

# Efficient Prompting for LLM-based Generative Internet of Things

Bin Xiao, Burak Kantarci, *Senior Member, IEEE*, Jiawen Kang, *Senior Member, IEEE*  
Dusit Niyato, *Fellow, IEEE*, Mohsen Guizani, *Fellow, IEEE*

**Abstract**—Large language models (LLMs) have demonstrated remarkable capacities on various tasks, and integrating the capacities of LLMs into the Internet of Things (IoT) applications has drawn much research attention recently. Due to security concerns, many institutions avoid accessing state-of-the-art commercial LLM services, requiring the deployment and utilization of open-source LLMs in a local network setting. However, open-source LLMs usually have more limitations regarding their performance, such as their arithmetic calculation and reasoning capacities, and practical systems of applying LLMs to IoT have yet to be well-explored. Therefore, we propose a text-based generative IoT (GIoT) system deployed in the local network setting in this study. To alleviate the limitations of LLMs and provide service with competitive performance, we apply prompt engineering methods to enhance the capacities of the open-source LLMs, design a Prompt Management Module and a Post-processing Module to manage the tailored prompts for different tasks and process the results generated by the LLMs. To demonstrate the effectiveness of the proposed system, we discuss a challenging Table Question Answering (Table-QA) task as a case study of the proposed system, as tabular data is usually more challenging than plain text because of their complex structures, heterogeneous data types and sometimes huge sizes. We conduct comprehensive experiments on two popular Table-QA datasets, and the results show that our proposal can achieve competitive performance compared with state-of-the-art LLMs, demonstrating that the proposed LLM-based GIoT system can provide competitive performance with tailored prompting methods and is easily extensible to new tasks without training.

**Index Terms**—Generative Internet of Things, Table Question Answering, Prompt Engineering, Large Language Model

## I. INTRODUCTION

**A**RTIFICIAL Intelligence of Things (AIoT), integrating Artificial Intelligence (AI) and Internet of Things (IoT) for efficient data analysis and intelligent decision making [1], have been widely discussed in many studies and applied in many scenarios, such as Healthcare [2], Smart Cities [3] and Industries [4]. Currently, task-specific machine learning and Deep Neural Network (DNN) models are the mainstream choices for AIoT providing services to IoT devices, which are often deployed on the cloud and edge servers [1]. With the development of Generative Artificial Intelligence (GAI),

B. Xiao and B. Kantarci are with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON, Canada. Emails: {bxiao103, burak.kantarci}@uottawa.ca

J. Kang is with Guangdong University of Technology, China, Email: kavinkang@gdut.edu.cn

D. Niyato is with Nanyang Technological University, Singapore, Email: dnnyiato@ntu.edu.sg

M. Guizani is with Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE, mohsen.guizani@mbzuai.ac.ae

especially large language models (LLMs), leveraging its remarkable general capacities to the IoT applications, termed as Generative Internet of Things (GIoT) [5], becomes a promising research direction [6], [7]. Different from task-specific models, LLMs, such as GPT-4 [8], have demonstrated their general capacities on a wide range of tasks, such as data analysis, code generation, reasoning, planning and many others, making it possible to provide various services with a single LLM model, maintaining competitive performance with tailored task-specific models.

Despite the remarkable capacities of LLMs, many issues need to be considered when integrating LLMs into IoT systems for an LLM-based GIoT system. First, following the Scaling Law [9], the capacities of an LLM are growing with its number of parameters and the scale of the training dataset, making LLMs often have a tremendous number of parameters, which lead to high hardware requirements for the training and inference. For example, popular open-source LLMs providing competitive performance on some public benchmarks often have around 70 billion parameters, such as Mixtral-8\*7B [10] and Llama-3-70B [11]. These numbers of parameters make it not practical to deploy LLMs on edge IoT devices. Even though deploying LLMs on the edge and cloud servers can be an option, the efficiency of an LLM-based GIoT system still needs to be carefully considered. Besides the hardware requirements and efficiency issues, data privacy and security issues hinder many institutions from accessing commercial state-of-the-art LLMs. For example, for a healthcare IoT application collecting private data from patients, it is necessary to avoid uploading collected data to public, commercial LLM services for data analysis because of privacy issues. Therefore, a safe, transparent, and controllable open-source LLM deployed in a local network is more suitable for many institutions, even though commercial LLMs can provide better services. At last, the performance and the scalability of a LLM-based GIoT system are another two critical considerations because both commercial and open-source LLMs have some inherent limitations [12]–[14], such as their hallucination issues, limited reasoning capacities for complex tasks. Typically, prompting methods and fine-tuning are two directions that can alleviate these limitations. Specifically, prompting methods improve the performance of an LLM by designing tailored prompts for different tasks to elicit the capacities of an LLM. For example, Chain of Thought (CoT) [14] is a popular prompting method to improve the LLM’s reasoning performance by providing reasoning rationales. Program of Thoughts (PoT) [13] is another prompting method generating Python code to offload

the reasoning and calculation tasks to the Python interpreter. Since these prompting methods focus on tailoring task-specific prompts for LLMs, they often lead to longer inference time because of their larger number of prompting tokens. By contrast, fine-tuning an LLM for a task can also improve the performance but requires label datasets and computation resources for the model training. As discussed in some studies [15], [16], fine-tuning an LLM for unseen tasks can increase the model's bias and degrade its general capacities.

Since practical solutions applying LLMs to the IoT setting have yet to be well-explored, even though there have been some studies [6], [7] discussing the potential and possible frameworks of such applications, we propose a practical LLM-based GIoT system in this study. Considering the discussed possible issues for an LLM-based GIoT system, we propose to deploy the open-source LLM on the edge server in a local network setting to address the and employ prompting methods to enhance the capacities of the proposed system for different tasks. Specifically, a Prompt Management Module and a Post-processing Module are proposed to be deployed in the edge server, in which the former is responsible for the selection, management and creation of prompts and demonstrations for the requests from IoT devices, and the latter is responsible for post-processing the results generated by the LLMs, as shown in Figure 1. With the proposed Prompt Management Module and Post-processing Module and a Task-specific Prompts Database to address the considerations in data privacy and security, system scalability, and enhance the capacities of the LLM by integrating prompting methods.

To demonstrate the effectiveness of the proposed LLM-based GIoT system, we implement a Semi-structured Table Question Answering (Table-QA) service in the proposed system, which is useful and challenging. Specifically, the Table-QA service aims to answer the query question based on the given table information. Since tables are widely used to summarize critical information in many data sources, such as visually rich documents and web pages [17], Table-QA [18]–[22] can be a useful service to provide analysis to the tabular data, and has also drawn much research attention recently. Typically, tables can be easily categorized into two groups: structured and semi-structured tables. Structured tables are usually from relational database systems with explicit schema describing their structures and data types, meaning the programming languages, such as SQL, can naturally process them. By contrast, tables from other sources, such as web pages and visually rich documents, are usually semi-structured without schema requirements, resulting in complex structures, heterogeneous data types and sometimes huge sizes, making the Table-QA task on these semi-structured tables more challenging. Therefore, we use the Table-QA problem on the semi-structured tables to verify the proposed LLM-based system, which is a more challenging setting. To provide competitive service as commercial LLMs, we propose a three-stage prompting method, including task-planning, task-conducting and task-correction stages, leveraging Python code to conduct reasoning steps. The proposed prompting method can improve the performance of open-source LLMs, which can also demonstrate that the proposed system can be easily

extended to new tasks with prompting methods.

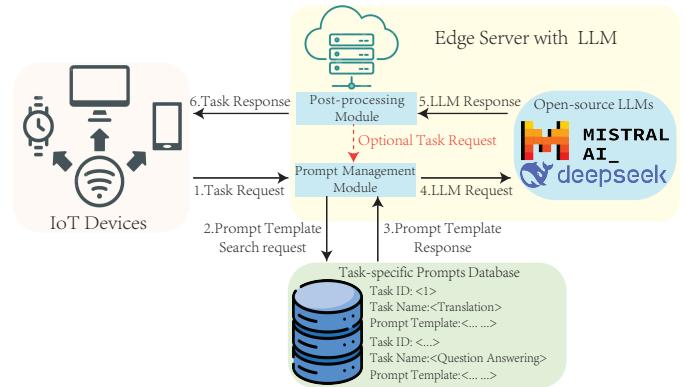


Fig. 1: Overall architecture of the proposed LLM-based GIoT system.

To sum up, the contributions of this study can be 3-fold:

- 1) We propose an LLM-based GIoT system with open-source LLMs deployed in the local network setting, which includes a Prompt Management Module, a Post-processing Module and a Task-specific Prompts Database to address the considerations in data privacy and security, system scalability, and enhance the capacities of the LLM by integrating prompting methods.
- 2) We discuss a challenging Table-QA problem to demonstrate the proposed LLM-based GIoT system and propose a three-stage prompting solution, including task-planning, task-conducting and task-correction stages, to alleviate the issues caused by the complex structures, heterogeneous data types, huge tables and the limitations of LLMs and also reduce the inference cost. A series of atomic operations is proposed to describe and measure the similarities of QA tasks and select proper demonstrations for prompt creation.
- 3) We conduct extensive experiments on the WikiTableQA and TabFact datasets with open-source LLMs to verify the proposed LLM-based GIoT system and the proposed prompting method. The experimental results demonstrate that our proposed prompting method can outperform the baseline methods by a large margin, achieving state-of-the-art performance. We also conduct comprehensive analyses that consider the performance of different prompting methods, the inference costs, and the behaviour of different open-source LLMs, which can be a guide for selecting LLMs and prompting methods for an LLM-based GIoT system.

The rest of this paper is organized as follows: Section II discusses related studies, including recent studies applying LLMs to the IoT systems, Table-QA solutions and prompting methods. Section IV describes our proposed LLM-based GIoT system and prompting solution. Section V shows the experimental results and discusses the design aspects of the proposed prompting method. At last, we draw our conclusion and possible future directions in Section VI.

