



# AUTOPREP: Natural Language Question-Aware Data Preparation with a Multi-Agent Framework

Meihao Fan  
Renmin University of China  
fmh1art@ruc.edu.cn

Ju Fan\*  
Renmin University of China  
fanj@ruc.edu.cn

Nan Tang  
HKUST (GZ)  
nantang@hkust-gz.edu.cn

Lei Cao  
University of Arizona  
caolei@arizona.edu

Guoliang Li  
Tsinghua University  
liguoliang@tsinghua.edu.cn

Xiaoyong Du  
Renmin University of China  
duyong@ruc.edu.cn

## ABSTRACT

Answering natural language (NL) questions about tables, known as Tabular Question Answering (TQA), is crucial because it allows users to quickly and efficiently extract meaningful insights from structured data, effectively bridging the gap between human language and machine-readable formats. Many of these tables are derived from web sources or real-world scenarios, which require meticulous data preparation (or data prep) to ensure accurate responses. However, preparing such tables for NL questions introduces new requirements that extend beyond traditional data preparation. This question-aware data preparation involves specific tasks such as column derivation and filtering tailored to particular questions, as well as question-aware value normalization or conversion, highlighting the need for a more nuanced approach in this context. Because each of the above tasks is unique, a single model (or agent) may not perform effectively across all scenarios. In this paper, we propose **AUTOPREP**, a large language model (LLM)-based multi-agent framework that leverages the strengths of multiple agents, each specialized in a certain type of data prep, ensuring more accurate and contextually relevant responses. Given an NL question over a table, AUTOPREP performs data prep through three key components. **Planner**: Determines a logical plan, outlining a sequence of high-level operations. **Programmer**: Translates this logical plan into a physical plan by generating the corresponding low-level code. **Executor**: Executes the generated code to process the table. To support this multi-agent framework, we design a novel Chain-of-Clauses reasoning mechanism for high-level operation suggestion, and a tool-augmented method for low-level code generation. Extensive experiments on real-world TQA datasets demonstrate that AUTOPREP can significantly improve the state-of-the-art TQA solutions through question-aware data preparation.

## PVLDB Reference Format:

Meihao Fan, Ju Fan, Nan Tang, Lei Cao, Guoliang Li, and Xiaoyong Du. AUTOPREP: Natural Language Question-Aware Data Preparation with a Multi-Agent Framework. PVLDB, 18(10): 3504 - 3517, 2025. doi:10.14778/3748191.3748211

\*Ju Fan is the corresponding author.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing [info@vldb.org](mailto:info@vldb.org). Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 18, No. 10 ISSN 2150-8097. doi:10.14778/3748191.3748211

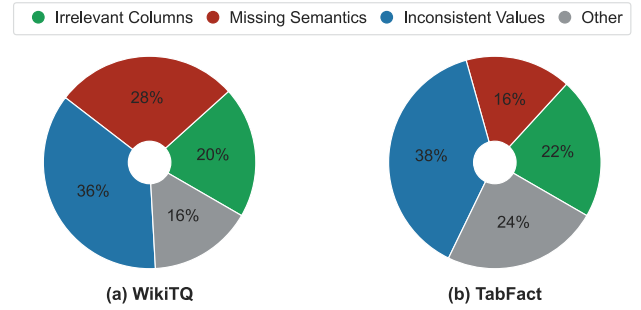


Figure 1: An error analysis of LLM-based TQA (using GPT-4) on two well-adopted datasets.

## PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/ruc-datalab/AutoPrep>.

## 1 INTRODUCTION

Tabular Question Answering (TQA) refers to the task of answering natural language (NL) questions based on provided tables [16, 17, 35, 39]. TQA empowers non-technical users such as domain scientists to easily analyze tabular data and has a wide range of applications, including table-based fact verification [15, 22, 23] and table-based question answering [37, 40]. As TQA requires NL understanding and reasoning over tables, state-of-the-art solutions [17, 50, 51, 56, 57, 59] mainly rely on large language models (LLMs).

As many tables in TQA originate from web sources or real-world data, they demand meticulous **data preparation** (or data prep) to produce accurate answers. Figure 1 shows an error analysis of an LLM-based approach (using GPT-4) across two TQA tasks: table-based question answering on the WikiTQ dataset [40] and table-based fact verification on the TabFact dataset [15] (More details of the error analysis can be found in our technical report [5]). The results indicate that 84% and 76% of the errors stem from inadequately addressing data prep issues, including *missing semantics*, *inconsistent values*, and *irrelevant columns*, as illustrated as follows.

**(1) Missing Semantics.** This data prep issue arises when a table lacks the necessary *semantics* to address the specific requirements of the NL question. That is, although some columns in the table may be related to the question, they do not directly provide the required semantic information. As shown in Figure 2a, the semantics needed for the NL questions, such as *country* and *GDP*, are not explicitly present in the tables. Therefore, to ensure accurate

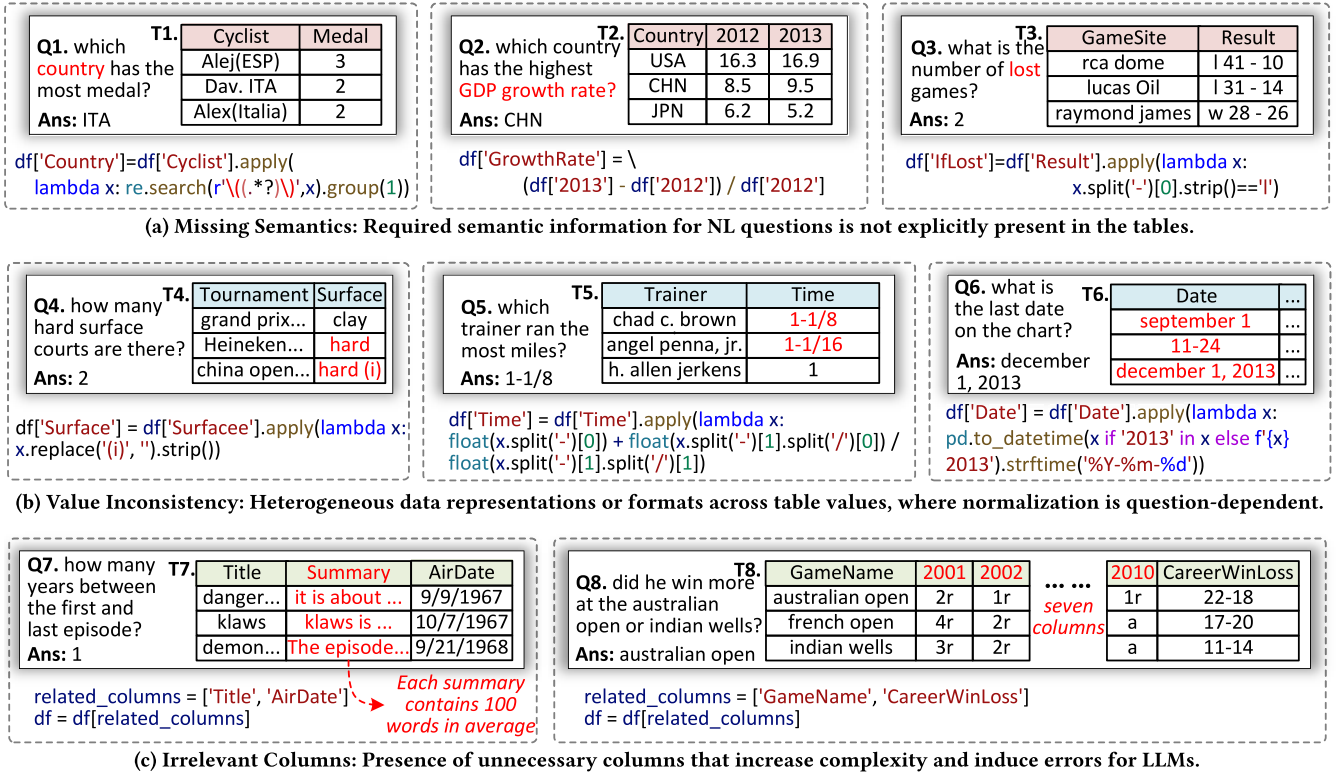


Figure 2: Examples of Data Preparation Issues for Table Question Answering.

responses in TQA, it is essential to perform **column derivation** from existing columns, such as extracting country information from text or applying mathematical operations across multiple columns.

(2) **Inconsistent Values.** This data preparation issue arises when the values within a particular column are inconsistent across different records due to varying data representations (e.g., inconsistent date formats), as illustrated in Figure 2b. This situation necessitates **column normalization**, which must be question-aware, as different questions may require different types of normalization. For instance, as  $Q_5$  requires comparing numbers in  $T_5$  but the corresponding columns are of string type, it is necessary to convert the strings in fraction format into numerical values.

(3) **Irrelevant Columns.** In some tables, although many columns are available, only a few are actually relevant to the NL question. For example, as shown in Figure 2c,  $Q_7$  does not require the Summary column in table  $T_7$ . Similarly, in  $T_8$ , although there are 12 columns, only 2 are needed to answer  $Q_8$ . Thus, **column filtering**, i.e., retaining only the columns directly related to the question, is crucial for TQA, as irrelevant columns add complexity and can mislead LLMs into producing incorrect answers.

This highlights the critical role of thorough data prep in ensuring accurate TQA. Note that, although other issues like missing values and duplicates are common in real-world data, they are less prevalent than the three highlighted issues in current TQA benchmarks. Nevertheless, we recognize their significance and leave the exploration of broader data prep challenges to future work.

**Question-Aware Data Preparation.** In this paper, we thus propose to study a new problem, namely *question-aware data preparation*. The novelty of this problem lies in the need to deeply understand the semantics of the NL question in order to guide data prep operations over the structured table. Unlike traditional data prep, which is performed *offline* and independent of downstream tasks, question-aware data prep is conducted *online* and must dynamically tailor the table to the specific demands of the question. This introduces new challenges due to the diversity and ambiguity of NL questions and the heterogeneity of tabular data, requiring precise semantic alignment between the question and the structured table. For example, as shown in Figure 2a, given question  $Q_1$  that requests country-specific information from a text column, table  $T_1$  needs to be transformed to extract and create a new Country column, which may not be considered in traditional data prep.

