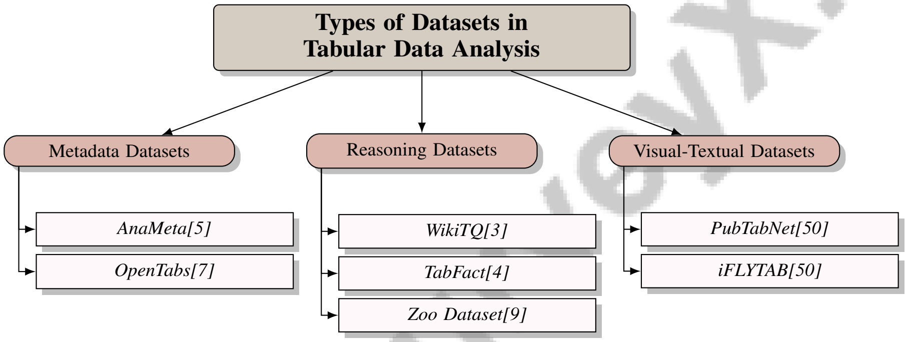


Figure 4: This figure illustrates the hierarchical categorization of datasets used in tabular data analysis, divided into metadata, reasoning, and visual-textual datasets, highlighting key examples from recent research.

### 5.2 Challenges in Dataset Compilation and Utilization

The compilation and utilization of datasets in tabular data analysis face significant challenges related to data quality, diversity, and real-world representativeness. A major challenge is the reliance on large labeled datasets, crucial for training robust models yet difficult to obtain, especially for complex queries and domain-specific applications [49]. Models like TableNet require manually annotated datasets for effective training, limiting their applicability across diverse document types [58].

The AnaMeta benchmark addresses the need for accurate field metadata understanding by providing comprehensive supervision labels [5]. However, variability in data interpretation and biases often result in model performance inconsistencies, necessitating robust evaluation metrics and methodologies [49].

Handling heterogeneous data highlights the importance of domain-specific understanding in tabular data analysis [33]. This understanding is vital for leveraging datasets in practical applications, accurately representing real-world data complexities. Dataset limitations can affect the generalizability of methods like automatic prompt generation, where outcomes depend on the training and evaluation datasets.

To overcome these challenges, research must prioritize methodologies that enhance dataset comprehensiveness, diversity, and real-world relevance. This includes developing large-scale datasets like FINDSum for summarizing long texts and multi-table data, employing advanced techniques to enrich tabular data in retrieval systems, and designing data lakes that integrate textual and tabular information for nuanced analyses. Such efforts ensure datasets capture a wide array of information, reflecting real-world complexities, thereby improving data-driven insights' accuracy and informativeness [12, 14, 73]. Strategies must be developed to reduce dependency on large labeled datasets, enhance domain-specific understanding, and establish robust evaluation frameworks to advance tabular data analysis.

### 5.3 Notable Datasets and Benchmarks

| Benchmark | Size | Domain | Task Format | Metric |
|---|---|---|---|---|
| TableBench[74] | 886 | Numerical Reasoning | Table Question Answering | ROUGE-L |
| LD-FD[75] | 4,908 | Clinical Trials | Dependency Preservation | Q-function, FDTool |
| CTSD[76] | 1,000,000 | Tabular Data | Synthetic Data Detection | AUC, Accuracy |
| LLM-Fairness[77] | 1,000,000 | Fairness IN Machine Learning | Classification | Accuracy, F1 |
| TabLM[78] | 284,807 | Finance | Binary Classification | Accuracy, F1 |
| CTSC[79] | 200 | Table Summarization | Summarization | ROUGE-L, BLEU |
| AnnotatedTables[80] | 405,616 | Tabular Data | Tabular Classification | AUROC, Cross Entropy |
| CauTabBench[81] | 17,000 | Causal Inference | Causal Inference Tasks | SHD, AUC |

Table 3: This table presents a comprehensive overview of significant benchmarks utilized in tabular data analysis, detailing their size, domain, task format, and evaluation metrics. These benchmarks serve as critical tools for assessing and enhancing language models and analytical methodologies, highlighting their application across diverse domains such as numerical reasoning, clinical trials, and finance.

The progression of tabular data analysis is significantly driven by notable datasets and benchmarks, which are critical for evaluating and enhancing language models and analytical methodologies. Datasets like WikiTableQuestions (WikiTQ) and FeTaQA (FTQ) provide detailed statistics on question-answer pairs and average tokens, essential for assessing language models' reasoning and interpretive abilities [57]. Table 3 provides a detailed overview of notable benchmarks that are pivotal in advancing the field of tabular data analysis, offering insights into their characteristics and applications.

