

Efficient Prompting for LLM-Based Generative Internet of Things

Bin Xiao¹, Burak Kantarci², *Senior Member, IEEE*, Jiawen Kang³, *Senior Member, IEEE*,
Dusit Niyato⁴, *Fellow, IEEE*, and Mohsen Guizani⁵, *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 an LLM-based Generative IoT (GloT) 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 Postprocessing 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 plaintext because of their complex structures, heterogeneous data types and sometimes huge sizes. We conduct comprehensive experiments on the two popular Table-QA data sets, and the results show that our proposal can achieve competitive performance compared with state-of-the-art LLMs, demonstrating that the proposed LLM-based GloT system can provide competitive performance with tailored prompting methods and is easily extensible to new tasks without training.

Index Terms—Generative Internet of Things (GloT), large language model (LLM), prompt engineering, table question answering (Table-QA).

I. INTRODUCTION

ARTIFICIAL 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 AI (GAI), especially large language models (LLMs), leveraging its remarkable general capacities to the IoT applications, termed as generative IoT (GloT) [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 GloT 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 data set, making LLMs often have a tremendous number of parameters, which leads 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-8x7B [10] and Llama-3-70B [11]. These numbers of parameters make it not practical to deploy LLMs on the edge IoT devices. Even though deploying LLMs on the edge and cloud servers can be an option, the efficiency of an LLM-based GloT 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 [12], [13]. 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

Received 13 June 2024; revised 22 August 2024; accepted 23 September 2024. Date of publication 4 October 2024; date of current version 25 December 2024. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the CREATE TRAVERSAL Program and NSERC DISCOVERY Program; in part by the National Research Foundation, Singapore; in part by the Infocomm Media Development Authority under its Future Communications Research & Development Programme, AI Singapore Programme under Grant FCP-NTU-RG-2022-010 and Grant FCP-ASTAR-TG-2022-003; in part by the Singapore Ministry of Education (MOE) Tier 1 (RG87/22); and in part by the NTU Centre for Computational Technologies in Finance (NTU-CCTF). (Corresponding author: Mohsen Guizani.)

Bin Xiao and Burak Kantarci are with the School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: bxiao103@uottawa.ca; burak.kantarci@uottawa.ca).

Jiawen Kang is with the School of Automation, Guangdong University of Technology, Guangzhou 510006, China (e-mail: kavinkang@gdut.edu.cn).

Dusit Niyato is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: dniyato@ntu.edu.sg).

Mohsen Guizani is with the Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE (e-mail: mohsen.guizani@mbzuai.ac.ae).

Digital Object Identifier 10.1109/IIOT.2024.3470210

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. Finally, the performance and the scalability of an LLM-based GIoT system are another two critical considerations because both commercial and open-source LLMs have some inherent limitations [14], [15], [16], such as their hallucination issues, limited reasoning capacities for complex tasks. Typically, prompting methods and supervised 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) [16] is a popular prompting method to improve the LLM's reasoning performance by providing reasoning rationales. Program of Thoughts (PoT) [15] is another prompting method generating the 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. In contrast, fine-tuning an LLM for a task can also improve the performance but requires the labeled data sets and computation resources for the model training. As discussed in some studies [17], [18], 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 mitigate these issues and employ prompting methods to enhance the capacities of the proposed system for different tasks. Specifically, a Prompt Management Module and a Postprocessing 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 postprocessing the results generated by the LLMs, as shown in Fig. 1. With the proposed Prompt Management Module and Postprocessing Module, the GIoT system can be easily extended to new tasks by adding tailored task-specific prompts in the Task-specific Prompts Database, providing competitive performance for a wide range of tasks.

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 [19], Table-QA [20], [21], [22], [23], [24] 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

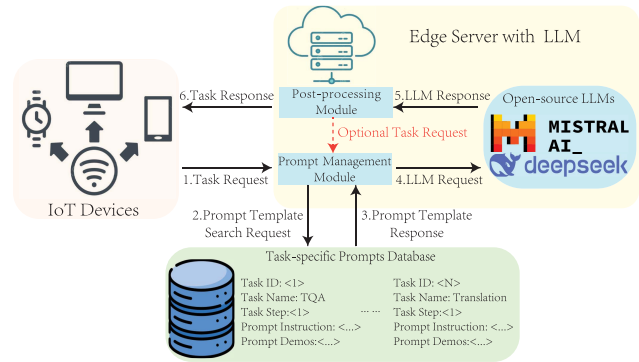


Fig. 1. Overall architecture of the proposed LLM-based GIoT system.

into two groups: 1) structured and 2) 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. In 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 the 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.

To sum up, the contributions of this study can be threefold as follows.

- 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 Postprocessing 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, 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 data sets 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 behavior of different open-source LLMs, which can be a guide for selecting LLMs and prompting methods for an LLM-based GIoT system.