However, building a system to support question-aware data prep is challenging due to the semantic alignment between the table and the NL question. Specifically, different questions may demand different types of data prep, e.g., derivation, normalization, or filtering. Furthermore, even if facing the same type of data prep issue, different questions may require different ways to handle it. For example, the method that normalizes the date to a unified format is clearly different from that of normalizing strings to integers. Although LLMs show promise for interpreting semantics in NL questions, simply applying them to address all potential data prep issues has shown ineffective, often leading to false negatives and false positives, as illustrated in Section 2.3.

**AUTOPREP: A Hierarchical, Multi-Stage Approach.** To address the challenges, we propose **AUTOPREP**, which features two key ideas. First, drawing inspiration from modern DBMS, particularly the distinction between logical operations and their physical implementations, AUTOPREP separates high-level, logical data prep operations from the concrete methods used for execution. Specifically, it introduces a planning stage that generates a *logical plan* for each question, consisting of a sequence of high-level data prep operations tailored to the question’s needs, such as column derivation, normalization and filtering shown in Figure 2. In the next stage, AUTOPREP maps these logical operations to the corresponding physical implementations. This separation allows AUTOPREP to break down complex data prep tasks into smaller, more manageable sub-tasks, making the process easier to solve. Moreover, this modular design makes AUTOPREP extensible, enabling the development of specialized implementations for each type of data prep operation or the introduction of new operation types.

Second, unlike conventional DBMSs, AUTOPREP does not predefine physical operations for each logical operation. Instead, for each data prep operation in the logical plan, AUTOPREP generates a physical implementation that is specialized to the specific question *on the fly*. By considering the unique requirements of the NL question, this customized implementation ensures that the data preparation is closely aligned with the needs of different questions.

Building on the above insights, we design the AUTOPREP system with a *multi-stage architecture*.

- **The Planning Stage:** Unlike traditional DBMS, where logical operations are already available beforehand via SQL queries, AUTOPREP has to determine the appropriate logical data prep operations by analyzing the semantic alignment between the table and the NL question. This process occurs during the planning stage, where the logical plan is formed.
- **The Programming Stage:** This stage converts the logical plan into a physical plan by generating low-level executable code, selecting the appropriate programming constructs (e.g., Python functions or APIs) for each operation, and customizing the code to align the table’s structure with the NL question’s semantics.
- **The Executing Stage:** This stage executes the generated code for each operation and returns any errors encountered to the Programming stage for debugging.

We implement AUTOPREP using the popular LLM-based Multi-Agent framework [25], which leverages multiple small, independent agents working collaboratively to solve complex problems.

More specifically, we design a PLANNER agent, which corresponds to the Planning Stage and suggests a tailored sequence of high-level operations to meet the specific needs of the question, leveraging the semantics understanding and reasoning capabilities of LLMs. The core technical idea behind this PLANNER agent is a novel *Chain-of-Clauses (CoC) reasoning* method. This method translates the NL question into an *Analysis Sketch*, which outlines how the table should be prepared to produce the answer, guiding the agent’s reasoning based on this sketch. Compared to the popular Chain of Thoughts (CoT) methods [51], which decompose questions into sub-questions, our approach more effectively captures the semantic relationships between questions and tables.

AUTOPREP also includes a set of PROGRAMMER agents, each of which synthesizes a question-specific implementation for a given logical data prep operation. However, existing LLM-based code synthesis often generates overly generic code that struggles to effectively address the *heterogeneity* challenges of tables. For instance, values may have diverse syntactic formats (e.g., “September 1” and “11-24” in  $T_6$ ) or semantic representations (e.g., “ITA” and “Italia” in  $T_1$ ), making it difficult to generate code tailored to these variations. To address this, we propose a *tool-augmented* approach that enhances the LLM’s code generation capabilities by incorporating predefined API functions, which allows the LLM to generate more specialized code that accounts for variations in table values. Furthermore, corresponding to the Executing stage, we design an EXECUTOR agent that executes the code to process the table.

**Contributions.** Our contributions are summarized as follows.

- (1) We introduce a novel problem of question-aware data preparation for TQA, which is formally defined in Section 2.
- (2) We propose AUTOPREP, an LLM-based multi-agent framework for question-aware data prep (Section 3). We develop effective techniques in AUTOPREP for the PLANNER agent (Section 4) and the PROGRAMMER agents (Section 5).
- (3) We conduct a thorough evaluation on data prep in TQA (Section 6). Extensive experiments show that AUTOPREP achieves new SOTA accuracy, outperforming existing TQA methods without data prep by 12.22 points on WikiTQ and 13.23 points on TabFact, and surpassing TQA methods with data prep by 3.05 points on WikiTQ and 1.96 points on TabFact.

## 2 QUESTION-AWARE DATA PREP FOR TQA

### 2.1 Tabular Question Answering

Let  $Q$  be a natural language (NL) question, and  $T$  a table consisting of  $m$  columns (i.e., attributes)  $\{A_1, A_2, \dots, A_m\}$  and  $n$  rows  $\{r_1, r_2, \dots, r_n\}$ , where  $v_{ij}$  denotes the value in the  $i$ -th row and  $j$ -th column of the table. The problem of **tabular question answering (TQA)** is to generate an answer  $Ans$ , in response to question  $Q$  based on the information in table  $T$ . By the purposes of the questions, there are two main types of TQA problems: (1) table-based fact verification [7, 15], which determines whether  $Q$  can be *entailed* or *refuted* by  $T$ , and (2) table-based question answering [37, 40], which extracts or reasons the answer to  $Q$  from  $T$ .

**EXAMPLE 1.** Figure 2 provides several examples of TQA. Consider table  $T_1$ , which contains medal information for cyclists from different countries, with two columns: Cyclist and Medal. Given the question  $Q_1$ , “Which **country** has the most medals?”, the answer should be ITA, as two Italian cyclists, “Dav” and “Alex”, have won a total of 4 medals, more than the ESP cyclist “Alej”. TQA often requires reasoning over tables. For instance, to answer question  $Q_2$ , we first need to calculate the “GDP growth rate” for all countries, then sort the countries by growth rate, and finally identify the country with the highest GDP growth rate, i.e., CHN.

### 2.2 Data Prep for TQA

In contrast to traditional data prep, *question-aware data prep for TQA* focuses on *adapting the table  $T$  to the specific informational needs*

of a given question  $Q$ , thereby enhancing the semantic alignment between the structured table and the NL question.

**Data Prep Operations.** To meet the new requirements of data prep for TQA, this paper defines high-level, logical **data prep operations** (or *operations* for short) to formalize the *question-aware* data prep tasks. Formally, an operation, denoted as  $o$ , encapsulates a specific question-aware data prep task that transforms table  $T$  into another table  $T'$ , i.e.,  $T' = o(T)$ .

As shown in Figure 1, the majority of TQA errors arise from inadequately addressing three key data prep issues: *missing semantics*, *inconsistent values*, and *irrelevant columns*. To address the challenges, we introduce three types of data prep operations.

- **Derive:** A data prep task that derives a new column for table  $T$  from existing columns, aimed at addressing the challenge of *missing semantics*. This task typically involves operations such as combining columns through arithmetic computations, extracting relevant values, etc.
- **Normalize:** a data prep task that normalizes types or formats of the values in a column of  $T$  based on the needs of  $Q$ , aimed at addressing the challenge of *inconsistent values*. This task typically involves value representation or format normalization, type conversion, etc.
- **Filter:** A data prep task that filters out columns in  $T$  that are not relevant to answer question  $Q$ , aimed at addressing the challenge of *irrelevant columns*. This is crucial for handling large tables to address the input token limitations and challenges in long-context understanding of LLMs [34].

Given a high-level operation  $o_i$ , we define  $f_i$  as its low-level *implementation*, either by calling a well-established algorithm from a known Python library or using a customized Python program to meet the requirements of  $o_i$ .

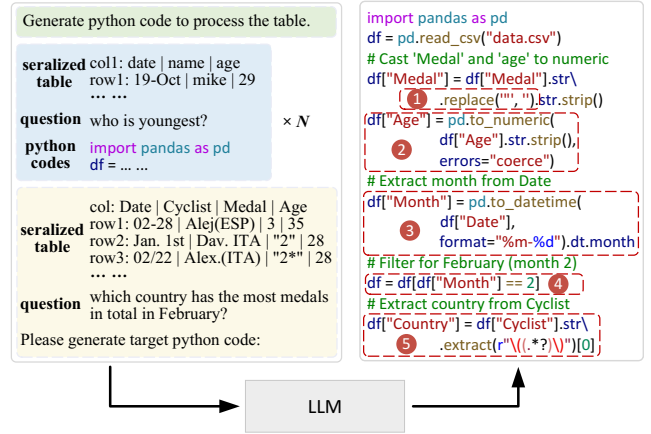
**EXAMPLE 2.** Figure 2 shows examples of question-aware data prep operations for TQA, along with their implementations in Python.

(a) **The Derive operation:** Figure 2a shows three examples of the Derive operations, i.e., extracting Country information from column Cyclist in  $T_1$ , computing new column GrowthRate using two columns in  $T_2$ , and inferring a status of IfLost by analyzing the scores in column Result.

(b) **The Normalize operation:** Figure 2b shows three examples of Normalize operations, i.e., normalizing the formats of Surface, Time and Date for tables  $T_4$ ,  $T_5$  and  $T_6$  respectively.

(c) **The Filter operation:** Figure 2c illustrates two examples of Filter operations in tables  $T_7$  and  $T_8$ . The Summary column in  $T_7$  contains an average of 100 words, increasing the difficulty for LLMs to identify the relevant AirDate column. Additionally, although  $T_8$  has 12 columns, only two are relevant to  $Q_8$ , and providing all columns may cause long-context understanding challenges of LLMs [34].

**Remarks.** While issues such as missing values and duplicates are common in real-world datasets, they are significantly less prevalent than the three highlighted challenges in existing TQA benchmarks. Thanks to AUTOPREP’s separation of logical operations and physical implementations, the set of data prep operations can be easily extended to accommodate new types of tasks. We leave the exploration of broader data prep challenges to future work.