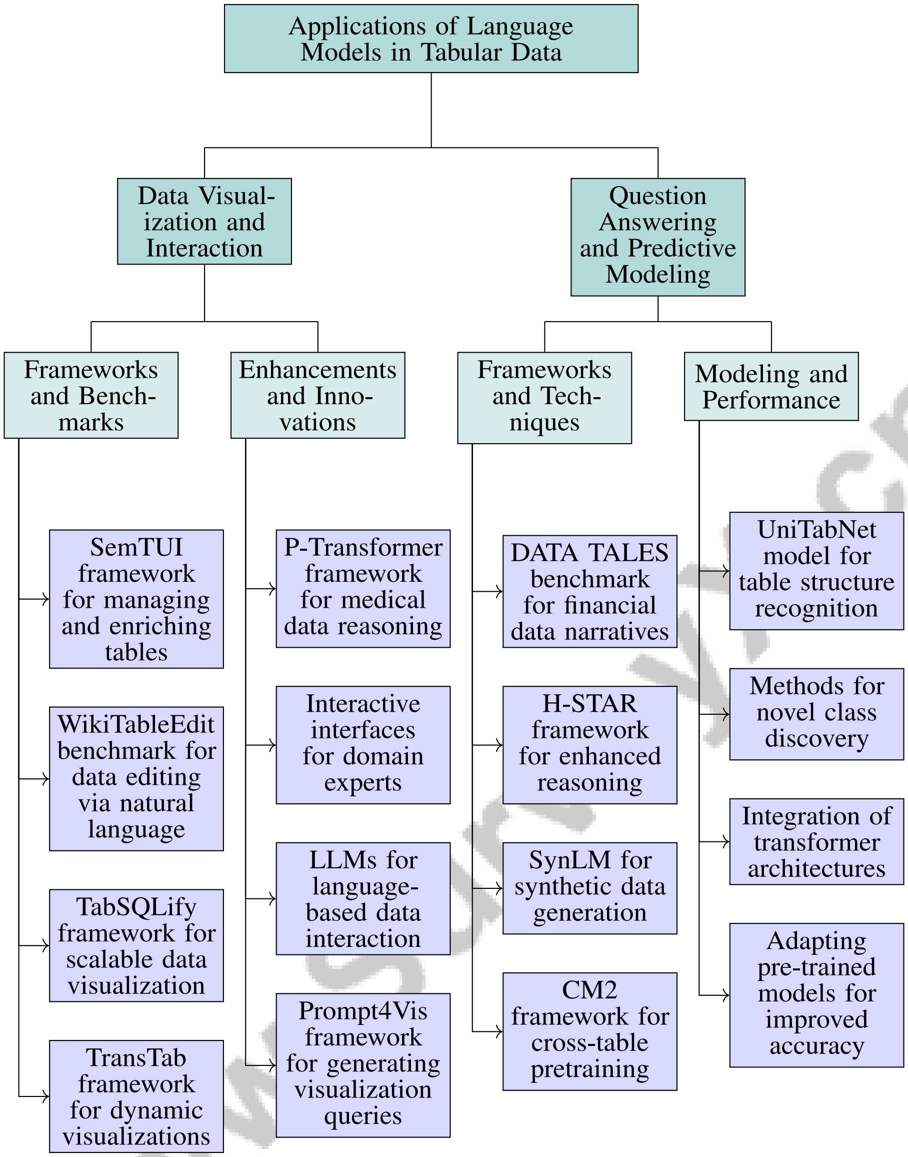
The CHAIN OF-TABLE framework demonstrates the importance of these datasets by achieving state-of-the-art results on benchmarks like WikiTQ, FeTaQA, and TabFact, showcasing its effectiveness in reasoning over tabular data [22]. These datasets push the boundaries of language models' capabilities in understanding and generating insights from structured data.

In semantic annotation, the C2 framework's evaluation using diverse datasets ensures a comprehensive assessment of its capabilities in annotating tabular data [46]. This underscores the necessity of diverse datasets in developing robust semantic annotation systems for real-world data complexities.

The TableParser system, focusing on automatic table parsing, utilizes datasets like ZHYearbooks-Excel, containing 16,041 tables, for training and fine-tuning, highlighting large-scale datasets' importance in enhancing table recognition and parsing accuracy [36].

These datasets and benchmarks provide a foundation for developing innovative methodologies, ensuring rigorous testing and validation across scenarios and applications. They play an indispensable role in shaping tabular data analysis, empowering researchers and practitioners to tackle intricate challenges, such as improving complex table queries' accuracy in retrieval-augmented generation systems and enhancing machine learning models' interpretability through innovative methods like Feature Vectors, which visualize feature interactions and provide deeper insights into data semantics [82, 12].

In recent years, the integration of language models into various domains has garnered significant attention, particularly in the context of tabular data analysis. As illustrated in Figure 5, the applications of these models can be categorized into two primary areas: data visualization and interaction, as well as question answering and predictive modeling. This figure not only highlights key frameworks and benchmarks but also underscores the enhancements and modeling techniques that have emerged. Such a comprehensive overview demonstrates the transformative impact of language models, providing a clear understanding of their potential to revolutionize how we engage with and interpret tabular data.