The remainder of this article is organized as follows. Section II discusses related studies, including recent studies applying LLMs to 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. Finally, 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], [25] summarizing the critical components of generative models and discussing the potential of applying generative models to IoT systems. Authors in study [5] first term the combination of generative models and IoT applications as 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 agents to cooperate and analyze the sensor data and user requests. LLMind [26] 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 [27] proposes an intelligent control framework integrating integrated terrestrial nonterrestrial 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 the large-scale data sets and then further fine-tuned LLMs with parameter-efficient fine-tuning methods, such as LORA [28] and LongLORA [29]. TableLLAMA [30] and TAT-LLM [22] are typical examples of applying instruction tuning for the table processing. TableLLAMA is a generalist model for tables fine tuned on Llama2 [31] with LongLORA. For fine-tuning TableLLAMA, a data set collection named TableInstruct is proposed by collecting table-based samples from 14 data sets 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: 1) extractor; 2) reasoner; and 3) executor. To construct the data set 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 LLMs often fail to process large tables and complex reasoning, many studies [22], [32], [33], [34] propose to decompose large tables into small ones in the extraction step and divide the complex reasoning task into more straightforward reasoning questions. For example, StructGPT [35] defines specialized interfaces for information extraction and linearizes the extracted subtables as inputs to the LLM for question answering. EEDP [36] is another example proposing a prompting method containing four steps: 1) elicit; 2) extract; 3) decompose; and 4) 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 [33] is a typical solution following the multistep 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 subquestions, and generates the final results based on the subresults and extracted evidence with in context learning (ICL). Binding [37] 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. Many studies have demonstrated that prompting LLMs with a few demonstrations and intermediate reasoning steps can elicit the reasoning capacities of LLMs, which often refer to ICL [38] and CoT [16]. Although ICL and CoT are useful, writing proper prompting demonstrations is nontrivial and time-consuming. Therefore, some studies [39], [40], [41] propose to use LLMs to generate demonstrations. Auto-CoT [39] 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 [40] 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 [42] 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. Multichain reasoning [43] is another ensemble method mixing information from multiple reasoning chains. Different from studies [42] to ensemble results of multireasoning chains, multichain 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 [15], PAL [14], and many other studies [44] leverage programming language to assist the reasoning steps. Some studies [45] 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 Fig. 1, an open-source LLM can be deployed in a local edge server to process the requests from

IoT devices. This is because fine-tuning an LLM for a specific task requires substantial data sets 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 Postprocessing Module, and a Task-specific Prompts Database, as shown in Fig. 1. Typically, a prompt often consists of instructions and demonstrations, in which the instructions are used to describe the task, and the demonstrations are the examples provided to the LLMs to show the desired outputs. 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 for the task-specific prompt instructions and demonstrations regarding the task, as steps 1–3 in Fig. 1. After obtaining the task-specific prompt instructions and demonstrations, 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 instructions and demonstrations in the Task-specific Prompt Database without training or fine-tuning to the LLM [46]. Since the results generated by the LLM are sometimes not ideal, we propose a Postprocessing 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 Postprocessing Module. Besides, programming language-aided prompting methods, such as PAL [14] and PoT [15], generate the Python code instead of the final results, which means that the Postprocessing 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 the HTTP protocol.

Following these discussed steps, Fig. 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: 1) Request Parsing; 2) Prompt Search; and 3) 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