**Figure 3: An LLM-based method with few-shot prompting for question-aware data prep.**

**Question-Aware Data Prep for TQA.** Given an NL question  $Q$  posed over table  $T$ , *question-aware data prep for TQA* is to generate a sequence of high-level operations  $O = \{o_1, o_2, \dots, o_{|O|}\}$  as a *logical plan*. Then, it generates a *physical plan*, where each operation  $o_i$  is implemented by low-level code  $f_i$ , such that these operations transform  $T$  into a new table  $T'$  that meets the needs of  $Q$ .

## 2.3 A Straightforward LLM-based Solution

A straightforward solution to question-aware data prep is to prompt an LLM to prepare tables, leveraging its ability to interpret the specific requirements of NL questions. For instance, consider the table in Figure 3 with columns Date, Cyclist, Medal and Age, and question: “Which country has the most medals in total in February”. A few-shot prompting strategy prompts an LLM with a task description and a few demonstrations, and requests the LLM to generate Python programs as shown in Figure 3.

However, this LLM-based solution may encounter the following limitations when performing question-aware data prep for TQA.

First, at the *logical-operation level*, given the inherent difficulties in understanding both NL questions and tables, it is challenging to accurately identify which data prep operations are specifically required to satisfy the needs of the NL question. This often leads to false negatives and false positives. For example, as shown in Figure 3, converting the Age column to a numerical format in code block ② is a false positive, as it is irrelevant to the question. In contrast, failing to normalize the Date column before extracting the month in code block ③ constitutes a false negative, as the method ignores the inconsistency in Date formats.

Second, at the *physical-operation level*, due to input token limitations and challenges in long-context understanding [34], it is not easy to fully understand all possible issues in a table, and thus may struggle to generate customized programs to correct issues. For example, in Figure 3, the normalization of Medal in code block ① overlooks certain corner cases (e.g., “2\*”), and the country extraction in code block ⑤ fails to handle “Dav.ITA”, which is formatted differently from other values.

Recent methods, such as CoTable [50] and ReAcTable [59], can improve few-shot prompting by employing techniques like Chain-of-Thoughts (CoT) and ReAct. However, these methods remain



insufficient to tackle the challenges, as they combine all diverse tasks, such as determining operations and implementing them, within a single LLM agent. Existing studies [30] have shown that a single LLM agent is often ineffective when tasked with handling a diverse range of operations, due to limited context length in LLMs and decreased inference performance with more input tokens.

### 3 AN OVERVIEW OF AUTOPREP

To address the limitations, we propose AUTOPREP, a *multi-agent* LLM framework that automatically prepares tables for given NL questions. Figure 4 provides an overview of our framework. Given an NL question  $Q$  posed over table  $T$ , AUTOPREP decomposes the data prep process into three stages:

- (1) **PLANNER Agent**: the **Planning** stage. It guides the LLM to suggest *logical* data prep operations  $O = \{o_1, o_2, \dots, o_{|O|}\}$ , which are tailored to specific question  $Q$ ,
- (2) **Multiple PROGRAMMER Agents** (e.g., **NORMALIZE**): the **Programming** stage. It directs the LLM to generate *physical* implementation  $f_i$  (e.g., Python code) for each operation  $o_i$  customized for the table  $T$ . Besides, it is also tasked for code debugging if any execution errors occur.
- (3) **An EXECUTOR Agent**: the **Executing** stage. It executes the generated code and reports errors if any bugs occur.

After that, an **ANALYZER** agent extracts the answer from the prepared table. This agent can either use LLMs as black-boxes or leverage them for code generation, which is *orthogonal* to the question-aware data prep problem studied in this paper. For simplicity, we use a Text-to-SQL strategy that translates the question into an SQL query over the prepared table to obtain the final answer, as shown in Figure 4. Note that other strategies could also be used by the agent in a “plug-and-play” manner.

**EXAMPLE 3.** Figure 4 illustrates how AUTOPREP supports data prep for an NL question posed over a table with 4 columns.

(a) The Planning stage: The **PLANNER** suggests the following high-level operations to address the specific NL question:

- $T'[\text{Country}] = \text{Derive}(\text{“Extract country code”}, \text{Cyclist})$  that extracts the country information from column *Cyclist*, producing a new *Country* column, in response to the “which country” part of the question.
- $\text{Normalize}(\text{“Case to INT”}, \text{Medal})$  that standardizes the value formats in the *Medal* column (e.g., removing quotation marks and asterisks) and then converts the strings to integers, as the question requires “the most medals”;
- $\text{Normalize}(\text{“Format date as \%m-\%d”}, \text{Date})$  standardizes the values in the *Date* column into a unified format to support the “in February” condition in the question.
- $T' = \text{Filter}([\text{Date}, \text{Country}, \text{Medal}], T)$  that filters out column *Age*, which is irrelevant to the question;

(b) The Programming stage: AUTOPREP designs specialized **PROGRAMMER** agents for each operation type, i.e., **DERIVE**, **NORMALIZE** and **FILTER**. Each specialized **PROGRAMMER** focuses on generating executable code for its assigned operations.

(c) The Executing stage: an **Executor** agent iteratively refines the generated code if any error occurs.

After these stages, AUTOPREP generates a prepared table  $T^*$ , which is then fed into an **ANALYZER** agent to produce the answer *ITA*.

**The PLANNER Agent.** The key challenge is how to suggest a *logical plan* that address specific NL questions. Even for the same table, different NL questions may require not only different logical operations but also varying sequences of those operations. To address this challenge, we propose a novel *Chain-of-Clauses (CoC) reasoning* method for the **PLANNER** agent. This method translates the NL question into an *Analysis Sketch*, outlining how the table should be transformed to produce the answer, thereby guiding the agent’s reasoning based on this sketch. More details of the method are given in Section 4.

**The PROGRAMMER Agents.** The key challenge is that a given logical operation can have multiple executable code alternatives (e.g., Python functions), and the difference in outcomes between the best and worst options can be substantial. For example, the **DERIVE** agent may generate an overly generic regular expression that extracts countries based on parentheses. Unfortunately, this code fails to correctly process “Dav. ITA”, which is formatted differently from other values. To tackle this challenge, we develop a *tool-augmented* approach that enhances the LLM’s code generation capabilities by utilizing predefined API functions. More details of our tool-augmented approach are discussed in Section 5.

**Remarks.** Our proposed AUTOPREP framework is extensible. When additional question-aware data prep operations are required, more specialized **PROGRAMMER** agents can be designed to handle them. The central **PLANNER** agent can then determine which operations should be performed and assign them accordingly.

## 4 THE PLANNER

### 4.1 A Direct Prompting Method

The most common way to generate a logical plan is to directly prompt an LLM using a typical in-context learning approach. The inputs are a question  $Q$ , a table  $T$ , a set  $\Sigma$  of specifications for each operation type, and an LLM  $\theta$ . Here, each specification  $\sigma \in \Sigma$  describes the purpose of an operation type, e.g., “an *Derive* operation creates a new column for a table based on existing columns, in response to the specific needs of a question”. The output of the algorithm is a set  $O$  of high-level operations

**EXAMPLE 4.** Figure 5(a) illustrates the direct prompting method, which produces two logical operations, *Filter* and *Normalize*. However, this logical plan might not be accurate, as discussed below.

(a) Incorrect operations: The *Filter* operation retains the *Country* column simply because the question mentions “which country.” However, it fails to recognize that the original table does not actually contain a *Country* column. Worse yet, it incorrectly filters out the *Cyclist* column, merely because *Cyclist* is not explicitly mentioned in the question. This mistake is critical, as the country information is implicitly embedded within the *Cyclist* values.

(b) Missing operations: Observing the ground-truth in Figure 4, we can see that the *Derive* operation on *Cyclist* is not generated, as the column has already been filtered out. Moreover, although the *Normalize* operation on *Medal* is generated, the *Normalize* operation on *Date* is missing. This is because the phrase “the most” in the

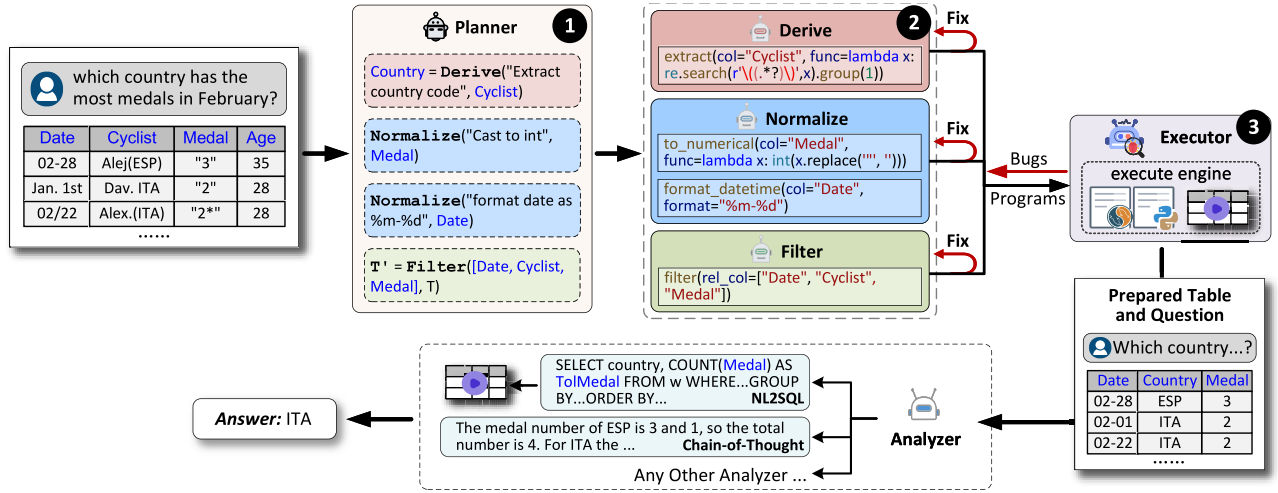


Figure 4: A Multi-agent Data Prep Framework for TQA (Top), which could be plugged to any other TQA methods (Bottom).

question suggests the need for type conversion of Medal, but there are no clues from the question to normalize Date, unless we observe from the table that the values are inconsistent.