## II. RELATED WORK

### A. Generative Models in IoT

As generative models, especially text-based LLMs, have demonstrated their remarkable capacities in a wide range of tasks, integrating their capacities into IoT applications has attracted the research community's attention. There have been some studies [5], [6], [23] summarizing the critical components of Generative Models and discussing the potential of applying Generative Models to IoT systems. Authors in study [5] firstly term the combination of generative models and IoT applications as Generative IoT (GIoT), discuss the foundations of generative models and the potential of GIoT applications, including Vision-based, Audio-based, Text-based and other GIoT applications. Similarly, authors in study [6] also point out various potential GIoT applications in many fields, such as Mobile Networks, Autonomous Vehicles, and many others. Besides discussing the insights and potentials of GIoT, some studies examine various aspects of applying LLMs in IoT settings. CASIT [7] is an LLM-agent-based IoT framework proposing a Sensor Interface to convert sensor data into natural language and design multiple types of LLM-based Agent to cooperate and analyze the sensor data and user requests. LLMind [24] introduces a framework that integrates various domain-specific AI modules and enables IoT device cooperation to enhance the LLM's capabilities for conducting complex tasks. Study [25] proposes an intelligent control framework integrating Integrated Terrestrial Non-terrestrial Networks, IoT and language models, identifying key components, potential applications and challenges. Overall, current studies regarding GIoT applications mainly focus on proposing abstractions of the systems, and practical cases need to be explored in more depth. Therefore, this study proposes an extensible LLM-based GIoT system using prompting methods and uses a challenging Table-QA problem as a case study to illustrate the proposed system.

### B. Table Question Answering

Since this study uses the Table-QA task as the case study of the proposed LLM-based GIoT system, we include recent studies using LLM to solve the Table-QA problem in this section. Specifically, most of the studies applying LLMs to the Table-QA task can be categorized into instruction tuning based and prompt engineering based approaches. Instruction-tuning approaches usually need to collect large-scale datasets and then further fine-tuned LLMs with parameter-efficient fine-tuning methods, such as LORA [26] and LongLORA [27]. TableLLAMA [28] and TAT-LLM [20] are typical examples of applying instruction tuning for the table processing. TableLLAMA is a generalist model for tables fine-tuned on Llama2 [29] with LongLORA. For fine-tuning TableLLAMA, a dataset collection named TableInstruct is proposed by collecting table-based samples from 14 datasets for 11 tasks. TAT-LLM is another fine-tuned LLAMA2 model specifically for the discrete reasoning over tables, which decomposes the Table-QA into three steps: Extractor, Reasoner and Executor. To construct the dataset for the model fine-tuning, TAT-LLM

proposes a template to guide LLM in generating data following the proposed step-wise pipeline.

On the other hand, prompting engineering based solutions focus on proposing proper prompts to the language model to elicit LLMs' capacities. Since Table-QA needs to extract relevant information and evidence from tables and perform reasoning, decomposing the Table-QA task into multiple steps is also widely adopted in prompting engineering-based solutions. Besides, as LLM often fails to process large tables and complex reasoning, many studies [20], [30]–[32] propose to decompose large tables into small tables in the extraction step and divide the complex reasoning task into more straightforward reasoning questions. For example, StructGPT [33] defines specialized interfaces for information extraction and linearizes the extracted sub-tables as inputs to the LLM for question answering. EEDP [34] is another example proposing a prompting method containing four steps: Elicit, Extract, Decompose and Predict.

Besides these studies decomposing the Table-QA into extraction and reasoning steps and dividing difficult extraction and reasoning tasks into simpler ones, programming languages, such as SQL and Python, are also widely leveraged in these solutions to overcome the limitation of LLMs in arithmetic calculation and reasoning. Dater [31] is a typical solution following the multi-step design and leveraging SQL to conduct reasoning. More specifically, Dater decomposes both large evidence and complex questions into relevant and simpler ones, applies SQL queries to produce numerical and logical reasoning results to the decomposed sub-questions, and generates the final results based on the sub-results and extracted evidence with ICL. Binding [35] is another example of applying Python and SQL for the table processing, which proposes a unified API to map tasks into executable programs.

All in all, as the complexities of Table-QA in the information extraction and reasoning, breaking Table-QA into multiple steps and decomposing difficult extraction and reasoning tasks into simpler ones have been the dominant solution, and external tools such as SQL and Python, can compensate the limitations of LLMs in arithmetic calculation and reasoning.

### C. Prompting Methods

This section focuses on recent prompting methods for LLMs because our proposed LLM-based GIoT system applies prompting methods to enhance the LLMs' capacities for different tasks. As some of these methods have also been applied to the Table-QA problem, some methods included in this section can be overlapped with the discussed studies in Section II-B. There have been many studies demonstrated that prompting LLMs with a few demonstrations and intermediate reasoning steps can elicit the reasoning capacities of LLMs, which often refer to In Context Learning (ICL) [36] and CoT [14]. Although ICL and CoT are useful, writing proper prompting demonstrations is non-trivial and time-consuming. Therefore, some studies [37]–[39] propose to use LLMs to generate demonstrations. Auto-CoT [37] proposes to prompt LLMs with Zero-shot CoT to generate reasoning rationales but finds that the generated rationales often contain mistakes. To

mitigate the wrong demonstration issue, Auto-CoT proposes clustering and sampling the questions first and then applying simple heuristics to sample simpler questions and rationales to construct demonstrations. Synthetic Prompting [38] is another typical solution for constructing demonstrations with LLMs, containing forward and backward processes. In the backward process of Synthetic Prompting, a topic word, a target complexity and a self-generated reasoning chain are used as conditions to generate the synthetic question, and the synthetic question is used in the forward process to generate the precise synthetic reasoning chain. Along with these studies focusing on demonstration generation with LLMs, ensemble methods are also effective for reasoning. For example, self-consistency decoding [40] first generates a set of candidate outputs from the LLM with different sampling methods, such as temperature sampling, and then aggregates the sampled outputs and uses the most consistent output as the final result. Multi-Chain Reasoning [41] is another ensemble method mixing information from multiple reasoning chains. Different from studies [40] to ensemble results of multi-reasoning chains, Multi-Chain Reasoning collects evidence from multiple reasoning chains and prompts the LLM to give the final answer.

In addition to these prompting methods, the integration of programming languages and other external tools holds great potential for overcoming the limitations of LLMs. For instance, LLMs often struggle with precise arithmetic calculations, especially division operations, and staying updated with the latest information. To address these issues, PoT [13], PAL [12], and many other studies [42] leverage Programming Language to assist the reasoning steps. Some studies [43] introduce the concept of Agent and call external tools by parsing the LLM outputs to enhance the LLM's capabilities.

### III. SYSTEM MODEL

As discussed in Section I, in this study, we consider the scenarios in which IoT devices cannot access commercial LLMs because of data privacy and security considerations, which is a practical setting in many institutions, such as hospitals. Alternatively, as shown in Figure 1, an open-source LLM can be deployed in a local edge server to process the requests from IoT devices. Because fine-tuning an LLM for a specific task requires substantial datasets and computational resources, limiting the scalability of extending to multiple tasks, limiting the scalability of extending to multiple tasks, we propose to use prompting methods to enhance the capacities for different tasks with a Prompt Management Module, a Post-processing Module and a Task-specific Prompts Database, as shown in Figure 1. For a text-based request from IoT devices, such as a translation task to a sentence, the proposed Prompt Management Module is responsible for parsing the request and searching the task-specific prompt template and demonstrations regarding the task, as steps 1, 2 and 3 in Figure 1. After obtaining the task-specific prompt template, the Prompt Management Module constructs the final prompt and sends the request to the LLM. In our design, the capacities to deal with new tasks can be easily extended to the system by adding new prompt templates and demonstrations in the

Task-specific Prompt Database without training or fine-tuning to the LLM. Since the results generated by the LLM are sometimes not ideal, we propose a Post-processing Module to process the results further. For example, some methods must prompt the LLM model multiple times to obtain the final results, whose intermediate results and Optional Task Request should be processed by the Post-processing Module. Besides, programming language-aided prompting methods, such as PAL [12] and PoT [13], generate the Python code instead of the final results, which means that the Post-processing Module should also be responsible for executing the generated Python code to obtain the final results. It is worth mentioning that, in the proposed solution, the IoT devices and the edge server can be easily connected with Wi-Fi and other wireless network protocols, and the communication between them can be easily implemented with HTTP protocol.

Following these discussed steps, Figure 2 shows the detailed components of the proposed two modules and the detailed workflow of the proposed GIoT system. Specifically, the Prompt Management Module consists of three components: Request Parsing, Prompt Search, and Prompt Generation. The Request Parsing component receives and parses the requests from the IoT devices and outputs Task ID, Task Step and parsed Data, in which Task ID and Task Step would be used in the Prompt Search component to search corresponding Prompt Instructions and Prompt Demonstrations from the Task Specific Prompts Database. The parsed Data generated by the Request Parsing component is the information that needs to be further processed by the LLM. For example, for a service of translating English to Chinese, the request from the IoT devices can be "*Task Name: Translation. Data: This is a sample translation service.*" Then the Data parsed by the Request Parsing component should be "*This is a sample translation service.*", which would be used by the Prompt Generation component to generate the final Task Prompt together with the Prompt Instruction and Prompt Demonstrations. It is worth mentioning that Prompt Generation can implement customize demonstration selection methods to optimize the performance of applying In Context Learning (ICL) [36]. For the Post-processing Module, since the output of the LLM cannot always follow the instructions, a Result Parsing function is needed to refine the outputs. For single-stage prompting methods, the result generated by the Result Parsing function can be the final result returned to the IoT device, as the path 1 shown in Figure 2. By contrast, prompting methods, such as PoT [13] and PAL [12], generate Python code and conduct the reasoning steps by an external Python Interpreter. Therefore, a Code Execution component should be included in the Post-processing Module. At last, a Optional Request Management component need to be implemented for the prompting methods with multiple-stages.