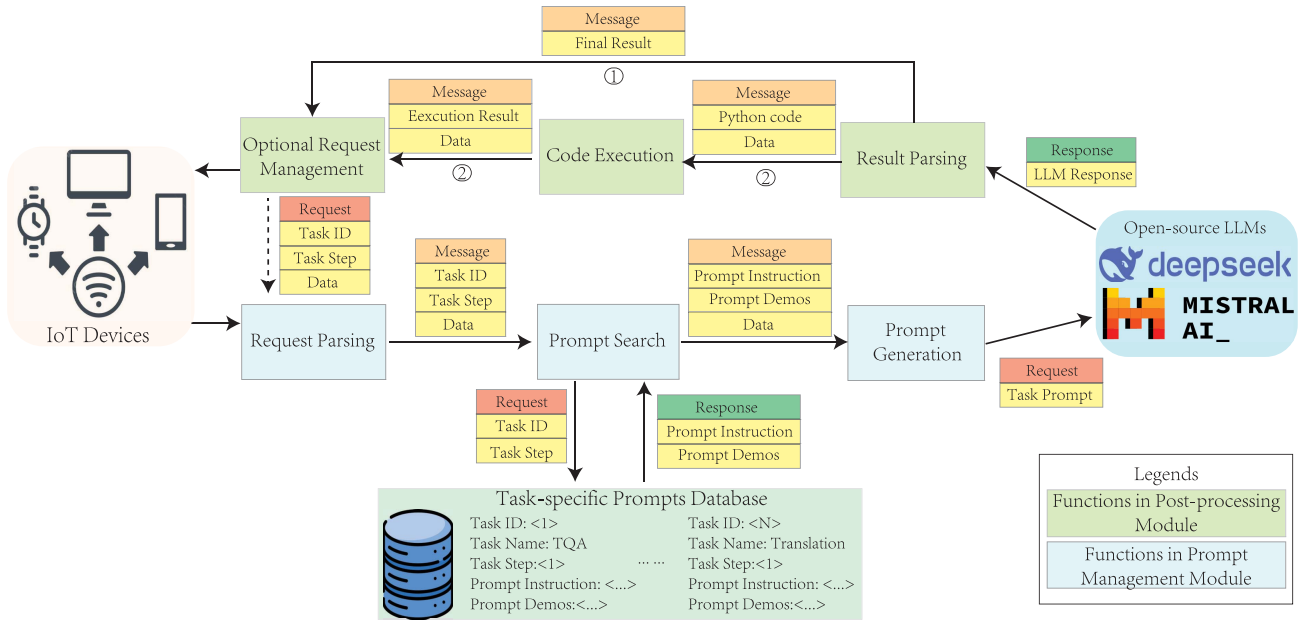


Fig. 2. Detailed workflow of the proposed LLM-based GIoT system.

Prompt Generation can implement customize demonstration selection methods to optimize the performance of applying ICL [38]. For the Postprocessing 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 Fig. 2. In contrast, prompting methods, such as PoT [15] and PAL [14], generate the Python code and conduct the reasoning steps by an external Python interpreter. Therefore, a code execution component should be included in the Postprocessing Module. Finally, an 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 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

A Table-QA service deployed in the proposed LLM-based GIoT system contains two key components: 1) a parametric LLM with parameters θ and 2) a retriever \mathcal{R} , in which the retriever \mathcal{R} is responsible for selecting proper instructions and demonstrations used in ICL. Therefore, for an input query pair, (t, q) , including a query table and a query question, the output answer after n rounds of prompting y_n can be defined as

$$y_{i+1} = f_{\theta}(y_i, t, q, \mathcal{R}_i), \quad \text{for } i = 0, 1, \dots, n-1$$

where y_0 is an empty string.

As mentioned in Section I, many prompting methods have been proposed to enhance LLMs' capacities. For example, CoT [16] is a popular prompting method providing reasoning rationales to elicit LLMs' reasoning capacities. PAL [14] and PoT [15] propose to offload the reasoning steps to the 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 [32]. Fig. 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 the Python variables. Applying Pandas library [47], [48] can alleviate this information extraction issue [49] by providing proper selection criteria, but introducing extra difficulties caused by the heterogeneous data types. Fig. 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. Fig. 4(a) shows an example from the WikiTableQA [50] data set, which contains several spanning cells across multiple columns and rows, and inconsistent data types, such as the column Season. Even though some methods [51], [52], [53] 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 the Python code to answer the question "What is the maximum

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 subtables in these stages to avoid the original large tables.