The above example clearly demonstrates that the key challenge in designing the PLANNER agent, as discussed earlier, is to capture the relationships between different parts of question  $Q$  and the columns in table  $T$ . For instance, as shown in Figure 4, the Normalize operation on the Date column is generated by considering two key factors: (1) the question contains "in February", and (2) the values in the Date column are inconsistent.

## 4.2 A Chain-of-Clauses Method

To address this challenge, we propose a *Chain-of-Clauses (CoC)* reasoning method that decomposes the entire process of logical plan generation into two phases, as shown in Figure 5(b).

- The first phase leverages an LLM to generate an SQL-like *Analysis Sketch*, outlining how the table should be transformed to produce the answer.
- The second phase iteratively examines different clauses in the Analysis Sketch, e.g., ORDER BY Date. At first, we use the syntax parser from Binder [17] to extract all UDFs and operation clauses in the sketch and sort them by the execution order. Then, for each clause, we associate the corresponding data in  $T$  (e.g., values in column Date) with it and provide several in-context learning demos to prompt the LLM for generating possible logical operations (e.g., Normalize).

Compared with existing CoT methods [51], which simply break down questions into sub-questions, our approach is more effective for logical plan generation. First, our method decomposes the question into a set of analysis steps over table, formalized as an SQL-like Analysis Sketch, which simplifies the task of logical plan generation. More importantly, our method *jointly* considers each analysis step along with the corresponding relevant columns (instead of the whole table) to prompt the LLM, effectively capturing the relationships between the question and the data.

**SQL-like Analysis Sketch.** An Analysis Sketch  $S$  is an SQL-like statement, which can be represented as follows.

```
SELECT A | agg(A) | f(A*, {A_i}) FROM T
WHERE Pred(A_i) AND ... AND Pred(A_j)
GROUP BY A ORDER BY A LIMIT n
```

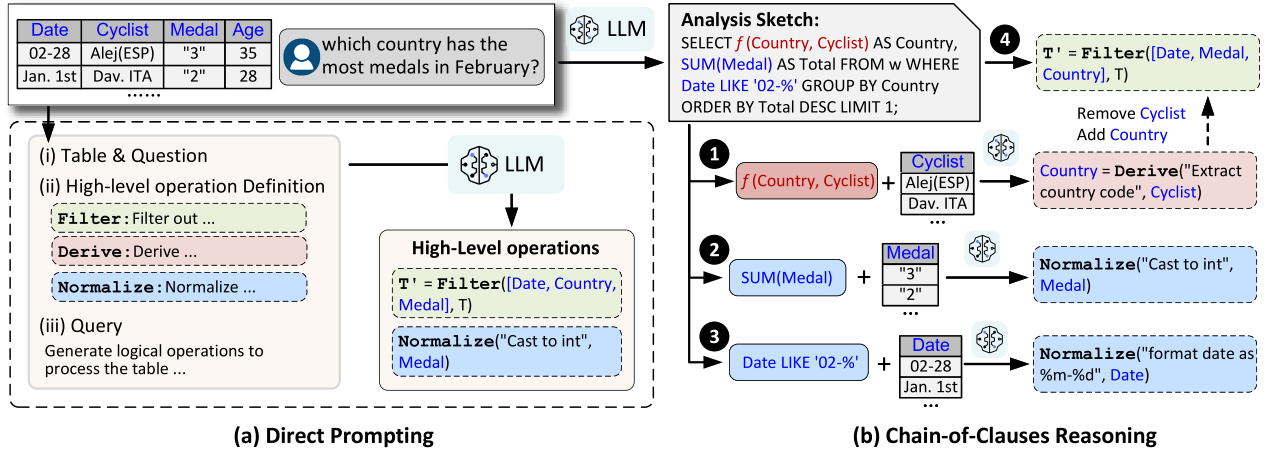
where  $agg(A)$  is an aggregation function (e.g., SUM and AVG) over column  $A$ ,  $Pred(A_i)$  denotes a predicate (i.e., filtering condition) over column  $A_i$ , such as Date LIKE '02-%'. Note that  $f(A*, \{A_i\})$  is a *user-defined function* (UDF) that maps existing columns  $\{A_i\}$  in the table into a new column  $A^*$ . Figure 5(b) shows an example Analysis Sketch with a UDF that specifies a new column Country, i.e.,  $f(\text{Country}, \text{Cyclist})$ . Obviously, this UDF is introduced because the Analysis Sketch contains a GROUP BY Country clause.

**Logical Operation Generation.** Given the constructed Analysis Sketch, we orchestrate the logical operations and determine their execution order by translating the sketch into actionable steps. Specifically, we leverage the syntax parser from Binder [17] to extract all UDFs and operation clauses from the sketch and sort them according to their intended execution sequence from the parser. Based on this parsed order, we then generate the corresponding logical operations, as illustrated below.

**EXAMPLE 5.** Figure 5(b) shows our proposed CoC reasoning method for the example question and table in Figure 4. Specifically, the method generate a set  $O$  of operations via the following two phases.

(a) **Phase I - Analysis Sketch Generation:** In this phase, the algorithm prompts the LLM  $\theta$  with several exemplars  $\{(Q_i, T_i, s_i)\}$  to generate an Analysis Sketch  $S$ , as shown in Figure 5(b).

(b) **Phase II - Operation Generation:** In this phase, the algorithm iteratively examines the clauses in Analysis Sketch  $S$  as follows. (1)  $f(\text{Country}, \text{Cyclist})$ : this clause and relevant columns are used to prompt the LLM to generate an Derive operation. (2) SUM(Medal): this clause and the values in column Medal are used to prompt the LLM to generate a Normalize operation. (3) Date LIKE '02-%': this clause and the values in column Date are used to prompt the LLM to generate a Normalize. Finally, the algorithm generates a Filter



**Figure 5: Our proposed approaches to logical plan generation in the PLANNER agent: (a) A straightforward direct prompting method, and (b) a more effective Chain-of-Clauses (CoC) reasoning method.**

operation that only selects columns relevant to the Analysis Sketch. Note that the Filter operation removes irrelevant columns, whereas the Derive operation creates new columns from existing ones.

## 5 THE PROGRAMMER & EXECUTOR

PROGRAMMER agents translate a high-level logical plan into a *physical plan* by generating low-level code, which is then passed to an EXECUTOR agent for code execution and interactive debugging. A straightforward approach is to prompt an LLM with the logical plan using in-context learning with several exemplars, asking the LLM to generate code for each high-level operation in the plan. The code is then executed iteratively, and if any runtime errors occur, the PROGRAMMER is prompted with the error messages for debugging.

However, the above method may have limitations, as the generated code may be overly generic and unable to effectively address the *heterogeneity* challenge of the tables. Specifically, many tables originate from web sources or real-world scenarios, where values have diverse syntactic (e.g., “19-Oct” and “9/14”) or semantic formats (e.g., “ITA” and “Italia”), making it difficult to generate code customized to these tables. For example, given a Normalize operation to standardize country formats, a generic function from an existing Python library may not be sufficient to transform country names (e.g., “Italia”) to their ISO codes (e.g., “ITA”). In such cases, customized functions, such as the `clean_country` function from an external library `dataprep` [41], are needed to address specific requirements of code generation.

To address these limitations, we introduce a *tool-augmented* method for the PROGRAMMER agents, enhancing the LLM’s code generation capabilities by utilizing pre-defined API functions, referred to as the *function pool* in this paper.

### 5.1 Tool-Augmented Method: Key Idea

Figure 6 provides an overview of our *tool-augmented* method for physical plan generation and execution. Given a table  $T$  and a set  $O = \{o_1, \dots, o_{|O|}\}$  of logical operations, the method *iteratively* generates and executes physical executable code for each individual operation  $o_i \in O$ , thus generating a sequence of intermediate tables

$T_1$  (i.e., the input  $T$ ),  $T_2, \dots, T_{|O|+1}$ . Specifically, in the  $i$ -th iteration, given a high-level operation  $o_i$  (e.g., Derive) and an intermediate table  $T_i$ , the method generates physical code and executes it to produce an updated table  $T_{i+1}$  through the following two steps.

**Function Selection and Argument Inference.** The method first prompts the LLM to select a specific function from a function pool  $\mathcal{F}$  corresponding to the operation type (e.g.,  $F_{\text{Der}}$ ) and infer the arguments (e.g., regular expression) for the selected function.

For instance, given the Derive operation over table  $T_i$  shown in Figure 6, the LLM selects a function from the pool  $\mathcal{F}_{\text{Der}}$  designed specifically for Derive, obtaining an `extract` function with two arguments: column and func. The LLM then generates preliminary code for these arguments, assigning `Cyclist` to the column argument and generating a lambda function with a specific regular expression for argument `func`.

Note that, in addition to selecting functions from the corresponding pool, our method may also prompt the LLM to write specific code for an operation (denoted as `CustomCode`) if no existing function is suitable to meet the operation’s requirements.

**Code Execution and Debugging.** Given the function generated in the previous step, the EXECUTOR agent then applies the function over table  $T_i$ . If any bugs occur during execution, it captures and summarizes error messages and returns to corresponding PROGRAMMER agent. For instance, as shown in Figure 6, the first lambda function for the argument `func` only extracts countries based on parentheses, which may produce incorrect results for the value “Dav.ITA”, as it is formatted differently from other values. In this case, based on the execution results, the EXECUTOR agent records the error log and examines relevant data to summarize reasons, which will be passed to the corresponding PROGRAMMER to modify the code. After the execution and debugging process, the method produces a new intermediate table  $T_{i+1}$  and proceeds to the next operation,  $o_{i+1}$ , i.e., a Normalize operation in Figure 6.

### 5.2 Search Space of Function Pools

This section presents a search space of function pools for different types of high-level operations formalized in the current AUTOPREP.