Applications of Language Models in Tabular Data
Data Visualization and Interaction
Question Answering and Predictive Modeling
Frameworks and Benchmarks
Enhancements and Innovations
Frameworks and Techniques
Modeling and Performance
SemTUI framework for managing and enriching tables
WikiTableEdit benchmark for data editing via natural language
TabSQLify framework for scalable data visualization
TransTab framework for dynamic visualizations
P-Transformer framework for medical data reasoning
Interactive interfaces for domain experts
LLMs for language-based data interaction
Prompt4Vis framework for generating visualization queries
DATA TALES benchmark for financial data narratives
H-STAR framework for enhanced reasoning
SynLM for synthetic data generation
CM2 framework for cross-table pretraining
UniTabNet model for table structure recognition
Methods for novel class discovery
Integration of transformer architectures
Adapting pre-trained models for improved accuracy

Figure 5: This figure illustrates the applications of language models in tabular data, categorized into data visualization and interaction, and question answering and predictive modeling. It highlights key frameworks, benchmarks, enhancements, and modeling techniques, demonstrating the transformative impact of language models in these domains.

## 6 Applications of Language Models in Tabular Data

### 6.1 Data Visualization and Interaction

Language models have revolutionized data visualization and user interaction with tabular data, facilitating intuitive exploration and interpretation of complex datasets. The SemTUI framework exemplifies this by enabling users to manage and semantically enrich tables through a web-based application, enhancing engagement [83]. Similarly, the WikiTableEdit benchmark demonstrates seamless integration of language models for data editing via natural language instructions [84].

A significant benefit of language models in data visualization is the reduction of input length for scalability, as shown by the TabSQLify framework, which enhances interpretability by focusing reasoning on relevant data, particularly in datasets with high column redundancy [3]. This adaptability is crucial for tailoring visualizations to specific user needs, improving data-driven insights.

AI-generated, for reference only.

The TransTab framework highlights language models' role in data visualization by excelling in various tasks, especially in supervised and transfer learning scenarios, thereby facilitating dynamic visualizations of variable-column tabular data [85]. Moreover, language models improve reasoning in semi-structured data contexts, as demonstrated by the P-Transformer framework's application in medical tasks like predicting surgical duration and patient outcomes [28]. An interactive interface developed by [33] further enhances domain experts' ability to visualize and interpret tabular data.

The integration of large language models (LLMs) into data visualization frameworks has transformed user engagement, moving beyond traditional querying to more intuitive, language-based interfaces. This enhances accessibility and effectiveness in deriving insights. With capabilities like those of ChatGPT, users can engage with data through natural language queries, leading to adaptive visualizations and improved reasoning. Recent studies highlight LLMs' effectiveness in generating visual representations from natural language descriptions, addressing challenges in visualizing complex datasets, and outperforming previous methods. Innovations such as the Prompt4Vis framework leverage in-context learning and schema filtering, achieving notable improvements in generating data visualization queries and setting new benchmarks [61, 49, 86].

As illustrated in Figure 6, the categorization of various frameworks, techniques, and innovations in the field of data visualization and interaction highlights key contributions and user interaction enhancements, further underscoring the transformative impact of these advancements.

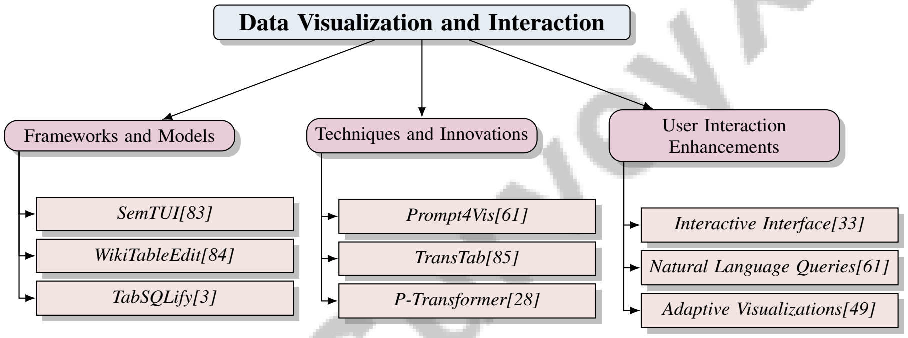

Data Visualization and Interaction
Frameworks and Models
Techniques and Innovations
User Interaction Enhancements
SemTUI[83]
WikiTableEdit[84]
TabSQLify[3]
Prompt4Vis[61]
TransTab[85]
P-Transformer[28]
Interactive Interface[33]
Natural Language Queries[61]
Adaptive Visualizations[49]

Figure 6: This figure illustrates the categorization of various frameworks, techniques, and innovations in the field of data visualization and interaction, highlighting key contributions and user interaction enhancements.

### 6.2 Question Answering and Predictive Modeling

Language models have significantly advanced question answering and predictive modeling with tabular data, enhancing both performance and interpretability. The DATA TALES benchmark illustrates this progress by evaluating models' abilities to transform complex financial data into coherent narratives, showcasing language models' potential in these domains [2]. The H-STAR framework, utilizing large language models like Gemini-1.5-Flash and GPT-4, enhances reasoning capabilities in question answering [4].