Club	Season	League		Cup ¹		Continental ²		Other ³		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)

```
#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)

Fig. 4. 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. (a) Sample table with a complex structure from the WikiTableQA data set. (b) Sample of generated Python code to answer the question *What is the maximum League Apps after 2004?*

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 Fig. 4(b), which is another failure example caused by the heterogeneous data types. Along with applying Python to enhance the LLMs, some studies [33], [37] 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 [37], [54], 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 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 [33], [55] propose decomposing huge tables into subtables for further reasoning, they need to prompt the full huge table into the LLM first. Some studies [32] 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 performance further but simultaneously drastically increase the inference cost.

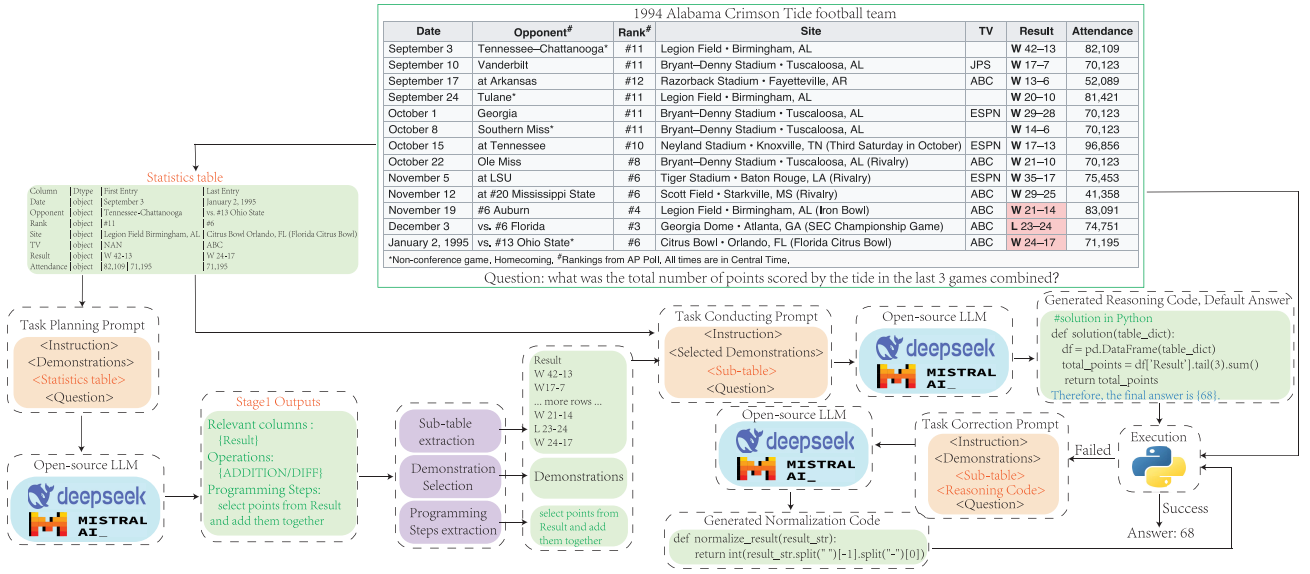


Fig. 5. Workflow of the proposed prompting solution. Notably, the question and the table are from the request of an IoT device. The Python interpreter is in the Postprocessing Module, and the stages of selecting demonstrations and creating prompts are in the Prompt Management Module.

Finally, most of these studies are based on ICL, using demonstrations to guide the LLM in conducting target tasks, while crafting and selecting proper demonstrations is still an open issue. Many studies [56], [57] 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 Fig. 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

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 Moncoutie (FRA)
Team | object | Caisse d' Epargne | Cofidis
Time | object | 5h 29' 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 below regarding " + <title> + " to answer the following one question:
<Statistics Table>
<Question>

Fig. 6. 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.

prompt open-source LLMs to generate the 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 Fig. 5, in which the question is the total number of points scored by the tide in the final three games combined. To answer this question, the statistics table of the given table contains information regarding the column names, data types, and the first and final entries, which are first generated as a part of the task-planning prompt, together with instructions, demonstrations, and questions, as shown in Fig. 6. Then, the task-planning prompt is fed into

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

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 Fig. 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,” and “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 Fig. 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 Fig. 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

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 Clásica de San Sebastián 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 | 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 Pellisoni (ITA) | Liquigas | s.t. | 15
6 | Denis Menchov (RUS) | Rabobank | s.t. | 11
7 | Samuel Sanchez (ESP) | Euskaltel-Euskadi | s.t. | 7
8 | Stéphane Goubert (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> +” to answer the following one question with Python code.

<Statistics Table>
<Column Details>
<Question>

Fig. 7. Task-conducting prompt. The statistics table and column details are highlighted with yellow, and the default answers are highlighted with blue. <title>, red <Statistics Table>, <Column Details>, and <Question> are from the table to be analyzed.

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. Fig. 7 shows an example of COUNT operation, which includes a statistics 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 Figs. 1 and 2.

V. EXPERIMENTS AND ANALYSIS

A. Data Sets and Experimental Settings

We evaluate our proposed prompting solution on the WikiTableQA [50] and TabFact [58] data sets, which are 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 data set, 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 data sets to evaluate the performance, which contains 4344 and 12828 QA pairs, respectively. TabFact data set further categorizes the test set into simple and complex sets, which include 4219 and 8609 QA pairs, respectively. For the simple set, the QA pairs are usually obtained from a single row or record in the table, reflecting unary facts without complex logical inference. In contrast, for the complex set, the QA pairs are created by information from multiple columns and rows and derived by complex semantic operations, such as argmax, argmin, and the table records are also rewritten to include more semantic understanding. Considering the large size of the TabFact data set, some studies [33], [37] conducted experiments on a small subset of TabFact, which contains 1005 simple and 1019 complex QA pairs. To compare with these studies, we also report the results on this small test subset of TabFact.

As the proposed LLM-based GIoT system is designed to be deployed in a local network, we use open-source LLMs, including Mixtral-8x7B,¹ Mistral-7B,² DeepSeek-67B,³ and DeepSeek-7B⁴ to conduct our experiments. Specifically, we deploy the LLMs with vLLM [59] on a workstation with 8 NVIDIA A100 40G GPUs. For the Mistral-7B and DeepSeek-7B models with default 16-bit parameters, one A100 40G GPU has provided enough RAM. By contrast, for the Mixtral8x7B and DeepSeek-67B with default 16-bit parameters, four A100 40G GPUs are needed for their deployment with tensor parallel.