### IV. CASE STUDY OF TABLE-QA FOR THE LLM-BASED GIOT SYSTEM

In this section, we use the Table-QA task as a case study to demonstrate the proposed LLM-based GIoT System. We first provide the problem formulation and examine the challenging

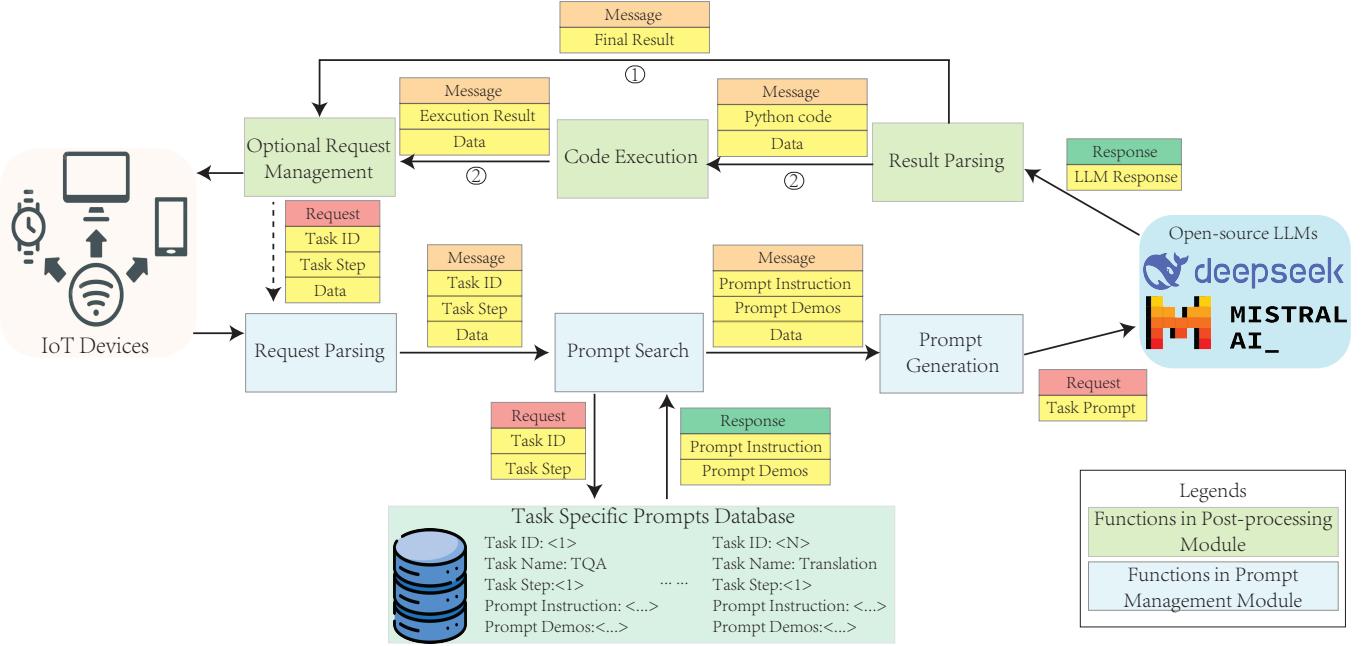


Fig. 2: Detailed workflow of the proposed LLM-based GIoT system.

aspects of the Table-QA task in Section IV-A. Then, we describe our proposed three-stage prompting method in detail in Section IV-B.

#### A. Problem Formulation and Analysis

For a Table-QA service deployed in the proposed LLM-based GIoT system, it contains two key components: a parametric a LLM with parameters  $\theta$  and a Retriever  $\mathcal{R}$ , in which the Retriever  $\mathcal{R}$  is responsible for selecting proper demonstrations used in ICL. Therefore, for a input query pair,  $(t, q)$ , including a query table and a query question, the output answer  $y$  can be  $y = f_{\theta}(t, q, \mathcal{R})$ .

As mentioned in Section I, many prompting methods have been proposed to enhance LLMs' capacities. For example, Chain of Thought (CoT) [14] is a popular prompting method providing reasoning rationales to elicit LLMs' reasoning capacities. PAL [12] and PoT [13] propose to offload the reasoning steps to Python codes. However, these typical prompting methods cannot perform well in the semi-structured Table-QA problem because of the complex structures, heterogeneous data types, and sometimes huge tables with numerous columns and rows. More specifically, prompting methods without applying programming languages must first extract the correct information from semi-structured tables before reasoning, which is challenging for LLMs, especially when the table is huge [30]. Figure 3 contains a failure example of CoT, which interprets 1936/37 as two years and fails to extract the correct Year 1953/54. Similarly, programming-aided solutions, such as PAL and PoT, can also suffer from this information extraction issue if we define the relevant information as Python variables. Applying Pandas Library [44], [45] can alleviate this information extraction issue [46] by providing proper selection criteria, but introducing extra difficulties caused by

the heterogeneous data types. Figure 3 shows a PoT example using Pandas Library, which fails to run because the values in column *Year* cannot be directly compared with 1936. Besides, complex structures of semi-structured tables can often lead to wrong results, especially for program-aided solutions. Figure 4a shows an example from WikiTableQA [47] dataset, which contains several spanning cells across multiple columns and rows, and inconsistent data types, such as the column *Season*. Even though some methods [48]–[50] can transform this table into a standard table by repeating table spanning cell into multiple single table cells so that SQL or Python Pandas library can process it, its structure still can lead to wrong results. For example, when an LLM is prompted to generate Python code to answer the question "What is the maximum League Apps after 2004?". Two *Totals* in the column *Season* will be compared with correct *Seasons*, which leads to wrong results. Besides, the generated code needs to compare the values from the column *Season*, which can lead to an error of execution because " $>$ " operation cannot be applied to the string "2011-12" and the integer 2004, as highlighted in Figure 4b, which is another failure example caused by the heterogeneous data types. Along with applying Python to enhance the LLMs, some studies [31], [35] apply SQL to the Table-QA problem. However, SQL is a programming language designed for structured tables, meaning that semi-structured tables need to be transformed into structured format first so that tables can be imported into the relational databases to execute SQL queries, which is another challenging task for the semi-structured Table-QA problem discussed in this study. Besides, as pointed out by some studies [35], [51], some questions are not answerable by merely using SQL. Therefore, considering the flexibility of Python and the limitations of applying SQL to the semi-structured Table-QA task, we apply

| New York Americans (soccer) |          |        |                  |                       |              |
|-----------------------------|----------|--------|------------------|-----------------------|--------------|
| Year                        | Division | League | Reg. Season      | Playoffs              | National Cup |
| 1931                        | 1        | ASL    | 6th (Fall)       | No playoff            | N/A          |
| Spring 1932                 | 1        | ASL    | 5th?             | No playoff            | 1st Round    |
| Fall 1932                   | 1        | ASL    | 3rd              | No playoff            | N/A          |
| Spring 1933                 | 1        | ASL    | ?                | ?                     | Final        |
| 1933/34                     | N/A      | ASL    | 2nd              | No playoff            | ?            |
| 1934/35                     | N/A      | ASL    | 2nd              | No playoff            | ?            |
| 1935/36                     | N/A      | ASL    | 1st              | Champion (no playoff) | ?            |
| 1936/37                     | N/A      | ASL    | 5th, National    | Did not qualify       | Champion     |
| 1937/38                     | N/A      | ASL    | 3rd(t), National | 1st Round             | ?            |
| 1938/39                     | N/A      | ASL    | 4th, National    | Did not qualify       | ?            |
| 1939/40                     | N/A      | ASL    | 4th              | No playoff            | ?            |
| 1940/41                     | N/A      | ASL    | 6th              | No playoff            | ?            |
| 1941/42                     | N/A      | ASL    | 3rd              | No playoff            | ?            |
| 1942/43                     | N/A      | ASL    | 6th              | No playoff            | ?            |
| 1943/44                     | N/A      | ASL    | 9th              | No playoff            | ?            |
| 1944/45                     | N/A      | ASL    | 9th              | No playoff            | ?            |
| 1945/46                     | N/A      | ASL    | 5th              | No playoff            | ?            |
| 1946/47                     | N/A      | ASL    | 6th              | No playoff            | ?            |
| 1947/48                     | N/A      | ASL    | 6th              | No playoff            | ?            |
| 1948/49                     | N/A      | ASL    | 1st(t)           | Finals                | ?            |
| 1949/50                     | N/A      | ASL    | 3rd              | No playoff            | ?            |
| 1950/51                     | N/A      | ASL    | 5th              | No playoff            | ?            |
| 1951/52                     | N/A      | ASL    | 6th              | No playoff            | ?            |
| 1952/53                     | N/A      | ASL    | 6th              | No playoff            | Semifinals   |
| 1953/54                     | N/A      | ASL    | 1st              | Champion (no playoff) | Champion     |
| 1954/55                     | N/A      | ASL    | 8th              | No playoff            | ?            |
| 1955/56                     | N/A      | ASL    | 6th              | No playoff            | ?            |

Question: how long did it take for the new york americans to win the national cup after 1936?

<CoT Prompt: Instructions, Demonstrations, [Full Table](#), Question>

Answer: The New York Americans won the national cup in 1936 and 1937. Therefore, it took them one year to win the national cup after 1936.

CoT

<PoT Prompt: Instructions, Demonstrations, [Full Table](#), Question>

Answer:

```
def solution(table_dict):
    ... More lines here ...
    winning_year = df[(df['League'] == 'ASL') & (df['National Cup'] == 'Champion') & (df['Year'] > 1936)][['Year']].min()
    return winning_year - 1936
```

PoT

<Task Planning Prompt: Instructions, Demonstrations, [Statistical Table](#), Question>

Answer: Relevant columns :{Year, National Cup}, Operations: {ADDITION/DIFF}, Programming Steps Demonstration selection and sub-table extraction

<Task Conducting Prompt: Instructions, Demonstrations, [SubTable](#), Question>

Answer:

```
def solution(table_dict): <----- reasoning function
    ... More lines here ...
    winning_year = df[(df['League'] == 'ASL') & (df['National Cup'] == 'Champion') & (df['Year'] > 1936)][['Year']].min()
    return winning_year - 1936
```

Therefore, the final answer is [1]. <----- default answer when Python function fails to run

<Task Correction Prompt: Instructions, Demonstrations, [SubTable](#), Error Code, Question>

```
def normalize_year(year): <----- data normalization function for error correction
    ...More lines here...
```

Data normalization and re-run Task-conducting Function Proposed

Fig. 3: Comparison of CoT, PoT and the proposed method. It is worth mentioning that many details are omitted due to space limitations. The proposed method contains task-planning, task-conducting, and task-correction stages, and it uses a statistical table and sub-tables in these stages to avoid the original huge tables.