Synthetic data generation techniques, such as SynLM, provide realistic datasets for training and evaluation, boosting language models' utility in question answering and predictive modeling [1]. LLMs' ability to simulate datasets revealing interesting patterns further supports their application, enabling nuanced data-driven insights [9].

The CM2 framework achieves state-of-the-art performance across downstream tasks, validating cross-table pretraining's feasibility and implications for improving question answering and predictive modeling [7]. Analysis of embeddings in relational tables reveals varying strengths among models, with some demonstrating robustness to changes in row and column order, critical for effective predictive modeling [6].

The UniTabNet model shows state-of-the-art performance in table structure recognition across datasets, validating its integration of visual and textual understanding in processing table images, essential for accurate question answering and predictive modeling [50]. Additionally, methods for discovering novel classes rely on the assumption that similar instances in latent space belong to the same class, facilitating informed clustering and enhancing predictive modeling accuracy [87].

These advancements underscore the transformative role of language models in question answering and predictive modeling with tabular data. By leveraging advanced modeling techniques and comprehensive evaluation frameworks, language models adeptly navigate tabular data's intricacies. This evolution, driven by integrating transformer architectures and large language models, enhances predictive performance and robustness across applications. Recent studies emphasize adapting pretrained models like BERT and introducing innovative approaches combining textual and symbolic reasoning, improving accuracy in interpreting complex data structures and offering insights into structural normalization methods, achieving state-of-the-art performance in tasks like table-based question answering [11, 15].

# 7 Automated Data Analysis and Insights Extraction

The advent of AI technologies has significantly transformed automated data analysis, particularly in the realm of tabular data processing and interpretation. This section explores how AI not only automates data analysis but also advances feature engineering, highlighting the frameworks and methodologies that drive these innovations. The following subsection will delve into the specific techniques and innovations that characterize automated data analysis and feature engineering, emphasizing their practical applications across various domains.

## 7.1 Automated Data Analysis and Feature Engineering

AI substantially enhances the efficiency of data analysis and feature engineering. The HySem framework, for instance, converts unstructured HTML tables into semantic JSON, streamlining feature engineering and facilitating efficient data analysis [88]. The P-Transformer utilizes prompt-based multimodal learning to generate patient embeddings, thereby improving prediction accuracy in medical data analysis [28]. Methodologies such as SAFS minimize computational inefficiencies and optimize feature selection, crucial for automating feature engineering [53]. The JarviX platform exemplifies AI's role in automating data analysis by leveraging large language models (LLMs) to enhance data processing and feature engineering, thus improving accuracy and efficiency [27].

The EPIC framework highlights the benefits of model-agnostic approaches in generating high-quality synthetic data, which enhances classification performance and facilitates automated analysis [89]. The DP-TabICL method underscores privacy in automated analysis by producing differentially private demonstration examples, ensuring performance while safeguarding sensitive information [90]. Small language models have shown potential in outperforming traditional machine learning methods on tabular data by reducing preprocessing requirements [45]. The semi-automated role-playing framework in the Tabular Embedding Model (TEM) underscores AI's capacity to generate diverse training questions, enhancing data analysis automation [91].

Interactive interfaces further aid automated data analysis by enabling domain experts to perform novel class discovery (NCD) and clustering algorithms without coding, expediting dataset curation and hypothesis testing. As illustrated in Figure 7, the hierarchical categorization of advancements in automated data analysis and feature engineering encompasses frameworks and models, platforms and interfaces, and innovative applications that leverage AI technologies to enhance efficiency and accuracy in processing tabular data. Future research should expand benchmarks to include diverse reasoning tasks and datasets, while developing robust evaluation methods to enhance automation in data analysis and feature engineering [49]. Collectively, these advancements highlight AI's transformative impact on automating data analysis and feature engineering, promoting efficient, transparent, and insightful processing of tabular data across various sectors.

## 7.2 Improving Insights Extraction

Enhancing insights extraction from tabular data through AI involves advanced methodologies that boost the interpretive capabilities of language models and AI systems. The CHAIN OF-TABLE framework employs iterative reasoning, improving insights extraction from complex datasets for more nuanced interpretations [22]. Frameworks such as TableNet and UniTabNet integrate semantic understanding into table processing, enhancing accuracy in data interpretation and insights extraction.

AI-generated, for reference only.

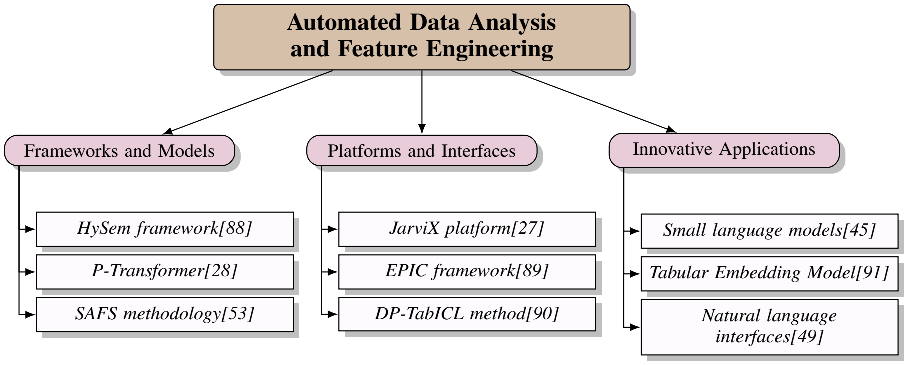


Figure 7: This figure illustrates the hierarchical categorization of advancements in automated data analysis and feature engineering. It highlights frameworks and models, platforms and interfaces, and innovative applications that leverage AI technologies to enhance efficiency and accuracy in processing tabular data.