Since the proposed solution in this study is a prompt engineering method, we include Direct Prompting, CoT [16], PoT [15], Binder [37], and Dater [33] as benchmarks, in which Binder and Dater are two solutions leveraging SQL. For the implementation of benchmark methods, we use the implementation of TableCoT⁵ [32] for the Direct Prompting and CoT. We reimplemented the PoT method following the example prompts reported in PoT [15]. The results of Dater and Binder are directly from study [33]. We use a simple beam search decoding for the experiments. It is worth mentioning that self-consistency and more complex sampling decoding methods are often used in state-of-the-art techniques, such as Dater and Binder, to improve the quality of generated texts, which can introduce significant inference overhead. Finally, 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 postprocess 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 [32]. It is worth mentioning that we use the default

TABLE II
EXPERIMENTAL RESULTS ON WIKITABLEQA DATA SET WITH EXACT MATCH ACCURACY AS METRIC

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

precision of parameters, namely bfloat16, for the LLMs in this section.

We term our proposed prompting solution as Tab-PoT, and the experimental results are shown in Tables II and 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 data set. 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 data set while consistently improving the performance on the TabFact data set.

B. Discussion and Analysis

1) *Ablation Study*: As discussed in Section IV, our proposed prompting solution consists of three stages: 1) task-planning; 2) task-conducting; and 3) task-correction. The task-planning stage can output the relevant columns and reasoning steps to answer the query question, which is also a step of table decomposition and question decomposition that can 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

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

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

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

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

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

TABLE III
EXPERIMENTAL RESULTS ON TABFACT DATA SET WITH EXACT MATCH ACCURACY AS METRIC. *full* AND *small* MEAN THE FULL AND SMALL VERSIONS OF TABFACT DATA SET

LLM	Method	Simple _{full}	Complex _{full}	All _{full}	Simple _{small}	Complex _{small}	All _{small}
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	90.09	78.91	82.58	91.34	80.27	85.77
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 DATA SET 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

TABLE V
IMPACT OF LLM QUANTIZATION METHODS. NOTABLY, THE LLM USED IN THIS TABLE IS MIXTRAL-8X7B

#GPU	Model Size	Parameter Precision	EM Acc	AVG Prompt Throughput	AVG Completion Throughput
4	87G	16-bit	63.33	743.8 tokens/s	58.8 tokens/s
4	48G	8-bit	62.20	553.7 tokens/s	39.7 tokens/s
2	48G	8-bit	62.20	377.6 tokens/s	26.5 tokens/s
4	25G	4-bit	60.52	581.3 tokens/s	49.3 tokens/s
2	25G	4-bit	60.52	427.7 tokens/s	35.9 tokens/s
1	25G	4-bit	60.52	286.8 tokens/s	24.3 tokens/s

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) *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 data set and compare its 16, 8, 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 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 achieves much higher performance, as shown in Tables V and II.

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 Fig. 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 Fig. 9. Finally, 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 Fig. 10. We conduct experiments on the WikiTableQA data set to compare the performance of these three types of implementations. Even though the implementation of applying the 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 the Pandas library can achieve better performance than using the standard library.

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 data set. As shown in Table VII, the proposed

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	David Rebellin (ITA)	Gerolsteiner	s.t.	25
4	Paolo Bettini (ITA)	Quick Step	s.t.	20
5	Franco Pellizzotti (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 Pellizzotti (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 Pellizzotti (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.

TABLE VI
COMPARISONS OF THREE TYPES OF THREE DIFFERENT TYPES OF IMPLEMENTATIONS FOR THE POT METHOD

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

Tab-PoT can introduce some overhead compared with other methods regarding the average prompt tokens and average completion tokens on the WikiTableQA data set. 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 data set. 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

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 Pellizzotti (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(sub_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.

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

computation operations and less inference time, as shown in Fig. 11. Since the WikiTableQA data set 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 [32], the LLM can perform 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].

As mentioned in Section V-A, we deployed the LLMs with vLLM [59] on a workstation with 8 NVIDIA A100 40G GPUs. We utilized a single GPU for the Mistral-7B and DeepSeek-7B and four GPUs for the Mixtral-8x7B and DeepSeek-67B. The throughput performances of each model (with respect to prompt and completion) are summarized in Table VIII. As shown in Table VIII, the throughput performance of the

Answer the questions based on the two given tables:

Rank	Cyclist	Team	Time	UCI ProTour Points
1	Alejandro Valverde (ESP)	Caisse d'Épargne	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 Pellizzotti (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:

```
#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.index[df["Cyclist"].str.contains("Alexandr Kolobnev")][0]
    # Extract the cyclist after Alexandr Kolobnev
    next_cyclist = df.loc[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)))
```

...More Lines...
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.

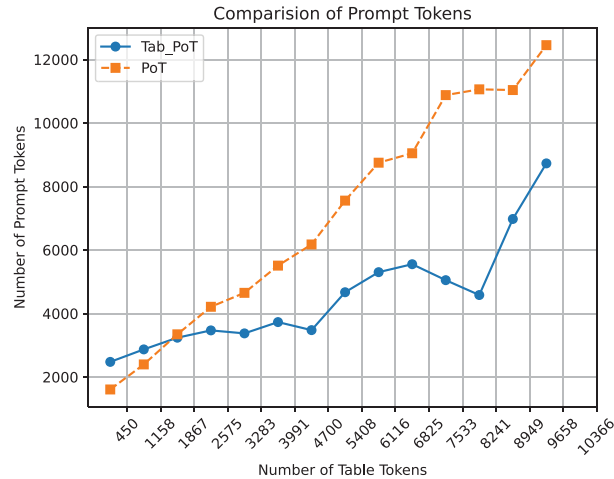


Fig. 11. Comparison of prompting tokens between PoT and Tab_PoT.

DeepSeek-67B is significantly lower than other models for both prompt and completion, which is caused by the overhead of using tensor parallelism. Specifically, we observed that the number of components of the DeepSeek-67B deployed in the CPUs is much larger than that of Mixtral-8x7B for the tensor parallelism, making it much slower than Mixtral-8x7B which also needs to be deployed using tensor parallelism. In contrast, the Mistral-7B and DeepSeek-7B, both of which can

TABLE VIII
COMPARISONS OF INFERENCE SPEEDS OF LLMs

Model	#GPU	AVG Prompt Throughput	AVG Completion Throughput
Mixtral-8x7B	4	743.8 tokens/s	58.8 tokens/s
Mistral-7B	1	633.0 tokens/s	57.7 tokens/s
DeepSeek-67B	4	320.7 tokens/s	20.1 tokens/s
DeepSeek-7B	1	705.2 tokens/s	59.4 tokens/s

be deployed in a single GPU, can show similar throughput performances. Finally, it is worth mentioning that the inference overheads caused by the Prompt Management Module and the Postprocessing Module are negligible, because they can be finished in milliseconds for our cases, while the LLM inference usually needs at least a few seconds.

VI. CONCLUSION AND FUTURE WORK

In this study, we propose an 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 subtables 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 the system. Besides, as the proposed GIoT system is designed to deploy in an 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. Besides extending the system to multiple modalities, improving computational efficiency and reducing hardware requirements are also critical to the system. Therefore, the caching mechanisms that cache intermediate results or frequently used outputs, model pruning that can reduce the size of LLM models, and many others are also promising directions.

REFERENCES

- [1] S. Hu, M. Li, J. Gao, C. Zhou, and X. Shen, "Adaptive device-edge collaboration on DNN inference in AIoT: A digital-twin-assisted approach," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 12893–12908, Apr. 2024.

- [2] M. Adil et al., "Healthcare Internet of Things: Security threats, challenges, and future research directions," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19046–19069, Jun. 2024.
- [3] J. Fan et al., "Understanding security in smart city domains from the ANT-centric perspective," *IEEE Internet Things J.*, vol. 10, no. 13, pp. 11199–11223, Jul. 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 Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2351–2383, 4th Quart., 2021.
- [5] J. Wen et al., "From generative AI to generative Internet of Things: Fundamentals, framework, and outlooks," *IEEE Internet Things Mag.*, vol. 7, no. 3, pp. 30–37, May 2024.
- [6] X. Wang et al., "IoT in the era of generative AI: Vision and challenges," 2024, *arXiv:2401.01923*.
- [7] N. Zhong et al., "CASIT: Collective intelligent agent system for Internet of Things," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19646–19656, Jun. 2024.
- [8] J. Achiam, "GPT-4 technical report," 2023, *arXiv:2303.08774*.
- [9] J. Kaplan et al., "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.
- [10] A. Q. Jiang et al., "Mixtral of experts," 2024, *arXiv:2401.04088*.
- [11] (Meta AI, New York, NY, USA). *Meta AI Blog: LLaMA 3*. 2024. Accessed: Jun. 10, 2024. [Online]. Available: <https://ai.meta.com/blog/meta-llama-3/>
- [12] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Apr. 2023.
- [13] B. Yan et al., "On protecting the data privacy of large language models (LLMs): A survey," 2024, *arXiv:2403.05156*.
- [14] L. Gao et al., "PAL: Program-aided language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 10764–10799.
- [15] W. Chen, X. Ma, X. Wang, and W. W. Cohen, "Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks," *Trans. Mach. Learn. Res.*, to be published. [Online]. Available: <https://openreview.net/forum?id=YfZ4ZPt8zd>
- [16] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.
- [17] T. Tu et al., "Towards lifelong unseen task processing with a lightweight Unlabeled data schema for AIoT," *IEEE Internet Things J.*, early access, May 6, 2024, doi: [10.1109/JIOT.2024.3396282](https://doi.org/10.1109/JIOT.2024.3396282).
- [18] 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," 2023, *arXiv:2309.16119*.
- [19] B. Xiao, M. Simsek, B. Kantarci, and A. A. Alkheir, "Table detection for visually rich document images," *Knowl.-Based Syst.*, vol. 282, Dec. 2023, Art. no. 111080.
- [20] J. Herzig, P. K. Nowak, T. Müller, F. Piccinno, and J. M. Eisenschlos, "TaPas: Weakly supervised table parsing via pre-training," 2020, *arXiv:2004.02349*.
- [21] 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," 2022, *arXiv:2207.03637*.
- [22] 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," 2024, *arXiv:2401.13223*.
- [23] Z. Yu, L. He, Z. Wu, X. Dai, and J. Chen, "Towards better chain-of-thought prompting strategies: A survey," 2023, *arXiv:2310.04959*.