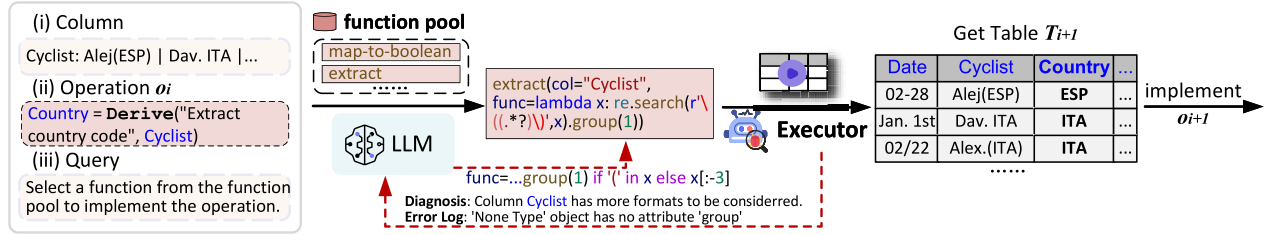


Figure 6: Our proposed tool-augmented method for physical plan generation and execution.

**Function Pool for Derive.** We define a specific PROGRAMMER agent, *i.e.*, DERIVE, to generate low-level code for the operation type Derive from the following function pool  $\mathcal{F}_{\text{Der}}$ .

(1) `extract`. This function extracts a substring from a column to generate a new column. It has two arguments: `column`, representing the name of the source column, and `func`, representing the substring extraction function. Using table  $T_1$  from Figure 2a as an example, to extract the country code, we would set `column` to `Cyclist` and `func` to a lambda function `lambda x: re.search(r'((.*?)\\', x).group(1))`.

(2) `calculate`. The function generates a new column through arithmetic operations of existing columns. It has two arguments: `columns`, representing a list of source columns and `func`, representing a lambda function with input to be a dictionary of values, since we may perform calculations on multiple columns. For example, to generate a new column `GrowthRate` for  $T_2$  in Figure 2a, we set `columns` to `['2012', '2013']` and `func` to a lambda function `lambda x: (x['2013'] - x['2012']) / x['2012']`.

(3) `map-to-boolean` and `concatenate`. Both functions generate a new column from multiple columns. The difference is that one generates boolean values, while the other concatenates strings. Both operators take two arguments, `columns` and `func`, similar to those in the `calculate` operator.

(4) `infer`. This function uses LLMs to deduce values that could not be processed by the above functions. It takes a list of columns and the failed target values as input, learning from demonstrations of successfully processed values to output the target values.

**Function Pool for Normalize.** We define a specific PROGRAMMER agent, *i.e.*, NORMALIZE, to generate low-level code for the operation type Normalize from the following function pool  $\mathcal{F}_{\text{Norm}}$ .

(1) `to-numerical`. This function standardizes a column to a numerical type, such as `int` or `float`. To use it, two arguments need to be specified: `column`, representing the source column, and `func`, representing a lambda function. For example, we can use this function to standardize the `Time` column to integers, with `column` as `Time` and `func` as the lambda function.

(2) `format-datetime`. The function standardizes the format of columns with `DateTime` values. It has two arguments: `column` and `format`. The `format` argument specifies the desired format for standardization. For example, to make the values in the `Date` column comparable, we use the `format-datetime` function with `column` set to `Date` and `format` set to `"%Y-%m-%d"`.

(3) `clean-string`. To clean string values in tables, we define function `clean-string` with two arguments: `column` and `trans_dict`. The `trans_dict` is a dictionary where each key  $k$  and its corresponding value  $v$  represent that  $k$  in `column` is replaced by  $v$ .

(4) `infer`. This function uses LLMs to deduce values that could not be processed by the functions above.

**Function Pool for Filter.** We design the FILTER agent to utilize a function pool  $\mathcal{F}_{\text{Filter}}$  with a single function, `filter-columns`. This function has one argument, `rel_columns`, which represents a list of question-related column names.

**Remarks.** The design of `CustomCode` and `infer` operations enables AUTOPREP to generalize to unseen tasks. For example, when encountering missing or clearly erroneous values, AUTOPREP can directly prompt an LLM to infer plausible replacements. Moreover, our search space is extensible: additional functions or external APIs can be easily integrated into the function pool, including specialized data preparation functions for various semantic types, such as those provided by the `dataprep` library [41].

## 6 EXPERIMENTS

### 6.1 Experimental Setup

**Datasets.** We evaluate AUTOPREP using the following well-adopted datasets. The statistics of the datasets are shown in Table 1.

(1) **WikiTQ** [40] contains complex questions annotated by crowd workers based on diverse Wikipedia tables. WikiTQ comprises 17,689 question-answer pairs in the training set and 4,344 in the test set with significant variation in table size.

(2) **TabFact** [15] provides a collection of Wikipedia tables, along with manually annotated NL statements. One requires to deduce the relations between statements and tables as “true” (a statement is entailed by a table) or “false” (a statement if refuted by a table). To reduce the experimental cost without losing generality, we choose the small test set provided by [57] which contains 2024 table-statement pairs. The tables in TabFact are much smaller than those in WikiTQ. The answers are binary with nearly equal proportions.

(3) **TabBench** [52] provides an evaluation of TQA capabilities within four major categories, including multi-hop fact checking (FC), multi-hop Numerical Reasoning (NR), Trend Forecasting and Chart Generation. We evaluate the generalization of AUTOPREP on FC and NR, where one requires to conduct more complex and multi-hop reasoning over tables for answering a question. Specifically, we directly use the prompts in WikiTQ to evaluate the generalization capabilities of AUTOPREP on TabBench.

**Baselines.** We consider the following three categories of baselines:

(1) **TQA Methods w/o Data Prep** directly generate the answer or utilize programs to extract the answer from the table without considering data prep. We implement four primary TQA baselines.



**Table 1: Statistics of Datasets.**

| Dataset    | # Rec. | # Row. | # Col. | Ans. Types             |
|------------|--------|--------|--------|------------------------|
| WikiTQ     | 22,033 | 4~753  | 3~25   | string / list (3.05%)  |
| TabFact    | 2,024  | 5~47   | 5~14   | true / false (49.60%)  |
| TableBench | 886    | 2~212  | 2~20   | string / list (31.49%) |

**End2EndQA (End2End)** [14] utilizes the in-context learning abilities of LLMs to generate the answer for TQA task based on the supervision of human-designed demonstrations. We implement End2End method with prompt and demonstrations provided by [17].

**Chain-of-Thought (CoT)** [51] prompts LLMs to generate the reasoning process step-by-step before generating the final answer. We implement CoT with the prompt provided by [14].

**NL2SQL** [44] first translates the question into an SQL program and then executes it to get the final answer from the table for the question. We use the prompt from [17] to implement NL2SQL.

**NL2Py** uses Python code to process and reason over the tables. To construct the prompt for NL2Py, we use TQA instances in NL2SQL prompt and manually write the Python code to process the table and generate the final answer.

(2) **Data Prep Baselines.** We consider the following data prep baselines. For each baseline, we first perform data prep and then use the above NL2SQL to extract answers from the prepared tables.

**Offline DataPrep (Off-Prep)** performs offline data prep operations for all tables in TQA. To this end, we have surveyed and consolidated the offline data prep operations, such as data cleaning, value normalization and column renaming, adopted by current SOTA TQA methods [17, 32, 50, 55, 60] to construct a comprehensive offline data prep pipeline, by utilizing the Pandas library [36] and the popular DataPrep toolkit [42].

**ICL-Prep** uses few-shot ICL demonstrations to guide LLMs in generating Python programs for data prep, as shown in Figure 3.

(3) **SOTA TQA Methods with Data Prep.** We investigate four SOTA TQA methods considering data prep tasks for TQA task.

**Dater** [57] addresses the TQA task by decomposing the table and question. It first selects relevant columns and rows to obtain a sub-table and then decomposes the origin question into sub-questions. Dater answers these sub-questions based on the sub-tables to generate the final answer. We use code in [2] for implementation.

**Binder** [17] enhances the NL2SQL method by integrating LLMs into SQL programs. It uses LLMs to incorporate external knowledge bases and directly answer questions that are difficult to resolve using SQL alone. We utilize the original code provided by [1].

**AutoTQA** [60] uses a multi-agent framework for TQA. Since the official code for AutoTQA is not publicly available, we reimplement it under the guidance of the authors (see our repository (<https://github.com/fmh1art/AutoPrep/src/model/autotqa>)). Given the high time and API costs, we evaluate AutoTQA on a sampled subset of 500 instances from each of WikiTQ and TabFact.

**ReAcTable** [59] uses the ReAct paradigm to extract relevant data from the table using Python or SQL code generated by LLMs. Once all relevant data is gathered, it asks the LLMs to predict the answer. We run the original code from [4] and keep all settings as default.

Notice that the original code does not include prompts for TabFact, we generate it based the WikiTQ prompt.

**Chain-of-Table (CoTable)** [50] enhances the table reasoning capabilities of LLMs by predefining several common atomic operations (including data prep operations) that can be dynamically selected by the LLM. These operations form an “operation chain” that represents the reasoning process over a table and can be executed either via Python code or by prompting the LLM. We implement CoTable using the original code from [3].

**Evaluation Metrics.** We consider both accuracy and cost.

**Accuracy.** We adopt the evaluator from Binder [17] to address cases where program executions are semantically correct but do not exactly match the golden answers.

**Cost.** We measure both the time and API cost for all methods. For time cost, we ensure a stable network environment and record the end-to-end processing time for each method on a single TQA instance. For API cost, we follow the official pricing guidelines ([https://api-docs.deepseek.com/quick\\_start/pricing/](https://api-docs.deepseek.com/quick_start/pricing/)) and calculate the cost based on the default LLM backbones used.

**Backbone LLMs.** We evaluate our methods using representative LLMs as backbones. For closed-source LLMs, we select DeepSeek [24] (DeepSeek-V2.5-Chat) and GPT3.5 [9] (GPT3.5-Turbo-0613). For open-source LLMs, we choose Llama3 [18] (Llama-3.1-70B-Instruct) and QWen2.5 [53] (QWen2.5-72B-Instruct) for evaluation.

**Experiment Settings.** We provide detailed prompts of each component in AUTOPREP in our technical report [5] due to the space limit. Moreover, for fair comparison, we set the maximum token input of all methods as 8192. Moreover, we set the temperature parameter of all methods to 0.01 for reproducibility.

## 6.2 Improvement of Data Prep for TQA

**Exp-1: Impact of question-aware data prep on TQA performance.** We integrate AUTOPREP into our four TQA baselines w/o data prep, and report the results in Table 2.