Language models like SynLM and H-STAR further demonstrate the potential for enhancing insights extraction by generating synthetic data that mirrors original datasets, facilitating more precise analysis [7]. Cross-table pretraining techniques in the CM2 framework bolster language models' generalization capabilities, enabling effective insights extraction across diverse datasets.

Interactive interfaces that allow domain experts to visualize and manipulate tabular data, such as the one developed by [33], enhance insights extraction by promoting intuitive exploration of data patterns. Additionally, simulating datasets with LLMs, as demonstrated by LLMDS, supports rapid hypothesis testing and latent pattern exploration, contributing to effective insights extraction [9]. The integration of advanced AI methodologies and interactive tools is crucial for optimizing insights extraction from tabular data. Techniques such as context enrichment and the Retrieval-Augmented Generation (RAG) architecture have significantly improved the accuracy of complex table queries, particularly in challenging formats like PDFs. Frameworks like SemTUI facilitate seamless dataset integration and enrichment from diverse sources, enhancing efficiency and accessibility in analysis. These innovations enable precise, efficient, and insightful analyses across various domains, transforming how organizations leverage tabular data for decision-making and trend analysis [92, 12, 83]. Ongoing research is essential for refining AI systems' capabilities in interpreting and extracting valuable insights from structured data.

### 7.3 Enhancements in Decision-Making

The integration of automated analysis into decision-making processes significantly enhances operational efficiency by delivering timely and accurate insights from tabular data. AI-driven methodologies, such as those in the JarviX platform, underscore the potential of LLMs to automate data processing and feature engineering, thereby supporting informed decision-making [27]. Synthetic data generation techniques, exemplified by the SynLM model, aid decision-making by creating realistic datasets that preserve privacy while retaining original data utility [1]. Frameworks like H-STAR improve decision-making by integrating symbolic and semantic reasoning, enhancing output interpretability and accuracy in complex data scenarios [4]. The LLM-driven Dataset Simulation (LLMDS) method facilitates rapid hypothesis testing and latent pattern exploration, fostering agile decision-making processes [9].

Interactive interfaces, such as the one developed by [33], empower domain experts to visualize and manipulate tabular data, enhancing both decision-making and result interpretation. The CHAIN OF-TABLE framework, employing iterative reasoning, further supports nuanced decision-making by improving insights extraction from complex datasets [22]. Advancements in automated analysis and AI-driven methodologies are pivotal for enhancing decision-making processes through accurate, efficient, and insightful analyses of tabular data. These developments highlight AI's profound influence on operational efficiency, enabling organizations to leverage sophisticated data management techniques, such as those in the AUDAL data lake, thus enhancing decision-making across diverse sectors. By integrating textual and tabular data and employing innovative methodologies like context enrichment for complex table queries, organizations can achieve more informed and timely insights, ultimately driving improved outcomes in their operations [12, 14].

# 8 Conclusion

## 8.1 Challenges and Future Directions

The domain of tabular data understanding integrated with language model applications faces numerous challenges that demand continuous exploration and innovation. A critical issue is enhancing the interpretability and robustness of large language models (LLMs) in handling tabular data, particularly when dealing with imbalanced labels and integrating multi-modal data. Future research should emphasize refining LLM embeddings to better capture the nuances of relational data.

Progress in synthetic data generation is contingent upon the advancement of pre-trained models and the optimization of hyper-parameters, ensuring that generated datasets mirror the original data's distribution while maintaining privacy. Moreover, expanding the CM2 framework to cover a broader range of data domains and exploring additional pretraining objectives are crucial for improving cross-table learning performance.

Efficiently managing larger datasets through optimized table preprocessing methods is another vital area, as demonstrated by frameworks like TabSQLify. Further exploration into optimization techniques to enhance the efficiency and scalability of table processing is warranted.

In table parsing, enhancing the precision of parsing complex table structures remains a priority. Future enhancements to models like TableNet should include additional functionalities for row identification and the incorporation of semantic features to improve performance. Similarly, broadening the application of frameworks like H-STAR to more complex data structures and evaluating their efficacy across different languages presents promising research avenues.

The reliability of simulated datasets also requires attention, with future efforts directed towards developing automated systems for continuous hypothesis generation and testing, thereby increasing the practical utility of simulated data. Enhancing the taxonomy of field metadata and improving LLMs' capabilities in metadata extraction are pivotal for advancing tabular data analysis.

Finally, exploring optimization strategies for embedding modules and undertaking comparative analyses with other machine learning algorithms are essential for advancing deep feature embedding in tabular data analysis. Addressing these challenges through a multidisciplinary lens can significantly improve the capabilities and application of language models in tabular data understanding, enabling more precise and insightful data-driven decision-making.

## References

[1] Alexandre Sablayrolles, Yue Wang, and Brian Karrer. Privately generating tabular data using language models, 2023.

[2] Yajing Yang, Qian Liu, and Min-Yen Kan. Datatales: A benchmark for real-world intelligent data narration, 2024.

[3] Md Mahadi Hasan Nahid and Davood Rafiei. Tabsqlify: Enhancing reasoning capabilities of llms through table decomposition, 2024.

[4] Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K. Reddy. H-star: Llm-driven hybrid sql-text adaptive reasoning on tables, 2024.

[5] Xinyi He, Mengyu Zhou, Mingjie Zhou, Jialiang Xu, Xiao Lv, Tianle Li, Yijia Shao, Shi Han, Zejian Yuan, and Dongmei Zhang. Anameta: A table understanding dataset of field metadata knowledge shared by multi-dimensional data analysis tasks, 2023.

[6] Tianji Cong, Madelon Hulsebos, Zhenjie Sun, Paul Groth, and H. V. Jagadish. Observatory: Characterizing embeddings of relational tables, 2024.

[7] Chao Ye, Guoshan Lu, Haobo Wang, Liyao Li, Sai Wu, Gang Chen, and Junbo Zhao. Towards cross-table masked pretraining for web data mining, 2024.

[8] Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alexander J. Smola. Benchmarking multimodal automl for tabular data with text fields, 2021.

[9] Miguel Zabaleta and Joel Lehman. Simulating tabular datasets through llms to rapidly explore hypotheses about real-world entities, 2024.

[10] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey, 2024.

[11] Tianyang Liu, Fei Wang, and Muhao Chen. Rethinking tabular data understanding with large language models, 2023.

[12] Uday Allu, Biddwan Ahmed, and Vishesh Tripathi. Beyond extraction: Contextualising tabular data for efficient summarisation by language models, 2024.

[13] Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, and Alan Ritter. Schema-driven information extraction from heterogeneous tables, 2024.

[14] Pegdwendé Sawadogo, Jérôme Darmont, and Camille Noûs. Joint management and analysis of textual documents and tabular data within the audal data lake, 2021.

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

[16] Jaehyun Nam, Kyuyoung Kim, Seunghyuk Oh, Jihoon Tack, Jaehyung Kim, and Jinwoo Shin. Optimized feature generation for tabular data via llms with decision tree reasoning, 2024.

[17] Ruiyu Wang, Zifeng Wang, and Jimeng Sun. Unipredict: Large language models are universal tabular classifiers, 2024.

[18] Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey, 2022.

[19] Aneta Koleva, Martin Ringsquandl, Mark Buckley, Rakebul Hasan, and Volker Tresp. Named entity recognition in industrial tables using tabular language models, 2022.

[20] Yilun Zhao, Haowei Zhang, Shengyun Si, Linyong Nan, Xiangru Tang, and Arman Cohan. Investigating table-to-text generation capabilities of llms in real-world information seeking scenarios, 2023.

[21] Hiroshi Iida, Dung Thai, Varun Manjunatha, and Mohit Iyyer. Tabbie: Pretrained representations of tabular data, 2021.

[22] 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. Chain-of-table: Evolving tables in the reasoning chain for table understanding, 2024.

[23] Stefan Hackmann, Haniyeh Mahmoudian, Mark Steadman, and Michael Schmidt. Word importance explains how prompts affect language model outputs, 2024.

[24] Sérgio Jesus, José Pombal, Duarte Alves, André Cruz, Pedro Saleiro, Rita P. Ribeiro, João Gama, and Pedro Bizarro. Turning the tables: Biased, imbalanced, dynamic tabular datasets for ml evaluation, 2022.

[25] Ashlesha Akella, Abhijit Manatkar, Brij Chavda, and Hima Patel. An automatic prompt generation system for tabular data tasks, 2024.

[26] Fionn Murtagh. Origins of modern data analysis linked to the beginnings and early development of computer science and information engineering, 2008.

[27] Shang-Ching Liu, ShengKun Wang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, Tsungyao Chang, and Jianwei Zhang. Jarvix: A llm no code platform for tabular data analysis and optimization, 2023.

[28] Yucheng Ruan, Xiang Lan, Daniel J. Tan, Hairil Rizal Abdullah, and Mengling Feng. P-transformer: A prompt-based multimodal transformer architecture for medical tabular data, 2024.

[29] Maria F. Davila R., Sven Groen, Fabian Panse, and Wolfram Wingerath. Navigating tabular data synthesis research: Understanding user needs and tool capabilities, 2024.

[30] Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Avantika Sharma, Neeraj Gunda, and Sai Krishna Bala. Evaluation of table representations to answer questions from tables in documents : A case study using 3gpp specifications, 2024.

[31] Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts), 2021.

[32] Akash Ghosh, B Venkata Sahith, Niloy Ganguly, Pawan Goyal, and Mayank Singh. How robust are the tabular qa models for scientific tables? a study using customized dataset, 2024.

[33] Colin Troisemaine, Joachim Flocon-Cholet, Stéphane Gosselin, Alexandre Reiffers-Masson, Sandrine Vaton, and Vincent Lemaire. An interactive interface for novel class discovery in tabular data, 2023.

[34] Yuqian Wu, Hengyi Luo, and Raymond S. T. Lee. Deep feature embedding for tabular data, 2024.

[35] Dmitry Medvedev and Alexander D'yakonov. New properties of the data distillation method when working with tabular data, 2020.

[36] Susie Xi Rao, Johannes Rausch, Peter Egger, and Ce Zhang. Tableparser: Automatic table parsing with weak supervision from spreadsheets, 2022.

[37] Yevgeni Berkovitch, Oren Glickman, Amit Somech, and Tomer Wolfson. Generating tables from the parametric knowledge of language models, 2024.

[38] Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data, 2020.

[39] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. Tablegpt2: A large multimodal model with tabular data integration, 2024.

[40] Shriyank Somvanshi, Subasish Das, Syed Aaqib Javed, Gian Antariksa, and Ahmed Hossain. A survey on deep tabular learning, 2024.

[41] Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox, 2024.

[42] Xinyi He, Mengyu Zhou, Xinrun Xu, Xiaojun Ma, Rui Ding, Lun Du, Yan Gao, Ran Jia, Xu Chen, Shi Han, Zejian Yuan, and Dongmei Zhang. Text2analysis: A benchmark of table question answering with advanced data analysis and unclear queries, 2023.

[43] Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets, 2025.

[44] Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C. Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, and Micah Goldblum. Transfer learning with deep tabular models, 2023.

[45] Benjamin L. Badger. Small language models for tabular data, 2022.

[46] Udayan Khurana and Sainyam Galhotra. Semantic annotation for tabular data, 2020.

[47] Andrej Tschalzev, Sascha Marton, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. A data-centric perspective on evaluating machine learning models for tabular data, 2024.

[48] Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data, 2023.

[49] Weixu Zhang, Yifei Wang, Yuanfeng Song, Victor Junqiu Wei, Yuxing Tian, Yiyan Qi, Jonathan H. Chan, Raymond Chi-Wing Wong, and Haiqin Yang. Natural language interfaces for tabular data querying and visualization: A survey, 2024.

[50] Zhenrong Zhang, Shuhang Liu, Pengfei Hu, Jiefeng Ma, Jun Du, Jianshu Zhang, and Yu Hu. Unitabnet: Bridging vision and language models for enhanced table structure recognition, 2024.

[51] Zhenrong Zhang, Jianshu Zhang, and Jun Du. Split, embed and merge: An accurate table structure recognizer, 2022.

[52] Bin Xiao, Murat Simsek, Burak Kantarci, and Ala Abu Alkheir. Table structure recognition with conditional attention, 2022.

[53] Girmaw Abebe Tadesse, William Ogallo, Celia Cintas, and Skyler Speakman. Model-free feature selection to facilitate automatic discovery of divergent subgroups in tabular data, 2022.

[54] Ruya Jiang, Chun Wang, and Weihong Deng. Seek and solve reasoning for table question answering, 2024.

[55] Mythreyi Velmurugan, Chun Ouyang, Yue Xu, Renuka Sindhgatta, Bemali Wickramanayake, and Catarina Moreira. Developing guidelines for functionally-grounded evaluation of explainable artificial intelligence using tabular data, 2024.

[56] Shraddha Barke, Christian Poelitz, Carina Suzana Negreanu, Benjamin Zorn, José Cambronero, Andrew D. Gordon, Vu Le, Elnaz Nouri, Nadia Polikarpova, Advait Sarkar, Brian Slininger, Neil Toronto, and Jack Williams. Solving data-centric tasks using large language models, 2024.

[57] Yuxiang Wang, Jianzhong Qi, and Junhao Gan. Accurate and regret-aware numerical problem solver for tabular question answering, 2025.

[58] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, and Lovekesh Vig. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images, 2020.

[59] Aleksa Bisercic, Mladen Nikolic, Mihaela van der Schaar, Boris Delibasic, Pietro Lio, and Andrija Petrovic. Interpretable medical diagnostics with structured data extraction by large language models, 2023.

[60] Paul Azunre, Craig Corcoran, Numa Dhamani, Jeffrey Gleason, Garrett Honke, David Sullivan, Rebecca Ruppel, Sandeep Verma, and Jonathon Morgan. Semantic classification of tabular datasets via character-level convolutional neural networks, 2019.

[61] Shuaimin Li, Xuanang Chen, Yuanfeng Song, Yunze Song, and Chen Zhang. Prompt4vis: Prompting large language models with example mining and schema filtering for tabular data visualization, 2024.

[62] Banooqa Banday, Kowshik Thopalli, Tanzima Z. Islam, and Jayaraman J. Thiagarajan. On the role of prompt construction in enhancing efficacy and efficiency of llm-based tabular data generation, 2024.

[63] Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. Tap4llm: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning, 2024.

[64] Daniel S Johnson and Igor L Markov. Efficient multi-stage inference on tabular data, 2023.

[65] Zhixin Guo, Minyxuan Yan, Jiexing Qi, Jianping Zhou, Ziwei He, Guanjie Zheng, and Xinbing Wang. Adapting knowledge for few-shot table-to-text generation, 2024.

[66] Sukriti Jaitly, Tanay Shah, Ashish Shugani, and Razik Singh Grewal. Towards better serialization of tabular data for few-shot classification with large language models, 2023.

[67] Hariharan Manikandan, Yiding Jiang, and J Zico Kolter. Language models are weak learners, 2023.

[68] Xiao Li, Yawei Sun, and Gong Cheng. Tsqa: Tabular scenario based question answering, 2021.

[69] Yazheng Yang, Yuqi Wang, Yaxuan Li, Sankalok Sen, Lei Li, and Qi Liu. Unleashing the potential of large language models for predictive tabular tasks in data science, 2025.

[70] Mingming Zhang, Zhiqing Xiao, Guoshan Lu, Sai Wu, Weiqiang Wang, Xing Fu, Can Yi, and Junbo Zhao. Aigt: Ai generative table based on prompt, 2024.

[71] Yotam Perlitz, Liat Ein-Dor, Dafna Sheinwald, Noam Slonim, and Michal Shmueli-Scheuer. Diversity enhanced table-to-text generation via type control, 2023.

[72] Irwin Deng, Kushagra Dixit, Vivek Gupta, and Dan Roth. Enhancing temporal understanding in llms for semi-structured tables, 2024.

[73] Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. Long text and multi-table summarization: Dataset and method, 2023.

[74] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, Guanglin Niu, Tongliang Li, and Zhoujun Li. Tablebench: A comprehensive and complex benchmark for table question answering, 2024.

[75] Chaithra Umesh, Kristian Schultz, Manjunath Mahendra, Saparshi Bej, and Olaf Wolkenhauer. Preserving logical and functional dependencies in synthetic tabular data, 2024.

[76] G. Charbel N. Kindji, Lina Maria Rojas-Barahona, Elisa Fromont, and Tanguy Urvoy. Cross-table synthetic tabular data detection, 2024.

[77] Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications, 2024.

[78] Kyoka Ono and Simon A. Lee. Text serialization and their relationship with the conventional paradigms of tabular machine learning, 2024.

[79] Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. Summarizing and exploring tabular data in conversational search, 2020.

[80] Yaojie Hu, Ilias Fountalis, Jin Tian, and Nikolaos Vasiloglou. Annotatedtables: A large tabular dataset with language model annotations, 2024.

[81] Ruibo Tu, Zineb Senane, Lele Cao, Cheng Zhang, Hedvig Kjellström, and Gustav Eje Henter. Causality for tabular data synthesis: A high-order structure causal benchmark framework, 2024.

[82] Amirata Ghorbani, Dina Berenbaum, Maor Ivgi, Yuval Dafna, and James Zou. Beyond importance scores: Interpreting tabular ml by visualizing feature semantics, 2021.

[83] Marco Ripamonti, Flavio De Paoli, and Matteo Palmonari. Semtui: a framework for the interactive semantic enrichment of tabular data, 2022.

[84] Zheng Li, Xiang Chen, and Xiaojun Wan. Wikitableedit: A benchmark for table editing by natural language instruction, 2024.

[85] Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables, 2022.

[86] Yang Wu, Yao Wan, Hongyu Zhang, Yulei Sui, Wucai Wei, Wei Zhao, Guandong Xu, and Hai Jin. Automated data visualization from natural language via large language models: An exploratory study, 2024.

[87] Colin Troisemaine, Joachim Flocon-Cholet, Stéphane Gosselin, Sandrine Vaton, Alexandre Reiffers-Masson, and Vincent Lemaire. A method for discovering novel classes in tabular data, 2022.

[88] Narayanan PP and Anantharaman Palacode Narayana Iyer. Hysem: A context length optimized llm pipeline for unstructured tabular extraction, 2024.

[89] Jinhee Kim, Taesung Kim, and Jaegul Choo. Epic: Effective prompting for imbalanced-class data synthesis in tabular data classification via large language models, 2025.

[90] Alycia N. Carey, Karuna Bhaila, Kennedy Edemacu, and Xintao Wu. Dp-tabicl: In-context learning with differentially private tabular data, 2024.

[91] Sujit Khanna and Shishir Subedi. Tabular embedding model (tem): Finetuning embedding models for tabular rag applications, 2024.

[92] Yuzhao Yang, Jérôme Darmont, Franck Ravat, and Olivier Teste. Automatic integration issues of tabular data for on-line analysis processing, 2020.

**Disclaimer:**

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.