- [24] B. Zhao et al., "Large language models are complex table parsers," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2023, pp. 14786–14802.
- [25] S. De, M. Bermudez-Edo, H. Xu, and Z. Cai, "Deep generative models in the Industrial Internet of Things: A survey," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 5728–5737, Sep. 2022.
- [26] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, "LLMind: Orchestrating AI and IoT with LLMs for complex task execution," 2023, *arXiv:2312.09007*.
- [27] B. Rong and H. Rutagemwa, "Leveraging large language models for intelligent control of 6G integrated TN-NTN with IoT service," *IEEE Netw.*, vol. 38, no. 4, pp. 136–142, Jul. 2024.
- [28] E. J. Hu et al., "LoRa: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
- [29] Y. Chen et al., "LongLoRa: Efficient fine-tuning of long-context large language models," 2023, *arXiv:2309.12307*.
- [30] T. Zhang, X. Yue, Y. Li, and H. Sun, "TableLLama: Towards open large generalist models for tables," 2023, *arXiv:2311.09206*.
- [31] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.
- [32] W. Chen, "Large language models are few (1)-shot table reasoners," in *Proc. Find. Assoc. Comput. Linguist.*, 2023, pp. 1120–1130.
- [33] 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 *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2023, pp. 174–184.
- [34] W. Lin, R. Billoshmi, B. Byrne, A. de Gispert, and G. Iglesias, "An inner table retriever for robust table question answering," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, 2023, pp. 9909–9926.
- [35] J. Jiang, K. Zhou, Z. Dong, K. Ye, X. Zhao, and J.-R. Wen, "StructGPT: A general framework for large language model to reason over structured data," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Dec. 2023, pp. 9237–9251. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.574>
- [36] P. Srivastava, M. Malik, and T. Ganu, "Assessing LLMs' mathematical reasoning in financial document question answering," 2024, *arXiv:2402.11194*.
- [37] Z. Cheng et al., "Binding language models in symbolic languages," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–27. [Online]. Available: <https://openreview.net/forum?id=IH1PV42cbF>
- [38] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.
- [39] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic chain of thought prompting in large language models," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–25. [Online]. Available: <https://openreview.net/forum?id=5NT8GFjUHkr>
- [40] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Synthetic prompting: Generating chain-of-thought demonstrations for large language models," 2023, *arXiv:2302.00618*.
- [41] K. Shum, S. Diao, and T. Zhang, "Automatic prompt augmentation and selection with chain-of-thought from Labeled data," in *Proc. Find. Assoc. Comput. Linguist.*, Dec. 2023, pp. 12113–12139. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.811>
- [42] X. Wang et al., "Self-consistency improves chain of thought reasoning in language models," 2022, *arXiv:2203.11171*.
- [43] O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant, "Answering questions by meta-reasoning over multiple chains of thought," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, Dec. 2023, pp. 5942–5966. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.364>
- [44] Y. Cao, S. Chen, R. Liu, Z. Wang, and D. Fried, "API-assisted code generation for question answering on varied table structures," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2023, pp. 14536–14548.
- [45] T. Schick et al., "Toolformer: Language models can teach themselves to use tools," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 68539–68551.
- [46] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different NLP tasks," 2024, *arXiv:2407.12994*.
- [47] The Pandas Development Team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.3509134>
- [48] W. McKinney, "Data structures for statistical computing in python," in *Proc. 9th Python Sci. Conf.*, 2010, pp. 56–61.
- [49] Y. Zhao, Y. Li, C. Li, and R. Zhang, "MultiHiert: Numerical reasoning over multi hierarchical tabular and textual data," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguist.*, May 2022, pp. 6588–6600. [Online]. Available: <https://aclanthology.org/2022.acl-long.454>
- [50] P. Pasupat and P. Liang, "Compositional semantic parsing on semi-structured tables," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguist. 7th Int. Joint Conf. Nat. Lang. Process.*, 2015, pp. 1470–1480.
- [51] B. Xiao, M. Simsek, B. Kantarci, and A. A. Alkheir, "Rethinking detection based table structure recognition for visually rich documents," 2023, *arXiv:2312.00699*.
- [52] J. Fernandes, B. Xiao, M. Simsek, B. Kantarci, S. Khan, and A. A. Alkheir, "TableStrRec: Framework for table structure recognition in data sheet images," *Int. J. Doc. Anal. Recognit.*, vol. 27, pp. 1–19, Sep. 2023.
- [53] B. Smock, R. Pesala, and R. Abraham, "PubTables-1M: Towards comprehensive table extraction from unstructured documents," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4634–4642.
- [54] 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," 2020, *arXiv:2010.11246*.

- [55] Y. Zhang, J. Henkel, A. Floratou, J. Cahoon, S. Deep, and J. M. Patel, "ReAcTable: Enhancing ReAct for table question answering," 2023, *arXiv:2310.00815*.