Python instead of SQL in our solution.

Besides the issues of applying prompting methods to the semi-structured Table-QA problem, current studies only focus on optimizing the prediction accuracy without considering their inference cost, which is critical for our proposed GIoT system. As mentioned, some tables can be huge, which can lead to long prompts and inference time. Even though some solutions [31], [52] propose decomposing huge tables into sub-tables for further reasoning, they need to prompt the full huge table into the LLM first. Some studies [30] truncate the huge tables, which can drastically reduce the inference cost but introduce the risk of losing critical information in the tables and sometimes make it impossible to give the correct answer. Besides, ensemble methods, such as self-consistency decoding and majority voting, are often employed to improve the performance further [31], [40], [52], but also drastically increasing the inference cost simultaneously.

Lastly, most of these studies are based on In Context Learning (ICL), using demonstrations to guide the LLM in conducting target tasks, while crafting and selecting proper demonstrations is still an open issue. Many studies [53], [54] pointed out that the content, number and order of demonstration can all influence the results. Therefore, we define a series of atomic operations to measure the complexity of a query question to the given table and describe the steps with the defined atomic operations, which can be a metric for the demonstration selection when crafting prompts. To mitigate the issues limiting the performance of LLMs, including the complex table structure, heterogeneous data types, huge tables, and the inherent limitations of LLMs in reasoning capacities, we propose a three-stage prompting solution containing task-

planning, task-conducting and task-correction stages, as shown in Figure 3. The task-planning stage prompts the LLM to analyze the statistical information of the given table and provide programming steps, data requirements, and relevant columns to solve the query question and generate a plan. Then, the task-conducting stage first generates a default answer as the final answer when the generated Python code fails to execute and then generates the Python code based on the plan from the first stage. When a task-conducting stage fails to execute, the heterogeneous data types are usually the reason. Therefore, we also include a task-correction stage, which can generate normalization functions to normalize the data and correct the error in the task-conducting stage. It is worth mentioning that our proposed method can avoid huge tables as a part of the prompt, which can drastically reduce the number of prompting tokens when the tables are huge.

### B. Proposed Prompting Method for Table-QA

1) Overall Workflow: As discussed in Sections I and IV-A, we use the challenging Table-QA task as the case study of the proposed LLM-based GIoT system and propose to prompt open-source LLMs to generate Python code for the semi-structured Table-QA problem to mitigate the issues caused by complex table structures, heterogeneous data types, huge tables, and limitations of LLMs. Specifically, we propose a three-stage prompting method, including task-planning, task-conducting and task-correction stages. The proposed workflow is shown in Figure 5, in which the question is *the total number of points scored by the tide in the last 3 games combined*. To answer this question, the statistics table of the given table contains information regarding the column names, data types,

| Club              | Season       | League |       | Cup <sup>1</sup> |       | Continental <sup>2</sup> |       | Other <sup>3</sup> |       | Total |       |
|-------------------|--------------|--------|-------|------------------|-------|--------------------------|-------|--------------------|-------|-------|-------|
|                   |              | Apps   | Goals | Apps             | Goals | Apps                     | Goals | Apps               | Goals | Apps  | Goals |
| Anzhi Makhachkala | 1999         | 40     | 2     |                  |       | -                        |       | -                  |       | 40    | 2     |
|                   | 2000         | 30     | 3     |                  |       | -                        |       | -                  |       | 30    | 3     |
|                   | 2001         | 14     | 0     |                  |       | -                        |       | -                  |       | 14    | 0     |
|                   | Total        | 84     | 5     |                  |       |                          |       |                    |       | 84    | 5     |
| CSKA Moscow       | 2001         | 12     | 1     |                  |       | -                        |       | -                  |       | 12    | 1     |
|                   | 2002         | 29     | 2     |                  |       | 2                        | 0     | -                  |       | 31    | 2     |
|                   | 2003         | 28     | 1     |                  |       | 2                        | 0     | 1                  | 0     | 31    | 1     |
|                   | 2004         | 26     | 1     |                  |       | 10                       | 0     | 1                  | 0     | 27    | 1     |
|                   | 2005         | 30     | 1     | 0                | 0     | 15                       | 0     | -                  |       | 45    | 1     |
|                   | 2006         | 30     | 1     | 6                | 0     | 7                        | 0     | 1                  | 0     | 44    | 1     |
|                   | 2007         | 27     | 0     | 5                | 0     | 7                        | 0     | 1                  | 0     | 40    | 0     |
|                   | 2008         | 23     | 0     | 3                | 0     | -                        |       | -                  |       | 18    | 0     |
|                   | 2009         | 10     | 0     | 1                | 0     | 6                        | 0     | 1                  | 0     | 18    | 0     |
|                   | 2010         | 11     | 0     | 1                | 0     | 6                        | 0     | 0                  | 0     | 18    | 0     |
|                   | 2011–12      | 13     | 0     | 1                | 0     | 1                        | 0     | 0                  | 0     | 15    | 0     |
|                   | 2012–13      | 0      | 0     | 1                | 0     | 0                        | 0     | -                  |       | 1     | 0     |
|                   | 2013–14      | 1      | 0     | 0                | 0     | 0                        | 0     | -                  |       | 1     | 0     |
|                   | Total        | 240    | 7     | 1                | 0     | 56                       | 0     | 5                  | 0     | 302   | 7     |
|                   | Career total | 324    | 12    | 1                | 0     | 56                       | 0     | 5                  | 0     | 386   | 12    |

(a) A sample table with a complex structure from the WikiTableQA dataset.

#solution in Python

```
import pandas as pd
def solution(table_dict):
    df = pd.DataFrame(table_dict)
    # Filter the rows after 2004
    df_after_2004 = df[df['Season'] > 2004]
    # Find the maximum League Apps
    max_league_apps = df_after_2004['League Apps'].max()
    return max_league_apps
```

(b) A sample of generated Python code to answer the question *What is the maximum League Apps after 2004?*.

Fig. 4: An example of a semi-structured table and its failed Python code because of the heterogeneous data types and table's complex structure. The Python code is generated by Mixtral-8x7B.

and the first and last entries, which are first generated as a part of the task-planning prompt, together with instructions, demonstrations, and questions, as shown in Figure 6. Then, the task-planning prompt is fed into the open-source LLM to make the reasoning plan, including Relevant Columns, Operations, and Programming Steps. It is worth mentioning that the statistics table can be far more compact than the original table when the original table is huge, which can reduce the number of prompt tokens. With the Relevant Columns, Operations and Programming Steps from the results of the task-planning step, the Relevant Columns are used to extract the contents of these columns, the Operations are used to select the demonstrations, and the Programming Steps are parts of the instructions to guide the reasoning function generation. With these processed results, the task-conducting prompt is constructed and fed into the open-source LLM to generate the reasoning Python function and a default answer, in which the default is treated as the final answer if the code fails to run. Notably, since we use the Pandas library in our solution, the original table needs to be represented as a dictionary of the list and fed into the generated function as a parameter. As running the generated Python code is almost cost-free compared with LLM inference, and some tables do not need to be normalized, we try to execute the reasoning function first. If there is no error from the execution, then the result of the

execution should be the final result. However, if the execution fails, we construct the task-correction prompt containing the content of Relevant columns and reasoning code to generate the normalization function, apply the normalization to the original table, and then feed the normalized original table as the parameter to the reasoning function to run it again. For the example in Figure 5, we need to extract the string "W 21-14", "L 23-24" and "W 24-17" first and then extract the scores "21", "23", "24" and convert them into integers, which beyond the capacities of the LLM. Therefore, the first run of the reasoning function would fail, even though its reasoning logic is correct. While the normalization function can correctly extract and convert the points from the string to integers, the second run of the reasoning function should be successful.

2) *Demonstration Crafting and Selection*: As shown in Figure 5, three prompts need to be constructed to answer a question; the demonstrations used in these three prompts are critical for the LLM's performance, especially for the task-conducting stage to generate the reasoning function. Therefore, we defined a series of atomic operations to describe the reasoning logic to answer questions, as shown in Table I. For the task-planning step, we craft a question-and-answer pair for each type of operation and instruct the LLM to generate the Relevant Columns, Operations and Programming Steps, as shown in Figure 6. For the task-conducting prompts, we construct two question and reasoning function pairs for each defined operation and use operation as the metric to select these pairs. Even though we prompt the LLM to select one defined operation as output, the LLM can generate results without following the instructions. Therefore, when the Operations cannot be matched, the default setting contains all the question and function pairs. Figure 7 shows an example of COUNT operation, which includes a Meta Information table, Column Details and two questions with their Python solutions and default answers. It is worth mentioning that the demonstrations and prompts discussed in this section are stored in the Task-specific Prompts Database, and using operations as metrics to select proper demonstrations is the function of the proposed Prompt Management Module, as shown in Figures 1 and 2.

## V. EXPERIMENTS AND ANALYSIS

### A. Datasets and Experimental Settings

We evaluate our proposed prompting solution on WikiTableQA [47] and TabFact [55] datasets, which are two datasets created by Wiki-tables without text context. WikiTableQA mainly contains compositional questions, such as questions requiring counting and ranking table contents. TabFact is a Fact Verification dataset, which can be treated as a special setting of a typical Table-QA problem whose answer set is {True, False}. Since the proposed solution of this study is based on ICL, which is a few-shot learning setting without any training stage, we only use the test set of these two datasets to evaluate the performance, which contains 4344 and 12828 QA pairs, respectively. TabFact dataset further categorizes the test set into simple and complex groups, which include 4219 and 8609 QA pairs, respectively. Considering the large

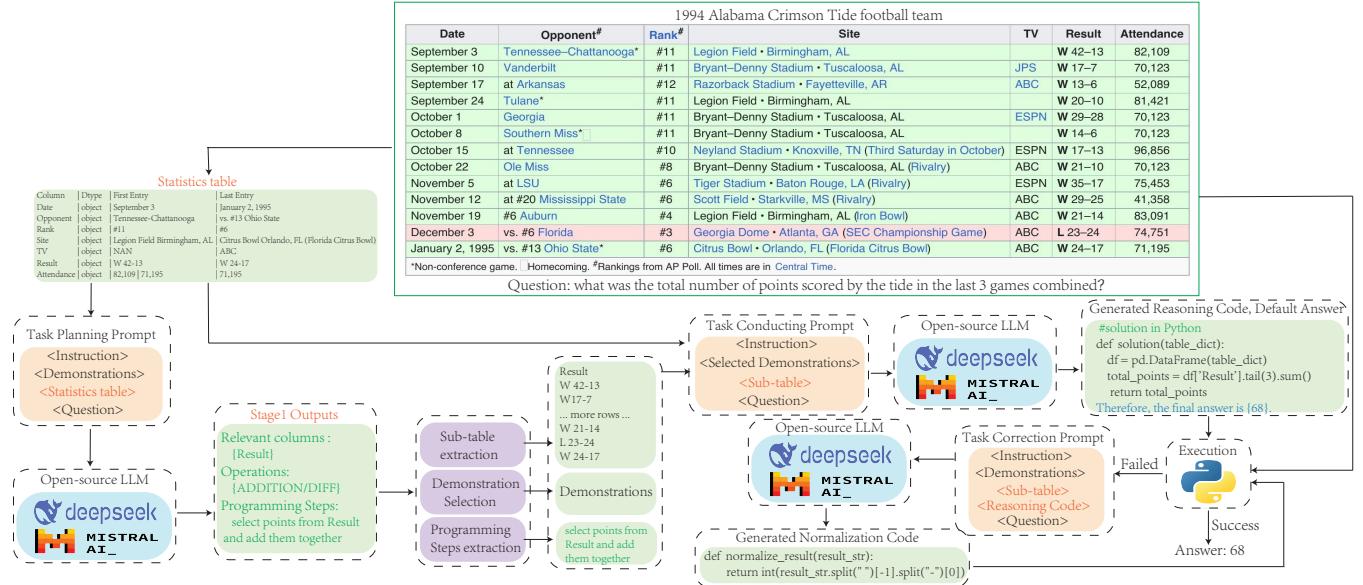


Fig. 5: The workflow of the proposed prompting solution. Notably, the question and the table are from the request of an IoT device. The Python interpreter is in the Post-processing Module, and the stages of selecting demonstrations and creating prompts are in the Prompt Management Module.

TABLE I: Defined operations for Demonstration selection for code generation.

| Operation Type | Operation Name | Description                                       |
|----------------|----------------|---|
| Reasoning      | SelectTable    | select a cell from the table based on a criteria  |
|                | ADDITION/DIFF  | addition or subtraction                           |
|                | TIMES/DIVISION | production or quotient of two numbers             |
|                | AVG            | average of several numbers                        |
|                | COUNT          | count the number based on a criteria              |
|                | MAX/MIN        | select the maximum/minimum one from given numbers |

You defined 5 types of operations to answer the questions based on the given table. The operations are defined as follows:

1. SelectTable: select the content from the given table based on some criteria
2. ADDITION/DIFF: addition or subtraction
3. AVG: average of several numbers
4. COUNT: count the number of spans
5. MAX/MIN: select the maximum/minimum one from given numbers

To answer the questions, one of these operations might be needed. You are going to analyze two tables. Select one operation from defined operations as needed Operation and also give Programming Steps and Relevant Columns to the questions.

Read the table below regarding "2008 Clasica de San Sebastian" to give the needed operations, reasoning steps and relevant columns to the following questions.

Column | Dtype | First Entry | Last Entry  
Rank | int64 | 1 | 10  
Cyclist | object | Alejandro Valverde (ESP) | David Moncoutié (FRA)

Team | object | Caisse d' Epargne | Cofidis  
Time | object | 5h 29' 10" | 10' + 2

UCI ProTour Points | int64 | 40 | 1

Question: who ranked at the 7th place?

Programming Steps: Column Rank contains the ranking information, column Cyclist contains the names of cyclists. To answer this question, select cyclist from column Cyclist ranked at the 7th place based on column Rank.

Operations: SelectTable.

Relevant Columns: {Rank, Cyclist}.

...More Lines...

Question: What is the average points of all cyclists listed in the table?

Programming Steps: Columns UCI ProTour Points contains the points information. To answer this question, calculate the mean value of all points from UCI ProTour Points

Operations: AVG.

Relevant Columns: {UCI ProTour Points}.

Read the table blow regarding " <title> " + " to answer the following one question:

<Statistics Table>

<Question>

Fig. 6: The task-planning prompt. The defined operations and the statistics table are highlighted with green and yellow. <title>, <Statistics Table> and <Question> are from the table to be analyzed.

size of TabFact dataset, some studies [31], [35] conducted experiments on a small subset of TabFact, which contains 1,005 simple and 1,019 complex QA pairs. To compare with these studies, we also report the results on this small test subset of TabFact.

TABLE II: Experimental results on WikiTableQA dataset with Exact Match Accuracy as metric.

| LLM          | Method  | EM Acc       |
|--------------|---------|--------------|
| Codex        | Binder  | 61.90        |
|              | Dater   | 65.90        |
|              | Direct  | 53.08        |
|              | CoT     | 53.48        |
| Mixtral-8x7B | PoT     | 40.40        |
|              | Tab-PoT | 63.33        |
|              | Direct  | 27.19        |
|              | CoT     | 30.46        |
| Mistral-7B   | PoT     | 27.66        |
|              | Tab-PoT | 52.12        |
|              | Direct  | 54.72        |
|              | CoT     | 55.57        |
| DeepSeek-67B | PoT     | 43.92        |
|              | Tab-PoT | <b>66.78</b> |
|              | Direct  | 33.86        |
|              | CoT     | 34.65        |
| DeepSeek-7B  | PoT     | 19.61        |
|              | Tab-PoT | 40.03        |

As the proposed LLM-based GIoT system is designed to be deployed in a local network, we use open-source

You have the following 2 tables' meta information and the detailed content of some columns. Answer the questions based on the 2 tables: Read the first table's meta information below regarding 2008 Clasica de San Sebastian to answer the following questions with Python codes.

**Meta Information:**  
 Column | Dtype | First Entry | Last Entry  
 Rank | int64 | 1 | 10  
 Cyclist | object | Alejandro Valverde (ESP) | David Moncoutie (FRA)  
 Team | object | Caisse d'Epargne | Cofidis  
 Time | object | 5h 29' 10" | +2  
 UCI ProTour Points | int64 | 40 | 1

**Column Details:**  
 Rank | Cyclist | Team | Time | UCI ProTour Points  
 1 | Alejandro Valverde (ESP) | Caisse d'Epargne | 5h 29' 10" | 40  
 2 | Alexander Kolobnev (RUS) | Team CSC-Saxo Bank | s.t. | 30  
 3 | Davide Rebellin (ITA) | Gerolsteiner | s.t. | 25  
 4 | Paolo Bettini (ITA) | Quick Step | s.t. | 20  
 5 | Franco Pellizotti (ITA) | Liquigas | s.t. | 15  
 6 | Denis Menchov (RUS) | Rabobank | s.t. | 11  
 7 | Samuel Sanchez (ESP) | Euskaltel-Euskadi | s.t. | 7  
 8 | Stephane Goubaert (FRA) | Ag2r-La Mondiale | + 2 | 5  
 9 | Haimar Zubeldia (ESP) | Euskaltel-Euskadi | + 2 | 3  
 10 | David Moncoutie (FRA) | Cofidis | + 2 | 1

Question: how many players got less than 10 points?

Answer:

```
#solution in Python
def solution(table_dict):
    import pandas as pd
    df = pd.DataFrame(table_dict)
    #count the number of points less than 10 based on column UCI ProTour Points
    less_than_10_points = df[df["UCI ProTour Points"] < 10].shape[0]
    return less_than_10_points
Therefore, the final answer is [4].

```

Question: how many cyclists are from Italy?

Answer:

```
#solution in Python
def solution(table_dict):
    import pandas as pd
    df = pd.DataFrame(table_dict)
    #count the number of cyclists who are from Italy
    num_italy_cyclist = df[df["UCI ProTour Points"].str.contains("ITA")].shape[0]
    return num_italy_cyclist
Therefore, the final answer is [3].

```

Read the second table's meta information below regarding "<title>" + "<code>" + " to answer the following one question with Python code.

<Meta Information>  
<Column Details>  
<Question>

Fig. 7: The task-conducting prompt. The Meta Information and Column Details are highlighted with yellow, and the default answers are highlighted with blue. <title>, <Meta Information>, <Column Details> and <Question> are from the table to be analyzed.

LLMs, including Mixtral-8x7B<sup>1</sup>, Mistral-7B<sup>2</sup>, DeepSeek-67B<sup>3</sup> and DeepSeek-7B<sup>4</sup> to conduct our experiments. Since the proposed solution in this study is a prompt engineering method, we include Direct Prompting, CoT [14], PoT [13], Binder [35] and Dater [31] as benchmarks, in which Binder and Dater are two solutions leveraging SQL. For the implementation of benchmark methods, we use the implementation of TableCoT<sup>5</sup> [30] for the Direct Prompting and CoT. We reimplemented the PoT method following the example prompts reported in PoT [13]. The results of Dater and Binder are directly from study [31]. We use beam search decoding for the experiments. At last, even though we employ ICL to provide demonstrations to guide the LLM output of the final results in a "}" sometimes LLMs can fail to follow this output format. Therefore, we use a simple answer alignment step to post-process the results with incorrect formats. More specifically, we employ a direct prompting method by providing a few demonstrations containing the question, answer and formatted answer following TableCoT [30]. It is worth mentioning that we use the default precision of parameters, namely bfloat16, for the LLMs in this section.

<sup>1</sup><https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1>

<sup>2</sup><https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2>

<sup>3</sup><https://huggingface.co/deepseek-ai/deepseek-llm-67b-chat>

<sup>4</sup><https://huggingface.co/deepseek-ai/deepseek-llm-7b-chat>

<sup>5</sup><https://github.com/wenhuchen/TableCoT>

We term our proposed prompting solution as Tab-PoT, and the experimental results are shown in Table II and Table III. The experimental results show that our proposed prompting solution can perform competitively compared with state-of-the-art methods. The Tab-PoT with DeepSeek-67B can achieve state-of-the-art performance, and the Tab-PoT with both Mixtral-8x7B and DeepSeek-67B can improve the original PoT method by at least 22% on the WikiTableQA dataset. Besides, applying LLMs with a larger number of parameters can significantly improve performance. The CoT does not show many benefits in improving performance compared with direct prompting on the WikiTableQA dataset while consistently improving the performance on the TabFact dataset.

## B. Discussion and Analysis

**1) Ablation Study:** As discussed in Section IV, our proposed prompting solution consists of three stages: task-planning, task-conducting and task-correction. The task-planning stage can output the relevant columns and reasoning steps to answer the query question, which can be treated as a step of table decomposition and question decomposition, which can also be applied to the conventional PoT. In the task-conducting stage, we prompt the LLM to generate the Python code and a default answer. The default answer is the final answer when the Python code fails to run even after the task-correction stage. Since the default answer is generated after the Python code, it can be treated as an implicit CoT where the Python code is the reasoning rationales in the CoT. Finally, the third stage generates normalization functions to correct the errors in the Python code caused by the heterogeneous data types, which rely on the relevant columns generated by the first stage. Therefore, in this section, we conduct four ablation experiments by applying task-planning, default answer in task-conducting, task-planning and task-correction, and task-planning and default answer in task-conducting to the conventional PoT. The experimental results are shown in Table IV, where Plan, Correction and Default represent applying task-planning, task-correction and the default answer in task-conducting stages. We use Mixtral-8x7B as the LLM, and the experimental results demonstrate that each of the proposed three components can significantly improve the PoT baseline.

**2) The impact of quantization:** Since the LLMs usually have very high hardware requirements for inference, quantization methods are widely used to compact LLMs using lower precision parameters. In this section, we conduct experiments to compare the performance of quantization versions of LLMs. Specifically, similar to the previous section, we also use Mixtral-8x7B to conduct experiments on the WikiTableQA dataset and compare its 16-bit, 8-bit and 4-bit versions of applying the proposed Tab-PoT solution and the experimental results are shown in Table V. For our proposed Tab-PoT, even though applying quantization methods can lead to worse performance, the performance of 8-bit and 4-bit versions is still competitive. It is worth mentioning that the 4-bit version shares a similar RAM footprint with the Mistral-7B model but

TABLE III: Experimental results on TabFact dataset with Exact Match Accuracy as metric. *full* and *small* mean the full and small versions of TabFact dataset.

| LLM          | Method  | Simple <sub>full</sub> | Complex <sub>full</sub> | All <sub>full</sub> | Simple <sub>small</sub> | Complex <sub>small</sub> | All <sub>small</sub> |
|--------------|---------|------------------------|-------------------------|---------------------|-------------------------|--------------------------|----------------------|
| Codex        | Binder  | -                      | -                       | -                   | -                       | -                        | 85.10                |
|              | Dater   | -                      | -                       | -                   | 91.20                   | 80.00                    | 85.60                |
| Mixtral-8x7B | Direct  | 80.59                  | 69.98                   | 73.47               | 81.29                   | 70.56                    | 75.89                |
|              | CoT     | 83.53                  | 74.94                   | 77.77               | 86.07                   | 74.19                    | 80.09                |
|              | PoT     | 73.33                  | 69.07                   | 70.47               | 76.02                   | 70.66                    | 73.32                |
|              | Tab-PoT | 86.49                  | 76.06                   | 79.49               | 88.36                   | 75.07                    | 81.67                |
| Mixtral-7B   | Direct  | 73.57                  | 64.83                   | 67.70               | 73.13                   | 62.71                    | 67.89                |
|              | CoT     | 73.31                  | 67.52                   | 69.43               | 73.73                   | 67.12                    | 70.41                |
|              | PoT     | 66.11                  | 63.58                   | 64.41               | 65.97                   | 66.54                    | 66.25                |
|              | Tab-PoT | 77.48                  | 66.83                   | 69.75               | 76.82                   | 66.93                    | 71.84                |
| DeepSeek-67B | Direct  | 84.36                  | 74.19                   | 77.53               | 84.68                   | 72.62                    | 78.61                |
|              | CoT     | 87.44                  | 78.00                   | 81.10               | 88.46                   | 76.84                    | 82.61                |
|              | PoT     | 74.43                  | 71.79                   | 72.65               | 76.32                   | 72.72                    | 74.51                |
|              | Tab-PoT | <b>90.09</b>           | <b>78.91</b>            | <b>82.58</b>        | <b>91.34</b>            | <b>80.27</b>             | <b>85.77</b>         |
| DeepSeek-7B  | Direct  | 59.16                  | 55.42                   | 56.65               | 59.50                   | 55.94                    | 57.71                |
|              | CoT     | 69.31                  | 61.18                   | 63.85               | 70.65                   | 59.76                    | 65.17                |
|              | PoT     | 63.71                  | 58.26                   | 60.06               | 64.88                   | 58.98                    | 61.91                |
|              | Tab-PoT | 70.42                  | 62.13                   | 64.86               | 70.75                   | 61.04                    | 65.86                |

TABLE IV: Ablation study results on the WikiTableQA dataset with Exact Match Accuracy as metric.

| Model      | Plan | Correction | Default | EM Acc |
|------------|------|------------|---------|--------|
| PoT        |      |            |         | 40.40  |
| Ablation 1 | ✓    |            |         | 46.52  |
| Ablation 2 |      |            | ✓       | 53.66  |
| Ablation 3 | ✓    | ✓          |         | 53.31  |
| Ablation 4 | ✓    |            | ✓       | 58.43  |
| Tab-PoT    | ✓    | ✓          | ✓       | 63.33  |

achieves much higher performance, as shown in Table V and Table II.

TABLE V: The impact of LLM quantization methods.

| LLM          | Parameter Precision | EM Acc |
|--------------|---------------------|--------|
|              | 16-bit              | 63.33  |
| Mixtral-8x7B | 8-bit               | 62.20  |
|              | 4-bit               | 60.52  |

3) *Analysis on different implementations of PoT*: Since Python is a flexible programming language that can implement a function with multiple implementations. In this section, we discuss the differences among these different types of implementations. More specifically, one straightforward implementation is using the Python Standard Library and following the extraction and reasoning steps, as shown in Figure 8. This method selects relevant data from the table, defines the data with a List of Tuples, and then conducts reasoning over the defined List of Tuples. One obvious drawback of this implementation method is that it needs to repeat the relevant columns in the Python code, which can be very large when the table contains a large number of rows, leading to more inference time. Therefore, a refined method can use the table as the parameter of the solution function, then as shown in Figure 9. At last, since Pandas is a widely used Python library to process tabular data, we can also use Pandas to finish the reasoning tasks with a table dictionary as the input, as shown in Figure 10. We conduct experiments on the WikiTableQA dataset to compare the performance of these three types of implementations. Even though the implementation of applying Python Standard Library can show some benefits regarding the

EM Accuracy, it requires more Prompt Tokens and Completion Tokens, as shown in Table VI, because this implementation needs to extract relevant from the table directly and define them as a Python dictionary, List or variables. On the other hand, both solutions introducing function parameters can reduce the number of prompting tokens and completion tokens, and applying Pandas Library can achieve better performance than using the standard library.

TABLE VI: Ablation study results on the WikiTableQA dataset with Exact Match Accuracy as metric.

| Method      | EM Acc | #AVG Prompt Tokens | #AVG Completion Tokens |
|-------------|--------|--------------------|------------------------|
| STDLib      | 44.96  | 2365               | 732                    |
| STDLib-Para | 31.17  | 2157               | 114                    |
| Pandas      | 40.40  | 2241               | 88                     |

4) *Analysis on inference cost*: Since the inference cost is highly correlated with the number of tokens, we use the prompt tokens and generated tokens as the metrics to measure the inference cost in this section. Therefore, we calculate the Average Prompt Tokens and Completion Tokens on the WikiTableQA dataset. As shown in Table VII, the proposed Tab-PoT can introduce some overhead compared with other methods regarding the average prompt tokens and average completion tokens on the WikiTableQA dataset. Since the proposed Tab-PoT contains three stages at most, each including instructions and demonstrations, it can reduce the number of Prompt Tokens only when the input table is huge. We group the number of table tokens into 15 bins and plot the relation between the number of table tokens and the prompt tokens on the WikiTableQA dataset. When the number of a table's tokens is larger than around 1867, our proposed Tab\_PoT can use fewer prompt tokens than PoT, which means fewer computation operations and less inference time, as shown in Figure 11. Since the WikiTableQA dataset contains a large portion of tables whose number of tokens is smaller than 1158, the average prompt tokens of the proposed Tab-PoT is still larger than the one of PoT overall, as shown in Table VII. As pointed out by some studies [30], the LLM can perform

Answer the questions based on the two given tables:

| Rank | Cyclist                  | Team               | Time      | UCI ProTour Points |
|------|--------------------------|--------------------|-----------|--------------------|
| 1    | Alejandro Valverde (ESP) | Caisse d'Epargne   | 5h 29' 10 | 40                 |
| 2    | Alexandr Kolobnev (RUS)  | Team CSC Saxo Bank | s.t.      | 30                 |
| 3    | Davide Rebellin (ITA)    | Gerolsteiner       | s.t.      | 25                 |
| 4    | Paolo Bettini (ITA)      | Quick Step         | s.t.      | 20                 |
| 5    | Franco Pellizotti (ITA)  | Liquigas           | s.t.      | 15                 |
| 6    | Denis Menchov (RUS)      | Rabobank           | s.t.      | 11                 |
| 7    | Samuel Sanchez (ESP)     | Euskaltel-Euskadi  | s.t.      | 7                  |
| 8    | Stephane Goubert (FRA)   | Ag2r-La Mondiale   | + 2       | 5                  |
| 9    | Haimar Zubeldia (ESP)    | Euskaltel-Euskadi  | + 2       | 3                  |
| 10   | David Moncoutie (FRA)    | Cofidis            | + 2       | 1                  |

Question: who is the cyclist after Alexandr Kolobnev in the table?

Answer:

```
def solution():
    # Define the table as a list of cyclist names
    table = [
        "Alejandro Valverde (ESP)",
        "Alexandr Kolobnev (RUS)",
        "Davide Rebellin (ITA)",
        "Paolo Bettini (ITA)",
        "Franco Pellizotti (ITA)",
        "Denis Menchov (RUS)",
        "Samuel Sanchez (ESP)",
        "Stephane Goubert (FRA)",
        "Haimar Zubeldia (ESP)",
        "David Moncoutie (FRA)"]
    # Find the cyclist after Alexandr Kolobnev
    kolobnev_index = table.index("Alexandr Kolobnev (RUS)")
    if kolobnev_index != -1 and kolobnev_index < len(table) - 1:
        next_cyclist = table[kolobnev_index + 1]
    return next_cyclist
else:
    return "Alexandr Kolobnev is either not in the table or is the last entry."
```

Therefore, the final answer with python code is {print(solution())}

Question: which country had the most cyclists finish within the top 10?

Answer:

```
def solution():
    # Define the table as a list of tuples containing cyclist names and their countries
    table = [
        ("Alejandro Valverde (ESP)", "ESP"),
        ("Alexandr Kolobnev (RUS)", "RUS"),
        ("Davide Rebellin (ITA)", "ITA"),
        ("Paolo Bettini (ITA)", "ITA"),
        ("Franco Pellizotti (ITA)", "ITA"),
        ("Denis Menchov (RUS)", "RUS"),
        ("Samuel Sanchez (ESP)", "ESP"),
        ("Stephane Goubert (FRA)", "FRA"),
        ("Haimar Zubeldia (ESP)", "ESP"),
        ("David Moncoutie (FRA)", "FRA")]
    # Count the occurrences of each country in the top 10
```

```
country_count = {}
for _, country in table:
    if country in country_count:
        country_count[country] += 1
    else:
        country_count[country] = 1
```

```
# Find the country with the most cyclists in the top 10
max_country = max(country_count, key=country_count.get)
return max_country
```

Therefore, the final answer with python code is {print(solution())}

...More Lines...

Read the table blow regarding " + <title> + " to answer the following one question:

<Full Table>

<Question>

Fig. 8: PoT implementation of applying Python Standard Library. Some lines are omitted due to the limited page.

well on small tables, meaning that we can easily extend the proposed Tab-PoT with other methods, such as CoT, by applying a threshold regarding the size of the input table, to reduce the number of prompt tokens and maintain the performance simultaneously. It is worth mentioning that the price of prompt tokens and completion tokens are different when using commercial LLMs, such as GPT-4 [8].

TABLE VII: Comparisons of Average Prompt Tokens and Completion Tokens.

| Method  | #AVG Prompt Tokens | #AVG Completion Tokens |
|---------|--------------------|------------------------|
| Direct  | 1405               | 10                     |
| CoT     | 1599               | 44                     |
| PoT     | 2241               | 88                     |
| Tab-PoT | 2685               | 192                    |

Answer the questions based on the two given tables:

| Rank | Cyclist                  | Team               | Time      | UCI ProTour Points |
|------|--------------------------|--------------------|-----------|--------------------|
| 1    | Alejandro Valverde (ESP) | Caisse d'Epargne   | 5h 29' 10 | 40                 |
| 2    | Alexandr Kolobnev (RUS)  | Team CSC Saxo Bank | s.t.      | 30                 |
| 3    | Davide Rebellin (ITA)    | Gerolsteiner       | s.t.      | 25                 |
| 4    | Paolo Bettini (ITA)      | Quick Step         | s.t.      | 20                 |
| 5    | Franco Pellizotti (ITA)  | Liquigas           | s.t.      | 15                 |
| 6    | Denis Menchov (RUS)      | Rabobank           | s.t.      | 11                 |
| 7    | Samuel Sanchez (ESP)     | Euskaltel-Euskadi  | s.t.      | 7                  |
| 8    | Stephane Goubert (FRA)   | Ag2r-La Mondiale   | + 2       | 5                  |
| 9    | Haimar Zubeldia (ESP)    | Euskaltel-Euskadi  | + 2       | 3                  |
| 10   | David Moncoutie (FRA)    | Cofidis            | + 2       | 1                  |

Question: who is the cyclist after Alexandr Kolobnev in the table?

Answer:

```
def solution(table, header_dict):
    # Define the sub_table as a list of cyclist names
    sub_table = [row[header_dict["Cyclist"]] for row in table]
    # Find the cyclist after Alexandr Kolobnev
    kolobnev_index = sub_table.index("Alexandr Kolobnev (RUS)")
    if kolobnev_index != -1 and kolobnev_index < len(table) - 1:
        next_cyclist = sub_table[kolobnev_index + 1]
    return next_cyclist
else:
    return "Alexandr Kolobnev is either not in the table or is the last entry."
```

Therefore, the final answer with python code is {print(solution(table, header\_dict))}

Question: which country had the most cyclists finish within the top 10?

Answer:

```
def solution(table, header_dict):
    # Define the sub_table as a list of cyclist names
    sub_table = [row[header_dict["Cyclist"]] for row in table]
    # Count the occurrences of each country in the top 10
    country_count = {}
```

```
for cyclist_info in sub_table:
    country_start_index = cyclist_info.find("(") + 1
    country_end_index = cyclist_info.find(")")
    country = cyclist_info[country_start_index:country_end_index]
```

```
if country in country_count:
    country_count[country] += 1
else:
    country_count[country] = 1
```

```
# Find the country with the most cyclists in the top 10
max_country = max(country_count, key=country_count.get)
return max_country
```

Therefore, the final answer with python code is {print(solution(table, header\_dict))}

...More Lines...

Read the table blow regarding " + <title> + " to answer the following one question:

<Full Table>

<Question>

Fig. 9: PoT implementation of applying Python Standard Library with parameters. Some lines are omitted due to the limited page.

## VI. CONCLUSION AND FUTURE WORK

In this study, we propose a LLM-based GIoT system, which can be deployed in a local network setting to address the security concerns of many scenarios. To demonstrate the proposed LLM-based GIoT system, we use a challenging semi-structured Table-QA problem as a case study and propose a three-stage prompting solution to alleviate the issues caused by complex structures, heterogeneous data types, huge tables, and LLMs' limitations. The proposed prompting solution uses a statistics table in the first stage and sub-tables in the following stages, which can reduce the inference cost and improve the performance when the original table is huge. We define a series of atomic operations to guide the demonstration crafting and selection, which can reduce reasoning errors. Besides, we use the task-correction stage to correct the failure code caused by the heterogeneous data types and use a default answer as the final answer when the generated Python code fails to run even after the task-correction step, which can be caused by the complex structure or the limitations of the LLM. As demonstrated in Section V, designing tailored prompting methods can improve the performance of open-source LLMs, achieving state-of-the-art performance, and the proposed LLM-based GIoT system can be easily extended by adding task-specific prompt instructions and demonstrations to

Answer the questions based on the two given tables:

| Rank | Cyclist                  | Team               | Time        | UCI ProTour Points |
|------|--------------------------|--------------------|-------------|--------------------|
| 1    | Alejandro Valverde (ESP) | Caisse d'Epargne   | [5h 29' 10] | 40                 |
| 2    | Alexandr Kolobnev (RUS)  | Team CSC Saxo Bank | [s.t.]      | 30                 |
| 3    | Davide Rebellin (ITA)    | Gerolsteiner       | [s.t.]      | 25                 |
| 4    | Paolo Bettini (ITA)      | Quick Step         | [s.t.]      | 20                 |
| 5    | Franco Pellizzetti (ITA) | Liquigas           | [s.t.]      | 15                 |
| 6    | Denis Menchov (RUS)      | Rabobank           | [s.t.]      | 11                 |
| 7    | Samuel Sanchez (ESP)     | Euskaltel-Euskadi  | [s.t.]      | 7                  |
| 8    | Stephane Goubert (FRA)   | Ag2r-La Mondiale   | [+ 2]       | 5                  |
| 9    | Haimar Zubeldia (ESP)    | Euskaltel-Euskadi  | [+ 2]       | 3                  |
| 10   | David Moncoutié (FRA)    | Cofidis            | [+ 2]       | 1                  |

Question: who is the cyclist after Alexandr Kolobnev in the table?

Answer:

```
#solution in Python
import pandas as pd
def solution(table_dict):
    df = pd.DataFrame(table_dict)
    # Find the index of Alexandr Kolobnev in the DataFrame
    kolobnev_index = df[df["Cyclist"].str.contains("Alexandr Kolobnev")][0]
    # Extract the cyclist after Alexandr Kolobnev
    next_cyclist = df.iloc[kolobnev_index + 1]["Cyclist"]
    return next_cyclist
```

Therefore, the final answer with python code is {print(solution(table\_dict))}

Question: which country had the most cyclists finish within the top 10?

Answer:

```
#solution in Python:
import pandas as pd
def solution(table_dict):
    df = pd.DataFrame(table_dict)
    #since country in the column Cyclist, extract the country code from the column Cyclist
    df["Country"] = df["Cyclist"].apply(normalize_cyclist)
    # Count the number of cyclists from each country in the top 10
    country_counts = df["Country"].value_counts()
    most_cyclists_country = country_counts.idxmax()
    return most_cyclists_country
#since country can not be obtained directly and is in "(" in column Cyclist, define the function which
#can extract country from the Cyclist
def normalize_cyclist(cyclist_info):
    import re
    # Regex pattern to extract content in ()
    pattern = r"\((.*?))"
    # Find the substring within parentheses
    match = re.search(pattern, cyclist_info)
    if match:
        return match.group(1)
    else:
        return ""
Therefore, the final answer with python code is {print(solution(table_dict))}
```

Read the table blow regarding '`+ <title> +`' to answer the following one question:

<Full Table>  
<Question>

Fig. 10: PoT implementation of applying Python Pandas Library. Some lines are omitted due to the limited page.

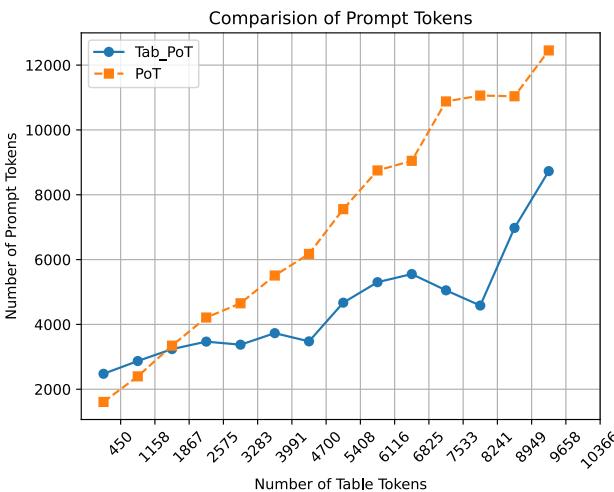


Fig. 11: Comparison of Prompting Tokens between PoT and Tab\_PoT.

the system. Besides, as the proposed GIoT system is designed to deploy in a edge server, applying quantization to the LLM can be a good option to reduce the hardware requirements, as discussed in Section V-B2. In this study, we focus on text data, while IoT devices can generate data in multiple data types, such as time series and images. Therefore, extending the current system to handle data in various modalities can be a further direction.

## ACKNOWLEDGEMENT

This work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under the CREATE TRAVERSAL and DISCOVERY programs.

## REFERENCES

- [1] S. Hu, M. Li, J. Gao, C. Zhou, and X. S. Shen, "Adaptive device-edge collaboration on dnn inference in aiot: A digital twin-assisted approach," *IEEE Internet of Things Journal*, 2023.
- [2] M. Adil, M. K. Khan, N. Kumar, M. Attique, A. Farouk, M. Guizani, and Z. Jin, "Healthcare internet of things: Security threats, challenges and future research directions," *IEEE Internet of Things Journal*, 2024.
- [3] J. Fan, W. Yang, Z. Liu, J. Kang, D. Niyato, K.-Y. Lam, and H. Du, "Understanding security in smart city domains from the ant-centric perspective," *IEEE Internet of Things Journal*, 2023.
- [4] J. Franco, A. Aris, B. Canberk, and A. S. Uluagac, "A survey of honeypots and honeynets for internet of things, industrial internet of things, and cyber-physical systems," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2351–2383, 2021.
- [5] J. Wen, J. Nie, J. Kang, D. Niyato, H. Du, Y. Zhang, and M. Guizani, "From generative ai to generative internet of things: Fundamentals, framework, and outlooks," *IEEE Internet of Things Magazine*, vol. 7, no. 3, pp. 30–37, 2024.
- [6] X. Wang, Z. Wan, A. Hekmati, M. Zong, S. Alam, M. Zhang, and B. Krishnamachari, "Iot in the era of generative ai: Vision and challenges," *arXiv preprint arXiv:2401.01923*, 2024.
- [7] N. Zhong, Y. Wang, R. Xiong, Y. Zheng, Y. Li, M. Ouyang, D. Shen, and X. Zhu, "Casit: Collective intelligent agent system for internet of things," *IEEE Internet of Things Journal*, 2024.
- [8] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257532815>
- [9] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
- [10] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.
- [11] Meta AI, "Meta ai blog: Llama 3," 2024, accessed: 2024-06-10. [Online]. Available: <https://ai.meta.com/blog/meta-llama-3/>
- [12] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10764–10799.
- [13] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *Transactions on Machine Learning Research*, 2023. [Online]. Available: <https://openreview.net/forum?id=YfZ4ZPt8zd>
- [14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [15] T. Tu, Z. He, Z. Zheng, Z. Zheng, J. Jiang, Y. Gong, C. Hu, and D. Cheng, "Towards lifelong unseen task processing with a lightweight unlabeled data schema for aiot," *IEEE Internet of Things Journal*, 2024.
- [16] J. Yin, J. Dong, Y. Wang, C. De Sa, and V. Kuleshov, "Modulora: Finetuning 3-bit llms on consumer gpus by integrating with modular quantizers," *arXiv preprint arXiv:2309.16119*, 2023.
- [17] B. Xiao, M. Simsek, B. Kantarci, and A. A. Alkheir, "Table detection for visually rich document images," *Knowledge-Based Systems*, vol. 282, p. 111080, 2023.

- [18] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. M. Eisenschlos, “Tapas: Weakly supervised table parsing via pre-training,” *arXiv preprint arXiv:2004.02349*, 2020.
- [19] Z. Jiang, Y. Mao, P. He, G. Neubig, and W. Chen, “Omnitab: Pretraining with natural and synthetic data for few-shot table-based question answering,” *arXiv preprint arXiv:2207.03637*, 2022.
- [20] F. Zhu, Z. Liu, F. Feng, C. Wang, M. Li, and T.-S. Chua, “Tat-llm: A specialized language model for discrete reasoning over tabular and textual data,” *arXiv preprint arXiv:2401.13223*, 2024.
- [21] Z. Yu, L. He, Z. Wu, X. Dai, and J. Chen, “Towards better chain-of-thought prompting strategies: A survey,” *arXiv preprint arXiv:2310.04959*, 2023.
- [22] B. Zhao, C. Ji, Y. Zhang, W. He, Y. Wang, Q. Wang, R. Feng, and X. Zhang, “Large language models are complex table parsers,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14786–14802.
- [23] S. De, M. Bermudez-Edo, H. Xu, and Z. Cai, “Deep generative models in the industrial internet of things: a survey,” *IEEE Transactions on Industrial Informatics*, vol. 18, no. 9, pp. 5728–5737, 2022.
- [24] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, “Llmind: Orchestrating ai and iot with llms for complex task execution,” *arXiv preprint arXiv:2312.09007*, 2023.
- [25] B. Rong and H. Rutagemwa, “Leveraging large language models for intelligent control of 6g integrated tn-tnr with iot service,” *IEEE Network*, 2024.
- [26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.
- [27] Y. Chen, S. Qian, H. Tang, X. Lai, Z. Liu, S. Han, and J. Jia, “Longlora: Efficient fine-tuning of long-context large language models,” *arXiv preprint arXiv:2309.12307*, 2023.
- [28] T. Zhang, X. Yue, Y. Li, and H. Sun, “Tablellama: Towards open large generalist models for tables,” *arXiv preprint arXiv:2311.09206*, 2023.
- [29] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [30] W. Chen, “Large language models are few (1)-shot table reasoners,” in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1120–1130.
- [31] Y. Ye, B. Hui, M. Yang, B. Li, F. Huang, and Y. Li, “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*, 2023, pp. 174–184.
- [32] W. Lin, R. Blloshmi, B. Byrne, A. de Gispert, and G. Iglesias, “An inner table retriever for robust table question answering,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 9909–9926.
- [33] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J. rong Wen, “Structgpt: A general framework for large language model to reason over structured data,” in *Conference on Empirical Methods in Natural Language Processing*, 2023.
- [34] P. Srivastava, M. Malik, and T. Ganu, “Assessing llms’ mathematical reasoning in financial document question answering,” *arXiv preprint arXiv:2402.11194*, 2024.
- [35] Z. Cheng, T. Xie, P. Shi, C. Li, R. Nadkarni, Y. Hu, C. Xiong, D. Raday, M. Ostendorf, L. Zettlemoyer, N. A. Smith, and T. Yu, “Binding language models in symbolic languages,” in *The Eleventh International Conference on Learning Representations*, 2023.
- [36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [37] Z. Zhang, A. Zhang, M. Li, and A. Smola, “Automatic chain of thought prompting in large language models,” in *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.
- [38] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, “Synthetic prompting: Generating chain-of-thought demonstrations for large language models,” *arXiv preprint arXiv:2302.00618*, 2023.
- [39] K. Shum, S. Diao, and T. Zhang, “Automatic prompt augmentation and selection with chain-of-thought from labeled data,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12113–12139. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.811>
- [40] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, “Self-consistency improves chain of thought reasoning in language models,” *arXiv preprint arXiv:2203.11171*, 2022.
- [41] O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant, “Answering questions by meta-reasoning over multiple chains of thought,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 5942–5966. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.364>
- [42] Y. Cao, S. Chen, R. Liu, Z. Wang, and D. Fried, “Api-assisted code generation for question answering on varied table structures,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 14536–14548.
- [43] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambrø, L. Zettlemoyer, N. Cancedda, and T. Scialom, “Toolformer: Language models can teach themselves to use tools,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 68539–68551.
- [44] T. pandas development team, “pandas-dev/pandas: Pandas.” Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>
- [45] Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.
- [46] Y. Zhao, Y. Li, C. Li, and R. Zhang, “MultiHierTT: Numerical reasoning over multi hierarchical tabular and textual data,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 6588–6600. [Online]. Available: <https://aclanthology.org/2022.acl-long.454>
- [47] P. Pasupat and P. Liang, “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)*, 2015, pp. 1470–1480.
- [48] B. Xiao, M. Simsek, B. Kantarci, and A. A. Alkheir, “Rethinking detection based table structure recognition for visually rich documents,” *arXiv preprint arXiv:2312.00699*, 2023.
- [49] J. Fernandes, B. Xiao, M. Simsek, B. Kantarci, S. Khan, and A. A. Alkheir, “Tablestrrec: framework for table structure recognition in data sheet images,” *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1–19, 2023.
- [50] B. Smock, R. Pesala, and R. Abraham, “PubTables-1M: Towards comprehensive table extraction from unstructured documents,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4634–4642.
- [51] T. Shi, C. Zhao, J. Boyd-Graber, H. Daumé III, and L. Lee, “On the potential of lexico-logical alignments for semantic parsing to sql queries,” *arXiv preprint arXiv:2010.11246*, 2020.
- [52] Y. Zhang, J. Henkel, A. Floratou, J. Cahoon, S. Deep, and J. M. Patel, “Reactable: Enhancing react for table question answering,” *arXiv preprint arXiv:2310.00815*, 2023.
- [53] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, “Complexity-based prompting for multi-step reasoning,” *arXiv preprint arXiv:2210.00720*, 2022.
- [54] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun, “Towards understanding chain-of-thought prompting: An empirical study of what matters,” in *The 61st Annual Meeting of the Association for Computational Linguistics*, 2023.
- [55] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang, “Tabfact: A large-scale dataset for table-based fact verification,” *arXiv preprint arXiv:1909.02164*, 2019.