As demonstrated, integrating AUTOPREP significantly improves the performance of all evaluated methods. Notably, NL2SQL shows the most substantial gains, achieving an average accuracy improvement of **12.22** on WikiTQ and **13.23** on TabFact across all LLM backbones. Similarly, NL2Py also shows notable improvements in its performance, after being integrated with AUTOPREP. This significant improvement is attributed to the sensitivity of NL2SQL and NL2Py to data incompleteness and inconsistency, which can lead to erroneous outcomes when performing operations on improperly formatted data. Thus, data prep operations, such as Derive and Normalize can solve these cases and improve the overall results.

Moreover, End2End and CoT methods also show considerable performance gains. These improvements are largely due to the filtering mechanism of AUTOPREP, which removes irrelevant columns, thereby simplifying the reasoning process for extracting answers from tables. Since “NL2SQL + AUTOPREP” achieves the best accuracy in most cases, we take its results as default for further comparison.

## 6.3 Data Prep Method Comparison

**Exp-2: Comparison of AUTOPREP with data prep baselines.** We compare AUTOPREP against two representative baselines: a

**Table 2: Improvement of data prep for TQA (the best results are in bold and the second-best are underlined).**

| Method     | DeepSeek                      |                               | GPT3.5                        |                               | Llama3                        |                               | QWen2.5                       |                               |
|------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|-------------------------------|
|            | WikiTQ                        | TabFact                       | WikiTQ                        | TabFact                       | WikiTQ                        | TabFact                       | WikiTQ                        | TabFact                       |
| End2End    | 56.65                         | 81.77                         | 52.56                         | 71.54                         | 58.72                         | 81.27                         | 60.01                         | 81.17                         |
| + AUTOPREP | 63.14 $\uparrow$ 6.49         | 82.11 $\uparrow$ 0.34         | 61.21 $\uparrow$ 8.65         | 71.79 $\uparrow$ 0.25         | 61.23 $\uparrow$ 2.51         | 84.19 $\uparrow$ 2.92         | 63.42 $\uparrow$ 3.41         | 83.05 $\uparrow$ 1.88         |
| CoT        | 54.95                         | 82.02                         | 53.48                         | 65.37                         | 40.75                         | 80.93                         | 59.67                         | 82.31                         |
| + AUTOPREP | 61.12 $\uparrow$ 6.17         | 82.26 $\uparrow$ 0.24         | 60.01 $\uparrow$ 6.53         | 74.36 $\uparrow$ 8.99         | 56.01 $\uparrow$ 15.26        | 83.65 $\uparrow$ 2.72         | 62.02 $\uparrow$ 2.35         | <u>85.67</u> $\uparrow$ 3.36  |
| NL2Py      | 59.35                         | 68.13                         | 53.59                         | 66.15                         | 50.12                         | 76.24                         | 53.02                         | 72.63                         |
| + AUTOPREP | <u>65.86</u> $\uparrow$ 6.51  | <u>87.35</u> $\uparrow$ 19.22 | <u>64.69</u> $\uparrow$ 11.1  | <u>84.83</u> $\uparrow$ 18.68 | <u>62.55</u> $\uparrow$ 12.43 | <u>85.42</u> $\uparrow$ 9.18  | <u>68.65</u> $\uparrow$ 15.63 | <b>85.72</b> $\uparrow$ 13.09 |
| NL2SQL     | 52.83                         | 70.21                         | 52.90                         | 64.71                         | 51.80                         | 75.15                         | 56.86                         | 80.09                         |
| + AUTOPREP | <b>66.09</b> $\uparrow$ 13.26 | <b>87.85</b> $\uparrow$ 17.64 | <b>64.75</b> $\uparrow$ 11.85 | <b>84.19</b> $\uparrow$ 19.48 | <b>63.72</b> $\uparrow$ 11.92 | <b>85.72</b> $\uparrow$ 10.57 | <b>68.72</b> $\uparrow$ 11.86 | 85.33 $\uparrow$ 5.24         |

**Table 3: Experimental results of AUTOPREP and TQA methods with data prep.**

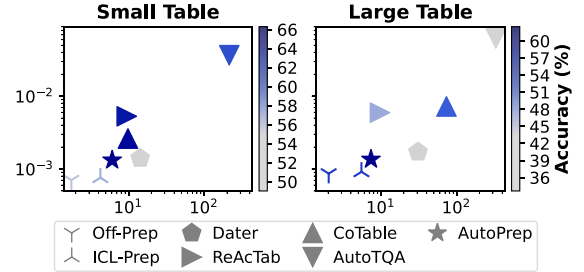
| Method    | DeepSeek     |              | GPT3.5       |              | Llama3       |              | QWen2.5      |              |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|           | WikiTQ       | TabFact      | WikiTQ       | TabFact      | WikiTQ       | TabFact      | WikiTQ       | TabFact      |
| Off-Prep  | 55.32        | 81.67        | 56.86        | 81.52        | 53.02        | 75.40        | 58.01        | 82.31        |
| ICL-Prep  | 56.54        | 80.53        | 55.71        | 73.91        | 50.05        | 75.20        | 57.00        | 80.14        |
| Dater     | 48.32        | 83.05        | 52.81        | 72.08        | 43.53        | 74.01        | 58.78        | 79.84        |
| Binder    | 56.81        | 82.81        | 56.74        | 79.17        | 50.51        | 78.16        | 55.43        | 81.72        |
| AutoTQA*  | 60.80        | 84.40        | 58.40        | 80.60        | 58.80        | 82.40        | 62.40        | 83.00        |
| ReAcTable | 64.13        | 85.71        | 51.80        | 72.80        | 58.01        | 80.00        | 60.15        | 81.67        |
| CoTable   | <u>64.53</u> | <u>86.22</u> | <u>59.94</u> | 80.20        | <u>62.22</u> | <u>85.62</u> | <u>64.41</u> | <u>83.20</u> |
| AUTOPREP  | <b>66.09</b> | <b>87.85</b> | <b>64.75</b> | <b>84.19</b> | <b>63.72</b> | <b>85.72</b> | <b>68.72</b> | <b>85.33</b> |

traditional offline data prep method Off-Prep and an LLM-based in-context learning data prep method ICL-Prep.

As shown in Table 3, AUTOPREP consistently outperforms both baselines on WikiTQ and TabFact. Specifically, AUTOPREP improves over Off-Prep by **10.02** and **5.55** on average, respectively. The performance gap can be attributed to the fact that Off-Prep lacks question guidance, making it ineffective in addressing question-specific issues such as missing semantics and irrelevant columns. Moreover, predefined normalization routines struggle with table heterogeneity, whereas AUTOPREP can generate specialized code for such cases via LLMs. Similarly, AUTOPREP outperforms ICL-Prep by **11.00** and **8.33** on average on the two datasets. This highlights that a multi-agent framework is essential for producing comprehensive data prep requirements and precise, executable programs.

**Exp-3: Comparison of AUTOPREP with previous SOTA TQA methods with data prep.** We compare AUTOPREP with TQA methods that integrate data prep tasks in their question-answering process. The results, shown in Table 3, highlight that AUTOPREP sets a new SOTA performance on both the WikiTQ and TabFact datasets across LLM backbones. Although CoTable achieves the best overall performance among existing SOTA methods, AUTOPREP outperforms it with an impressive average improvement of **3.05** on WikiTQ and **1.96** on TabFact. These improvements are largely attributed to our multi-agent framework, which effectively addresses the question-aware data prep challenges. Furthermore, when compared to AutoTQA, AUTOPREP shows significant gains of **5.81** on WikiTQ and **2.77** on TabFact. This is because AutoTQA lacks comprehensive data prep, leading to execution errors or incorrect answers.

The results demonstrate that data prep is inherently complex and cannot be solved with a one-size-fits-all solution. Instead, a more effective strategy involves specialized LLM-based agents for



**Figure 7: Efficiency evaluation.** The  $x$ -axis and  $y$ -axis represent the time cost (in seconds) and monetary cost (in dollars) for processing a single instance, respectively. The color intensity of each scatter point reflects the overall accuracy.

each type of data prep task, coordinated by a centralized planning agent. Moreover, AUTOPREP covers a broader range of data prep operations, filling gaps (e.g., Derive and Normalize) that previous solutions have not fully addressed.

Moreover, we also evaluate AUTOPREP on tables with various sizes. We find that, by employing multiple agents and program-based operations, AUTOPREP maintains stable performance as table size grows, ensuring that each data prep task is handled effectively without overwhelming a single model. More details on the experimental results and analysis can be found in our technical report [5].

As DeepSeek achieves the best accuracy at lower cost, we select it as our default backbone LLM for subsequent experiments.

## 6.4 Evaluation on Efficiency

**Exp-4: How efficient is AUTOPREP compared with other methods?** To evaluate the efficiency of AUTOPREP, we compare it with

Table 4: Evaluating Extensibility for New Operation Types.

| Method    | TransTQ      | WikiTQ       |              |
|-----------|--------------|--------------|--------------|
|           |              | origin       | after        |
| ICL-Prep  | 58.92        | 56.54        | 55.00        |
| ReAcTable | 51.04        | <u>64.13</u> | <u>63.03</u> |
| AutoTQA   | <u>59.75</u> | 60.80        | 61.20        |
| AUTOPREP  | <b>68.88</b> | <b>66.09</b> | <b>66.28</b> |

other TQA methods in terms of time cost, monetary cost, and accuracy. The results for both small and large tables are shown in Figure 7. As shown, AUTOPREP achieves the best overall performance with insignificant time and monetary costs. Specifically, its cost is lower than all methods except Off-Prep and ICL-Prep. However, AUTOPREP significantly outperforms these two baselines in accuracy, achieving improvements of at least 9.36 and 5.77 points on small and large tables, respectively. Furthermore, as table size increases, AUTOPREP maintains a much more stable time and cost overhead compared to other state-of-the-art TQA methods.

This efficiency stems from the program-based data prep mechanism of AUTOPREP, which avoids high latency and API expenses associated with directly processing tables via LLMs. For instance, in the column derivation task on Table  $T_1$  (Figure 2a), CoTable invokes LLMs to generate a full list of country codes. In contrast, AUTOPREP only generates a physical operator extract and executes it to produce the column, which results in significantly lower time and monetary costs, particularly on large tables.

## 6.5 In-Depth Exploration of System Capabilities

**Exp-5: Evaluating generalization on unseen datasets.** As discussed in Section 6.1, we evaluate the generalization capabilities of AUTOPREP on the TabBench datasets. Specifically, we directly use the designed prompting strategies on the WikiTQ dataset, and examine whether these strategies can be generalized to TabBench.

As shown in Figure 8, AUTOPREP achieves the highest overall accuracy among all methods. Specifically, compared with the second best method ReAcTable, AUTOPREP improves by 5.28, indicating the strong generalization capabilities of our method. The main reason is that AUTOPREP utilizes tool-augmented method for physical plan generation and execution, generalizing well on unseen datasets.

**Exp-6: Evaluating extensibility on datasets requiring new logical operation types.** To further evaluate the extensibility of AUTOPREP, we evaluate it on a more complex dataset involving logical operations not originally supported. Since no existing TQA dataset presents sufficiently complex data quality issues, we construct a new dataset TransTQ with the assistance of LLMs, which is available on our GitHub repository. Specifically, we prompt an LLM to identify tables from WikiTQ where inverse transformation operations (e.g., pivot and stack [33]) can be applied. These operations are then executed to generate non-relational table formats, introducing new challenges for TQA on table transformation.

As shown in Table 4, without explicit guidance for handling table transformations, all baseline TQA methods exhibit poor accuracy. In contrast, after incorporating a new logical operation Transform, AUTOPREP achieves the best performance, outperforming the second-best method by 9.31 points. Moreover, we also

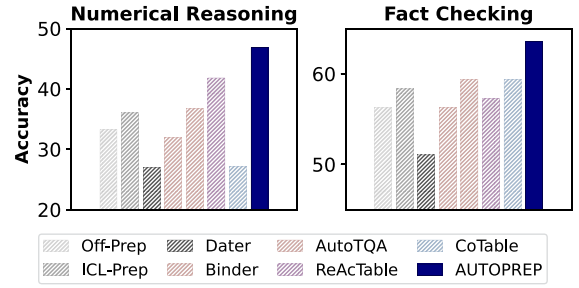


Figure 8: Evaluating generalization on the TabBench dataset.

validate that adding this new operation does not degrade the performance of AUTOPREP on the original WikiTQ dataset, confirming that its multi-agent architecture is well extensible to support the seamless integration of new data prep functionalities.

## 6.6 Ablation Studies

**Exp-7: Evaluation on the Planner agent.** We compare two Planner variants with different high-level operation suggestion methods, namely Direct Prompting and our proposed Chain-of-Clauses method, and report the results in Table 5a.

We observe that Chain-of-Clauses outperforms Direct Prompting by 7.92, 4.50, 5.27 in accuracy on WikiTQ, TabFact and TabBench respectively. This indicates the superiority of our proposed method in generating more accurate logical operations. Moreover, we find that the performance improvement on WikiTQ is more significant than that on other datasets. This is because the WikiTQ dataset has relatively large tables and complex questions, which could make the logical operation suggestion problem more challenging to be solved using Direct Prompting.

**Exp-8: Evaluation on the Programmer Agents.** We evaluate the low-level operation generation methods in the PROGRAMMER agents and keep other settings of AUTOPREP as default.

As illustrated in Table 5b, our proposed tool-augmented method achieves better performance compared with the code generation method. Considering the logical operations input to the Programmer agents are the same, we can conclude that selecting a function from a function pool and then completing its arguments can generate more accurate and high-quality programs to implement the high-level operations. Moreover, when using demonstrations constructed from same table-question pairs for the prompt of programmer agents, Tool-augmented method can save 18.07% input tokens for low-level operation generation. We also record the error ratio of these two methods, as shown in Table 5b. The probability of bugs in our tool-augmented method is greatly reduced (e.g., from 4.51% to 1.50%), demonstrating its effectiveness.

**Exp-9: Contribution of each agent in AUTOPREP.** We ablate each PROGRAMMER agent including FILTER, DERIVE and NORMALIZE and compare the performance with AUTOPREP.

As shown in Table 5c, for WikiTQ, without column derivation, the accuracy drops the most (4.30). For TabFact and TabBench, the Normalize matters the most with an accuracy drop by 4.06 and 8.52. This is because that WikiTQ has more instances requiring string extraction or calculation to generate new columns for answering the question, while normalization is a primary issue in TabFact

**Table 5: Experimental Results of Ablation Studies.****(a) Evaluation on the PLANNER agent.**

| Method           | WikiTQ       | TabFact      | TabBench     |
|------------------|--------------|--------------|--------------|
| Direct Prompting | 58.17        | 83.35        | 44.83        |
| Chain-of-Clauses | <b>66.09</b> | <b>87.85</b> | <b>50.10</b> |

**(b) Evaluation on the PROGRAMMER Agents.**

| Method     | Metric           | WikiTQ       | TabFact      | TabBench     |
|------------|------------------|--------------|--------------|--------------|
| Code       | Acc $\uparrow$   | 62.82        | 81.97        | 47.26        |
| Generation | Err $\downarrow$ | 4.51%        | 4.50%        | 4.73%        |
| Tool       | Acc $\uparrow$   | <b>66.09</b> | <b>87.85</b> | <b>50.10</b> |
| Augmented  | Err $\downarrow$ | <b>1.50%</b> | <b>0.05%</b> | <b>1.01%</b> |

**(c) Contribution of Each PROGRAMMER Agent in AUTOPREP.**

| Method      | WikiTQ        | TabFact       | TabBench      |
|-------------|---------------|---------------|---------------|
| AUTOPREP    | <b>66.09</b>  | <b>87.85</b>  | <b>50.10</b>  |
| - FILTER    | 62.78 (-3.31) | 84.98 (-2.87) | 47.67 (-2.43) |
| - DERIVE    | 61.79 (-4.30) | 85.67 (-2.18) | 45.44 (-4.66) |
| - NORMALIZE | 62.02 (-4.07) | 83.79 (-4.06) | 41.58 (-8.52) |

and TabBench. Moreover, for all datasets, each agent plays an essential role in data preparation for TQA, which brings accuracy improvement by at least **2.43**, **2.18** and **4.06**.

## 7 RELATED WORK

**Tabular Question Answering.** Most of the SOTA solutions for TQA rely on LLMs [9, 24], as TQA requires NL understanding and reasoning over tables. There are two types of methods for TQA named *Direct Prompting* [14, 51] and *Code Generation* [44]. Previous TQA methods, like Dater [57], Binder [17], CoTable [50] and ReAcTable [59] also consider data prep in their question answering process. Specifically, Dater [57] prompts LLMs to select relevant columns related to answering the question, targeted at solving filtering tasks. Binder [17] proposes to integrate SQL with an LLM-based API to incorporate external knowledge, which may partly address derivation tasks. Similarly, CoTable [50] addresses the filtering tasks and derivation tasks by designing operators which are implemented by LLM completion. ReAcTable [59] uses few-shot demonstrations to instruct LLMs to generate python code or SQL to address the derivation and filtering tasks.

However, previous TQA methods do not address the column normalization tasks, which account for the most significant error types, as indicated in Figure 1. Second, the methods utilize a single LLM agent for both data prep and answer reasoning. Given that data prep is a complex challenge, these one-size-fits-all solutions may not achieve satisfactory performance.

**Traditional Data Preparation.** Data prep techniques are widely used across various tasks [10, 20, 49]. For training machine learning models for data analytics, Auto-Weka [48] leverages Bayesian optimization to identify data prep operations. Auto-Sklearn [21, 29] and TensorOBOE [54] apply meta-learning to discover promising operations. Alpine Meadow [45] introduces an exploration-exploitation strategy, while TPOT [38] uses a tree-based representation of data prep and genetic programming optimization techniques. Several studies [8, 19, 27, 58] explore reinforcement learning techniques.

HAIPipe [13] integrates both human-orchestrated and automatically generated data prep operations.

We propose AUTOPREP to generate *question-aware* data prep operations for TQA tasks. Traditional data prep methods typically operate at an *offline stage*, independent of any specific downstream question or query. In contrast, question-aware data prep studied in AUTOPREP tailors the tables to the specific needs of the question during the *online stage*, directly addressing the challenge of aligning the table’s structure with the NL question’s semantics.

**LLMs-based Multi-Agent Framework.** A multi-agent LLM framework refers to a well-designed hierarchical structure consisting of multiple LLM-based agents and scheduling algorithms [25]. Compared with single-agent methods based on prompting techniques, such structures are better suited for handling complex tasks like software development, issue resolution, and code generation [12, 28, 30, 31, 43, 46]. The structure of multi-agent frameworks can be categorized into equi-level [47], hierarchical [6, 26], and nested [11] structures. AutoTQA [60] proposes a multi-agent framework with hierarchical structure for supporting Tabular Question Answering.

We adopt the hierarchical structure because our design introduces a Planner agent that decomposes the overall data preparation task into three distinct sub-tasks, each handled by specialized agents. This top-down coordination naturally aligns with the principles of the hierarchical multi-agent framework. Moreover, while AutoTQA focuses on improving table analysis, this paper focuses on the performance bottleneck caused by data prep issues in TQA. Our proposed framework, AUTOPREP, addresses question-aware data prep for TQA and can be integrated as a plugin into current TQA approaches to further improve the overall performance.

## 8 CONCLUSION AND FUTURE WORK

In this paper, we have introduced AUTOPREP, an LLM-based multi-agent framework to support data prep for TQA tasks. AUTOPREP consists of three stages: (1) the Planning stage, which suggests logical data prep operations, (2) the Programming stage, which generates physical implementations for each logical operation, and (3) the Executing stage, which executes the Python code and reports error messages. We propose a Chain-of-Clauses method to generate high-quality logical plans and a Tool-augmented method for effective physical plan generation. Extensive experiments on real datasets demonstrate the superiority of AUTOPREP.

For future work, we identify three promising directions. First, we aim to extend the system’s data prep capabilities by integrating a broader set of operations to handle more complex issues such as missing values, duplicates, etc. Second, AUTOPREP currently focuses on single-table question answering, and extending its capabilities to support multi-table TQA is essential for tackling more realistic and complex scenarios. The third direction is to enable question-aware data preparation over enterprise datasets, which are more complex than existing TQA benchmarks.

## ACKNOWLEDGMENT

This paper was partly supported by the NSF of China (62436010, 62441230, 62525202 and 62232009), National Key R&D Program of China (2023YFB4503600), NSF DBI-2327954, Guangdong Provincial Project 2023CX10X008 and Amazon Research Awards.



## REFERENCES

- [1] 2023. *Code of Binder*. <https://github.com/clang-ai/Binder>
- [2] 2023. *Code of Dater*. <https://github.com/AlibabaResearch/DAMO-ConvAI>
- [3] 2024. *Code of Chain-of-Table*. <https://github.com/google-research/chain-of-table>
- [4] 2024. *Code of ReAcTable*. <https://github.com/yunjiazhong/ReAcTable>
- [5] 2025. *Technical Report*. <https://github.com/ruc-datalab/AutoPrep/blob/main/pdf/report.pdf>
- [6] Sanjeevan Ahilan and Peter Dayan. 2019. Feudal multi-agent hierarchies for cooperative reinforcement learning. *arXiv preprint arXiv:1901.08492* (2019).
- [7] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *35th Conference on Neural Information Processing Systems, NeurIPS 2021*. Neural Information Processing Systems foundation.
- [8] Laure Berti-Equille. 2019. Learn2Clean: Optimizing the Sequence of Tasks for Web Data Preparation. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 2580–2586. <https://doi.org/10.1145/3308558.3313602>
- [9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS 2020*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.).
- [10] Chengliang Chai, Nan Tang, Ju Fan, and Yuyu Luo. 2023. Demystifying Artificial Intelligence for Data Preparation. In *SIGMOD*, Sudipto Das, Ippokratis Pandis, K. Selçuk Candan, and Sihem Amer-Yahia (Eds.). ACM, 13–20. <https://doi.org/10.1145/3555041.3589406>
- [11] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).
- [12] Dong Chen, Shaoxin Lin, Muhan Zeng, Daoguang Zan, Jian-Gang Wang, Anton Cheshkov, Jun Sun, Hao Yu, Guoliang Dong, Artem Aliev, et al. 2024. CodeR: Issue Resolving with Multi-Agent and Task Graphs. *arXiv preprint arXiv:2406.01304* (2024).
- [13] Sibe Chen, Nan Tang, Ju Fan, Xuemi Yan, Chengliang Chai, Guoliang Li, and Xiaoyong Du. 2023. HAIPipe: Combining Human-generated and Machine-generated Pipelines for Data Preparation. *Proc. ACM Manag. Data* 1, 1 (2023), 91:1–91:26. <https://doi.org/10.1145/3588945>
- [14] Wenhui Chen. 2022. Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv:2210.06710* (2022).
- [15] Wenhui Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164* (2019).
- [16] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2022. HiTab: A Hierarchical Table Dataset for Question Answering and Natural Language Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1094–1110.
- [17] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2023. Binding Language Models in Symbolic Languages. In *International Conference on Learning Representations (ICLR 2023)*(01/05/2023-05/05/2023, Kigali, Rwanda).
- [18] 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* (2024).
- [19] Ori Bar El, Tova Milo, and Amit Somech. 2020. Automatically Generating Data Exploration Sessions Using Deep Reinforcement Learning. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo (Eds.). ACM, 1527–1537. <https://doi.org/10.1145/3318464.3389779>
- [20] Meihao Fan, Xiaoyue Han, Ju Fan, Chengliang Chai, Nan Tang, Guoliang Li, and Xiaoyong Du. 2024. Cost-Effective In-Context Learning for Entity Resolution: A Design Space Exploration. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*. IEEE, 3696–3709. <https://doi.org/10.1109/ICDE60146.2024.00284>
- [21] Matthias Feurer, Katharina Eggenberger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074* 24 (2020), 8.
- [22] Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-Operations Aware Fact Verification via Sentence-Table Cloze Pre-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 4971–4983. <https://doi.org/10.18653/V1/2022.EMNLP-MAIN.331>
- [23] Zihui Gu, Ruixue Fan, Xiaoman Zhao, Meihui Zhang, Ju Fan, and Xiaoyong Du. 2022. OpenTFV: An Open Domain Table-Based Fact Verification System. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 2405–2408. <https://doi.org/10.1145/3514221.3520163>
- [24] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming - The Rise of Code Intelligence. *CoRR* abs/2401.14196 (2024). <https://doi.org/10.48550/ARXIV.2401.14196> arXiv:2401.14196
- [25] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. 2024. LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578* (2024).
- [26] Keegan Harris, Steven Wu, and Maria Florina Balcan. 2023. Stackelberg games with side information. In *Multi-Agent Security Workshop@ NeurIPS'23*.
- [27] Yuval Hefetz, Roman Vainshtein, Gilad Katz, and Lior Rokach. 2020. DeepLine: AutoMeFT Tool for Pipelines Generation using Deep Reinforcement Learning and Hierarchical Actions Filtering. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 2103–2113. <https://doi.org/10.1145/3394486.3403261>
- [28] Sirui Hong, Xiauwu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023).
- [29] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.
- [30] Yoichi Ishibashi and Yoshimasa Nishimura. 2024. Self-organized agents: A llm multi-agent framework toward ultra large-scale code generation and optimization. *arXiv preprint arXiv:2404.02183* (2024).
- [31] Md Ashrafur Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. MapCoder: Multi-Agent Code Generation for Competitive Problem Solving. *arXiv preprint arXiv:2405.11403* (2024).
- [32] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406* (2022).
- [33] Peng Li, Yeye He, Cong Yan, Yue Wang, and Surajit Chaudhuri. 2023. Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples. *Proceedings of the VLDB Endowment* 16, 11 (2023), 3391–3403.
- [34] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. 2024. Long-context LLMs Struggle with Long In-context Learning. *CoRR* abs/2404.02060 (2024). <https://doi.org/10.48550/ARXIV.2404.02060> arXiv:2404.02060
- [35] Weizheng Lu, Jing Zhang, Ju Fan, Zihao Fu, Yueguo Chen, and Xiaoyong Du. 2025. Large language model for table processing: a survey. *Frontiers Comput. Sci.* 19, 2 (2025), 192350. <https://doi.org/10.1007/S11704-024-40763-6>
- [36] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference, Stéfan van der Walt and Jarrod Millman (Eds.)*. 51 – 56.
- [37] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics* 10 (2022), 35–49.
- [38] Randal S Olson and Jason H Moore. 2016. TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on automatic machine learning*. PMLR, 66–74.
- [39] Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. MultiTabQA: Generating Tabular Answers for Multi-Table Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6322–6334.
- [40] Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1470–1480.
- [41] Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M. Rzeszotarski, and Jiannan Wang. 2021. DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. In *SIGMOD*. ACM, 2271–2280. <https://doi.org/10.1145/3448016.3457330>
- [42] Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M. Rzeszotarski, and Jiannan Wang. 2021. DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20–25, 2021, Virtual Event, China*.
- [43] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

- arXiv preprint arXiv:2307.07924* 6 (2023).
- [44] Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022. Evaluating the text-to-sql capabilities of large language models. *arXiv preprint arXiv:2204.00498* (2022).
  - [45] Zeyuan Shang, Emanuel Zraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, and Tim Kraska. 2019. Democratizing Data Science through Interactive Curation of ML Pipelines. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska (Eds.). ACM, 1171–1188. <https://doi.org/10.1145/3299869.3319863>
  - [46] Wei Tao, Yucheng Zhou, Wenqiang Zhang, and Yu Cheng. 2024. MAGIS: LLM-Based Multi-Agent Framework for GitHub Issue Resolution. *arXiv preprint arXiv:2403.17927* (2024).
  - [47] Mikhail Terekhov, Romain Graux, Eduardo Neville, Denis Rosset, and Gabin Kolly. 2023. Second-order Jailbreaks: Generative Agents Successfully Manipulate Through an Intermediary. In *Multi-Agent Security Workshop@ NeurIPS'23*.
  - [48] Chris Thornton, Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 847–855.
  - [49] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Chengliang Chai, Guoliang Li, Ruixue Fan, and Xiaoyong Du. 2022. Domain Adaptation for Deep Entity Resolution. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 443–457. <https://doi.org/10.1145/3514221.3517870>
  - [50] 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. 2024. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398* (2024).
  - [51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
  - [52] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. 2024. TableBench: A Comprehensive and Complex Benchmark for Table Question Answering. *arXiv preprint arXiv:2408.09174* (2024).
  - [53] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).
  - [54] Chengrun Yang, Jicong Fan, Ziyang Wu, and Madeleine Udell. 2020. AutoML Pipeline Selection: Efficiently Navigating the Combinatorial Space. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 1446–1456. <https://doi.org/10.1145/3394486.3403197>
  - [55] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).
  - [56] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. [https://openreview.net/forum?id=WE\\_vluYUL-X](https://openreview.net/forum?id=WE_vluYUL-X)
  - [57] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. 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*. 174–184.
  - [58] Yuxin Zhang, Meihao Fan, Ju Fan, Mingyang Yi, Yuyu Luo, Jian Tan, and Guoliang Li. 2025. Reward-sql: Boosting text-to-sql via stepwise reasoning and process-supervised rewards. *arXiv preprint arXiv:2505.04671* (2025).
  - [59] Yunjia Zhang, Jordan Henkel, Avriella Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M Patel. 2024. ReAcTable: Enhancing ReAct for Table Question Answering. *Proceedings of the VLDB Endowment* 17, 8 (2024), 1981–1994.
  - [60] Jun-Peng Zhu, Peng Cai, Kai Xu, Li Li, Yishen Sun, Shuai Zhou, Haihuang Su, Liu Tang, and Qi Liu. 2024. AutoTQA: Towards Autonomous Tabular Question Answering through Multi-Agent Large Language Models. *Proc. VLDB Endow.* 17, 12 (2024), 3920–3933. <https://www.vldb.org/pvldb/vol17/p3920-zhu.pdf>