- [56] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot, "Complexity-based prompting for multi-step reasoning," 2022, *arXiv:2210.00720*.
- [57] B. Wang et al., "Towards understanding chain-of-thought prompting: An empirical study of what matters," in *Proc. Workshop Math. Empir. Underst. Found. Models*, 2023, pp. 1–21. [Online]. Available: <https://openreview.net/forum?id=L9UMeoeU2i>
- [58] W. Chen et al., "TabFact: A large-scale dataset for table-based fact verification," 2019, *arXiv:1909.02164*.
- [59] W. Kwon et al., "Efficient memory management for large language model serving with PagedAttention," in *Proc. ACM SIGOPS 29th Symp. Oper. Syst. Princ.*, 2023, pp. 611–626.



Bin Xiao received the bachelor's degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2012, the master's degree from National Chiao Tung University, Hsinchu, Taiwan, in 2016, and the Ph.D. degree with the University of Ottawa, Ottawa, ON, Canada, in 2024 under the supervision of Prof. Kantarci.

His research interests include machine learning, natural language processing, and Internet of Things.

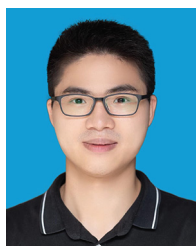


Burak Kantarci (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Istanbul Technical University, Istanbul, Türkiye.

He is a Full Professor and the Founding Director of the Smart Connected Vehicles Innovation Centre and the Next Generation Communications and Computing Networks Research Lab, Ottawa, ON, Canada. He is the author/co-author of 300+ publications in established journals and conferences and 15 book chapters. Continuously listed among the top-cited scientists in telecommunications and

networking based on the data reported by Stanford University since 2020, based on data collected from Microsoft Academic Graph, research.com has listed him among Canada's Top computer scientists. He has been a keynote/invited speaker or panelist in 40 events.

Prof. Kantarci holds an Exemplary Editor Award from IEEE Communications Surveys and Tutorials in 2021 and multiple Best Paper Awards from various conferences, most recently from IEEE Globecom 2021, Wireless World Research Forum 2022, IEEE ICC 2023, and IEEE VCC 2023. He is a recipient of the Minister's Award of Excellence from Ontario Ministry of Colleges and Universities in 2021. He is the recipient of the 2023 Technical Achievement Award of IEEE ComSoc Communications Software Technical Committee. From 2019 to 2020, he chaired the Communications Systems Integration and Modeling Technical Committee of the Institute of Electrical and Electronics Engineers. He has been the general chair, the program chair, or the track chair in 30+ international conferences. He is an Editor of IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING and an Associate Editor of IEEE NETWORKING LETTERS and *Vehicular Communications* (Elsevier). He was a Distinguished Speaker of the Association of Computing Machinery from 2019 to 2021. He is currently a Distinguished Lecturer of the IEEE Communications Society and IEEE Systems Council.



Jiawen Kang (Senior Member, IEEE) received the M.S. and the Ph.D. degrees from Guangdong University of Technology, Guangzhou, China, in 2015 and 2018, respectively.

He is currently a Full Professor with Guangdong University of Technology. He has been a Postdoctoral Fellow with Nanyang Technological University, Singapore, from 2018 to 2021. His research interests mainly focus on blockchain, metaverse, and AIGC in wireless communications and networking.



Dusit Niyato (Fellow, IEEE) received the B.Eng. degree from the King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Winnipeg, MB, Canada, in 2008.

He is a Professor with the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include areas of mobile generative AI, edge intelligence, decentralized machine learning, and incentive mechanism design.



Mohsen Guizani (Fellow, IEEE) received the B.S. degree (with Distinction), M.S., and Ph.D. degrees in electrical and computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1987, and 1990, respectively.

He is currently a Professor of Machine Learning with the Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. Previously, he worked with different institutions in USA. He is the author of 11 books, more than 1000 publications and several US patents. His research interests

include applied machine learning, artificial intelligence, smart city, Internet of Things, intelligent autonomous systems, and cybersecurity.

Dr. Guizani has won several research awards, including the 2015 IEEE Communications Society Best Survey Paper Award, the Best ComSoc Journal Paper Award in 2021, and the five Best Paper Awards from ICC and Globecom Conferences. He is also the recipient of the 2017 IEEE Communications Society Wireless Technical Committee Recognition Award, the 2018 AdHoc Technical Committee Recognition Award, and the 2019 IEEE Communications and Information Security Technical Recognition Award. He served as the Editor-in-Chief for *IEEE Network* and is currently serving on the editorial boards for many IEEE transactions and magazines. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the TAOS Technical Committee. He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science from 2019 to 2022. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer.