

# Self-Correction Distillation for Structured Data Question Answering

Yushan Zhu<sup>1,2</sup>, Wen Zhang<sup>1\*</sup>, Long Jin<sup>1</sup>, Mengshu Sun<sup>3</sup>, Ling Zhong<sup>3</sup>, Zhiqiang Liu<sup>1</sup>, Juan Li<sup>1</sup>,  
Lei Liang<sup>3</sup>, Chong Long<sup>2</sup>, Chao Deng<sup>2</sup>, Junlan Feng<sup>2</sup>

<sup>1</sup>Zhejiang University, China

<sup>2</sup>JiUTIAN Research, Beijing, China

<sup>3</sup>Ant Group, China

zhuyushan@cmjt.chinamobile.com, {zhang.wen, longjin}@zju.edu.com

## Abstract

Structured data question answering (QA), including table QA, Knowledge Graph (KG) QA, and temporal KG QA, is a pivotal research area. Advances in large language models (LLMs) have driven significant progress in unified structural QA frameworks like TrustUQA. However, these frameworks face challenges when applied to small-scale LLMs since small-scale LLMs are prone to errors in generating structured queries. To improve the structured data QA ability of small-scale LLMs, we propose a self-correction distillation (SCD) method. In SCD, an error prompt mechanism (EPM) is designed to detect errors and provide customized error messages during inference, and a two-stage distillation strategy is designed to transfer large-scale LLMs' query-generation and error-correction capabilities to small-scale LLM. Experiments across 5 benchmarks with 3 structured data types demonstrate that our SCD achieves the best performance and superior generalization on small-scale LLM (8B) compared to other distillation methods, and closely approaches the performance of GPT4 on some datasets. Furthermore, large-scale LLMs equipped with EPM surpass the state-of-the-art results on most datasets.

## 1 Introduction

Structured data question answering (QA) (Zhang et al. 2025) encompasses diverse sub-fields, including table QA (Liu et al. 2022; Xie et al. 2022), knowledge graph QA (KGQA) (Saxena, Tripathi, and Talukdar 2020; Jiang et al. 2023c; Luo et al. 2024a) and temporal KGQA (Liu et al. 2023; Shang et al. 2022). The goal is to retrieve answers from structured data given natural language queries. Leveraging the advanced reasoning capabilities of large language models (LLMs), recent approaches employ them as agents to generate reasoning paths (Sun et al. 2024; Luo et al. 2024b; Xu et al. 2024; Cheng et al. 2024) or decompose questions into structured queries (Li et al. 2023b; Luo et al. 2024a; Jiang et al. 2023a; Zhang et al. 2025).

In recent years, a growing research trend focuses on unified frameworks for multi-type structured QA, with notable advances such as StructGPT (Jiang et al. 2023a), ReadI (Cheng et al. 2024), and TrustUQA (Zhang et al.

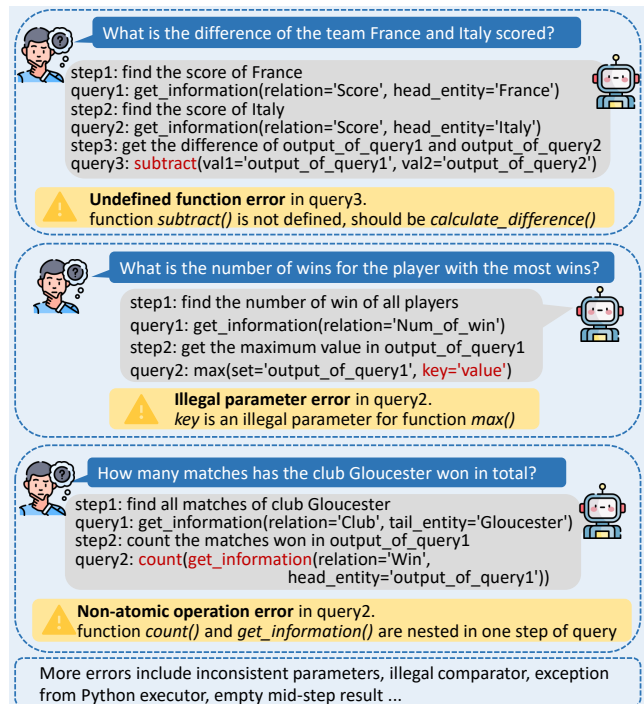


Figure 1: Examples of errors in small-scale LLM-generated queries with TrustUQA (Zhang et al. 2025).

2025). These methods rely on API-based, closed-source LLMs (e.g., GPT3.5/GPT4) for query generation or inference. However, in real-world applications, many users prefer independently deployed LLM solutions due to concerns over data privacy or API service stability. Considering two practical factors in structured data QA scenarios: first, many scenarios lack the hardware capacity to deploy large-scale LLMs (over 100B parameters); second, they do not require all the general AI capabilities such as writing or solving complex math problems. Thus, a small-scale LLM (under 10B parameters) that supports flexible deployment and excels in structured data QA is essential.

However, directly adapting unified structured QA frameworks such as TrustUQA (Zhang et al. 2025) to small-scale LLM presents significant challenges. Small-scale LLM

\*Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

demonstrates notable limitations in structured query generation, particularly in accurately identifying and decomposing interdependent subproblems within complex questions. As illustrated in Figure 1, small-scale LLM exhibits various errors when processing complex queries, including: calling undefined functions, supplying illegal parameters, improperly nesting function calls within single query step, etc.

This work introduces Self-Correction Distillation (SCD), a novel Chain of Thought (CoT) distillation framework (Li et al. 2023a; Wang et al. 2023) to transfer structured QA capabilities from a large-scale teacher LLM to a small-scale student LLM.

First, we propose an Error Prompt Mechanism (EPM) that facilitates iterative error correction during the inference process. EPM classifies LLM-generated query errors into different types and provides type-specific feedback. This closed-loop mechanism continuously monitors outputs, detects errors, and provides targeted corrective feedback to guide LLM through correction cycles iteratively. Second, SCD employs a two-stage distillation strategy consisting of teacher-distillation and self-distillation stages. In the teacher-distillation stage, the student learns from both correct teacher-generated queries and CoT-based error correction trajectories through supervised fine-tuning. This stage helps transfer the query-generation and error-correction capabilities from teacher to student. The subsequent self-distillation stage enables autonomous capability refinement, where the student model iteratively improves its performance by learning from its own outputs. This stage prevents recurring errors while enhancing student’s ability to generate correct queries directly. The two-stage approach transfers teacher expertise and promotes student self-improvement, addressing both error correction and prevention in small-scale LLM.

We conduct extensive experiments on 5 benchmarks spanning 3 types of structured data, following the evaluation protocol of TrustUQA (Zhang et al. 2025), the latest unified structured QA framework. Experimental results demonstrate that: (1) the proposed method outperforms all existing distillation approaches on small-scale LLM (8B parameters); (2) our 8B LLM achieves performance competitive with large-scale LLM (GPT4) on certain datasets; and (3) when equipped with EPM, large-scale LLM (GPT4) surpasses current state-of-the-art performance across most benchmarks. Through comprehensive model analysis, we further validate our method’s strong generalization capability to unseen data. The key contributions of this work are threefold:

- We introduce Self-Correction Distillation (SCD), a novel CoT distillation framework that improves small-scale LLMs’ structured data QA performance by transferring expertise from large-scale LLMs.
- In SCD, we design an Error Prompt Mechanism (EPM) to detect errors in LLM-generated queries and provide type-specific corrective feedback, along with a two-stage distillation strategy to enhance small-scale LLMs’ query generation and error correction abilities.
- We experimentally demonstrate that our method surpasses state-of-the-art CoT distillation approaches in en-

hancing small-scale LLM’s structured QA capability, while exhibiting strong generalization performance on unseen dataset.

## 2 Related Works

### 2.1 Structured Data QA

Structured data question answering (QA) includes sub-fields of table QA (Liu et al. 2022; Xie et al. 2022), knowledge graph question answering (KGQA) (Saxena, Tripathi, and Talukdar 2020; Jiang et al. 2023c; Luo et al. 2024a) and temporal KGQA (Liu et al. 2023; Shang et al. 2022). The goal is to answer a natural-language question using structured data. The field employs two main paradigms: NL2Answer and NL2Query. NL2Answer approaches generate answers directly through language models. Some methods (Jiang et al. 2023c; Sun et al. 2018; Jiang et al. 2023b) train text encoders for question or reasoning representation. TableGPT (Zha et al. 2023) encodes table embedding and uses it to prompt-tuning LLM. NL2Query translates questions into formal queries, offering greater interpretability through transparent execution (Zhang et al. 2025). Representation works include ZeroNL2SQL (Gu et al. 2023) and DIN-SQL (Pourreza and Rafiei 2023) for table QA, KB-BINDER (Li et al. 2023b) for KG QA. Recent advances have focused on unified frameworks for cross-domain structured QA. StructGPT (Jiang et al. 2023a) fetches, linearizes, and inputs structured knowledge into an LLM for direct answering. ReadI (Cheng et al. 2024) generates and refines a reasoning path, gathers KG evidence, then uses LLMs to produce the final answer. TrustUQA (Zhang et al. 2025) is the latest unified structured data QA framework. TrustUQA adopts an LLM-friendly representation method called Condition Graph, and a two-level LLM-based querying method with dynamic demonstration retrieval.

### 2.2 CoT Distilling

Knowledge distillation (Hinton, Vinyals, and Dean 2015) has been widely proven to be effective in transferring knowledge from a larger model (teacher) to a smaller model (student). The main idea is to optimize the parameters of student by making its output (Zhu et al. 2022), distribution (Feng et al. 2021), hidden state (Jiao et al. 2020), or attention (Tang et al. 2019), etc. close to that of teacher. However, the black-box nature of current mainstream large-scale LLMs (e.g., GPT3.5, GPT4) limits the application of these methods.

Many studies perform distillation by fine-tuning smaller models directly on the output or chain of thought (CoT) of large-scale LLMs (Ho, Schmid, and Yun 2023; Hsieh et al. 2023; Fu et al. 2023), referred to as CoT distillation. SCoTD (Li et al. 2023a) introduces symbolic distillation to transfer CoT reasoning ability. PERsD (Chen et al. 2023a) addresses the knowledge gap between student and teacher for code generation. Instead of fine-tuning student with correct code from teacher, PERsD enables teacher to refine the student’s incorrect code. KPOD (Feng et al. 2024) imitates human cognitive patterns, training student from simple tasks and gradually to more challenging tasks. SCOTT (Wang et al. 2023) enables the student to reason faithfully through

	Error Type	Example	Error Message
PE	Undefined function name	<code>subtract(set1=..., set2=...)</code>	<i>The function 'subtract' is not defined! Please call one of: ['get_information', 'min', 'mean', 'max', 'count', 'sum', 'keep', 'set_intersection', 'set_union'...</i>
	Illegal parameter	<code>max(set=..., key=...)</code>	<i>For function 'max', parameter name 'key' is illegal, which must be in ['set'].</i>
	Inconsistent parameters	<code>gi(he=..., relation=..., te=...)</code>	<i>For function 'get_information', it is not allowed to assign values to parameters ['head_entity', 'relation', 'tail_entity'] at the same time.</i>
	Illegal comparator	<code>gi(te=..., relation=..., he&lt;...)</code>	<i>In function 'get_information', comparison '&lt;' for 'head_entity' is illegal, non-equal comparators are only allowed for parameters 'tail_entity' and 'value'.</i>
	Non-atomic operation	<code>sum(set=set_negation(...))</code>	<i>The query is not an atomic operation: functions 'sum' and 'set_negation' are nested. Please make sure that each step is atomic.</i>
	Other non-standard expressions	<code>sum(set=[output_of_query1, 2])</code>	<i>Parsing the passed parameter value '[output_of_query1, 2]' failed. Please ensure that the format of the query is correct</i>
EE	Exception from Python executor	<code>sum(set=output_of_query1) //output_of_query1 must be numeric</code>	<i>Exception from Python in function 'sum': unsupported operand type(s) for +: 'int' and 'str'. //a string-type element in output_of_query1</i>
	Empty mid-step result	<code>gi(relation='Hometown', te='Utah')</code>	<i>For query1, the result=set(), empty output_of_query1 may affect subsequent query and final result. Please verify the correctness of entity or relation.</i>
		<code>//query1: find the entity from Utah //Plausible rel: Hometown, Colleges</code>	

Table 1: Error Messages. PE(Parsing Error), EE(Execution Error), *gi*(*get\_information*), *he*(*head\_entity*), *te*(*tail\_entity*)

counterfactual training. MCC-KD (Chen et al. 2023b) focuses on CoT reasoning diversity and consistency. To mitigate the impact of teacher hallucinations or errors on the student, (Yang et al. 2024) filters out the incorrect CoT and conclusions. OmniD (Bao et al. 2024) uses multiple sub-models to create a consensus-based pseudo-label. Some works (Li et al. 2024; Zhang et al. 2024) also aim to extract useful knowledge from incorrect answers generated by teacher.

### 2.3 LLM Correction

LLM error correction aims to address reliability challenges in LLM-generated content (Pan et al. 2023). While early approaches like DIN-SQL (Pourreza and Rafiei 2023) introduces general prompting techniques for SQL error correction, their reliance on intrinsic LLM capabilities without external validation limits effectiveness (Huang et al. 2024a). Recent work has consequently shifted toward environmental and execution feedback mechanisms, broadly categorized into prompt-based and training-based methods.

Prompt-based approaches include SPINACH’s graph-exploration agent (Liu et al. 2024), LDB’s block-level verification (Zhong, Wang, and Shang 2024), QueryAgent’s environmental feedback (Huang et al. 2024b), SALAM’s dynamic demonstration retrieval (Wang and Li 2023), few-shot prompted self-correction for SQL generation (Chen et al. 2024), and KB-Coder’s code-style context learning (Nie et al. 2024). Although avoiding training, these methods face limited demonstration coverage and poor adaptability to small-scale LLMs with weak in-context learning.

Training-based solutions include reinforcement learning methods like LEDEX (Jiang et al. 2024) and PPOCoder (Shojaee et al. 2023), which suffer from reward design complexity and training instability. CoT distillation methods like PERsD (Chen et al. 2023a) offer more stable training through teacher-guided correction, though their student-centric error sampling restricts exposure to diverse error types and risks data scarcity.

In this work, we reduce errors in small-scale LLM for structured QA task by synergistically combining feedback-enriched prompting with CoT distillation.

## 3 Preliminary

We briefly outline the workflow of the latest unified structured data QA method TrustUQA (Zhang et al. 2025). Given a question  $Q$  and structured data with schema  $S$ , TrustUQA first converts structured data into Condition Graph (CG), a general and expressive data representation proposed in TrustUQA. TrustUQA then employs a Two-layer Function-based CG Query method for querying. In the first layer, LLM breaks the question down into multiple query steps and write simplified functions like *get\_information()* for each step, referred to as LLM queries. (The prompt template P-Q to generate LLM queries according to the given  $Q$  and  $S$  is detailed in Appendix A.) The names and vocabularies of these functions (head entity, tail entity, relation, key, and value) are general and more understandable to LLM. In the second layer, each LLM query is translated into an execution query according to predefined rules, which are then executed over CG to retrieve the final result from structured data.

In this work, we formally define the collection of all steps and LLM queries for a question from the first layer as a “query”. The main focus of this work is to optimize the ability of small-scale LLM to write correct “query”. The second layer, which handles the parsing, translation, and execution of queries from the first layer, is defined as “query executor”.

## 4 Method

In this section, we introduce our error prompt mechanism (EPM), multi-round correction process with EPM, and two-stage distillation strategy.

### 4.1 Error Prompt Mechanism (EPM)

Error prompt mechanism (EPM) is embedded in the query executor to detect errors and report customized error messages. EPM focuses on two main categories of errors in LLM-generated queries: parsing errors and execution errors.

**Parsing Error Message** Parsing error refers to the errors found during the parsing phase before query execution, including undefined function names, illegal parameters, inconsistent parameters, illegal comparators, and non-atomic operations in LLM-generated queries. The first six rows of Ta-

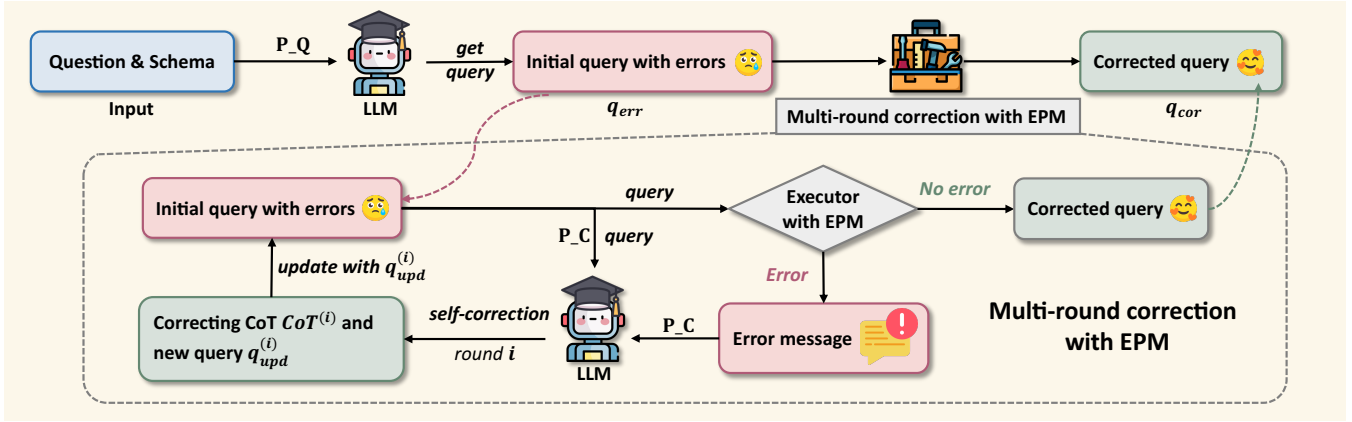


Figure 2: Multi-round Correction with EPM

ble 1 are examples of each parsing error and the corresponding error message.

**Execution Error Message** Execution error, usually caused by queries’ logic error, is found during the execution of a query after it has successfully passed the parsing process. There are two main types: 1) exception from Python executor refers to errors that can be caught by Python’s try-except statement during execution, and 2) empty mid-step result refers to the result of a mid-step query is empty. This means that the query for that step or the previous steps may be incorrect. Examples of each execution error and the error messages are in the last two rows of Table 1.

Note that false negatives may occur in the second type of execution error when the correct answer to a query is “null/none”. In this case, the correction iteration stops after reaching the maximum correction time.

**Other errors** The rest are errors unrelated to the LLM, e.g. entities or relations in the LLM-generated query are retrieved incorrectly, resulting in incorrect query results. Since such errors cannot be solved by optimizing LLM-related modules and their proportion is not high, we do not solve them in this work.

## 4.2 Multi-round Correction with EPM

As depicted in Figure 2, for a given question and data schema, LLM first writes the initial query. The query executor then parses and executes this initial query. If no error occurs, the initial query is output as the final correct query. Otherwise, the initial query is recorded as  $q_{err}$ , and the EPM reports the corresponding error message described in Section 4.1. LLM is prompted to analyze the error’s cause and correct the initial wrong query based on the error message. After updating the initial query, the query executor parses and executes it again. The above process iterates until no error occurs or exceeds the maximum correction time  $MCT$ . We denote the error analysis by LLMs as  $CoT^{(i)}$  and the updated query as  $q_{upd}^{(i)}$  after the  $i$ -th correction round. Assuming that after  $n$  ( $n \leq MCT$ ) rounds, the query  $q_{upd}^{(n)}$

can be parsed and executed without error, we output it as the final correct query  $q_{cor}$ .

## 4.3 Two-stage Distillation Strategy

Two-stage distillation strategy includes *teacher-distillation* and *self-distillation* stages, where large-scale LLM serves as teacher and small-scale LLM serves as student.

**The Teacher-distillation Stage** In the first stage, the student learns query generation and error correction capabilities from teacher. The teacher generates data for supervised fine-tuning the student.

In left part of Figure 3, firstly, to fully explore and utilize the errors of different models, both student and teacher generate initial queries for a given question and data schema. The initial query with errors from student and teacher are denoted as  $q_{err}^s$  and  $q_{err}^t$ , and the error message reported by EPM is denoted as  $err^{(0)}$ . Then  $q_{err}^s$  or  $q_{err}^t$  and  $err^{(0)}$  are input into teacher for multiple rounds of error correction. In the  $i$ -th correction round, the error analysis by teacher is denoted as  $CoT_t^{(i)}$ , and the updated query is denoted as  $q_{upd}^{t(i)}$ . After  $n$  rounds of correction, if the query  $q_{upd}^{t(n)}$  can be executed without errors and yields the correct answer, it is denoted as  $q_{cor}^t$ . Note that  $q_{cor}^t$  represents all correct queries from the teacher, both those generated directly by teacher and those corrected from teacher-generated or student-generated wrong queries.

To train the student’s query generation ability, question  $Q$  and data schema  $S$  are input into student, and  $q_{cor}^t$  is the target output:

$$\mathcal{L}_q = - \sum_j \log P_{\mathcal{M}}(q_{cor(j)}^t | \text{P-Q}(Q, S); q_{cor(<j)}^t) \quad (1)$$

where  $q_{cor(j)}^t$  is the  $j$ -th token of  $q_{cor}^t$ , P-Q is the prompt template (Zhang et al. 2025) of query generation, and more details are in Appendix A. To train the student’s error correction ability, for each correction round  $i$ , question  $Q$ , data schema  $S$ , wrong query  $q_{upd}^{t(i-1)}$  from the last round, and the

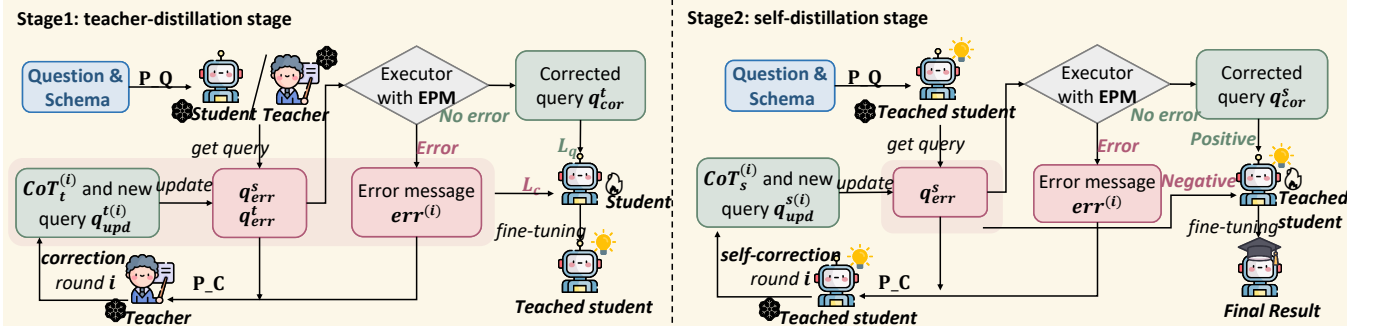


Figure 3: Two-stage Distillation Strategy

error message  $err^{(i-1)}$  are input into student, and error analysis  $CoT_t^{(i)}$  from teacher and the final correct query  $q_{cor}^t$  are the target output:

$$\mathcal{L}_c = - \sum_{i=1}^n \sum_j \log P_{\mathcal{M}}([CoT_t^{(i)}; q_{cor}^t]_{(j)} | \text{P\_C}(Q, S, q_{upd}^{t(i-1)}, err^{(i-1)}); [CoT_t^{(i)}; q_{cor}^t]_{(<j)}) \quad (2)$$

where  $q_{upd}^{t(i)} = q_{err}^t$  or  $q_{err}^s$ ,  $[x; y]$  represents text concatenation of  $x$  and  $y$ , P\_C is the prompt template of error correction, and more details are in Appendix A.

Note that for each correction round  $i$ , we use the final correct  $q_{cor}^t$  as the target output for student, rather than the updated query  $q_{upd}^{t(i)}$  of that round. This is because non-final-round queries still contain errors, and learning to generate them does not contribute to the LLM’s QA ability. Moreover, this design makes different correction rounds become tasks of varying difficulty: In later rounds, the degree of match between  $CoT_t^{(i)}$  and  $q_{cor}^t$  is higher, allowing student to generate the correct query more easily based on  $CoT_t^{(i)}$ . In earlier rounds,  $CoT_t^{(i)}$  and  $q_{cor}^t$  match less, and  $CoT_t^{(i)}$  only describes part of the clues from  $q_{upd}^{t(i-1)}$  to  $q_{cor}^t$ . So besides  $CoT_t^{(i)}$ , student must also fully understand the association between data schema and problem decomposition, which is more challenging.

The final loss of the first distillation stage is

$$\mathcal{L}_1 = \mathcal{L}_q + \mathcal{L}_c \quad (3)$$

**The Self-distillation Stage** After the first stage, student acquires basic query generation and error correction capabilities. In this stage, student improves itself based on its current output instead of learning from teacher, aiming to write correct queries with minimal corrections.

In right part of Figure 3, for a given question and data schema, student writes the initial query, denoted as  $q_{err}^s$  if an error occurs in the query executor, with error message  $err^{(0)}$ . Then the wrong query  $q_{err}^s$  and error message  $err^{(0)}$  are feedback to itself for multiple rounds of error correction. The updated query after the  $i$ -th correction round is  $q_{upd}^{s(i)}$ . After  $n$  rounds, query  $q_{upd}^{s(n)}$  is denoted as  $q_{cor}^s$  if it can be executed with no error and yields the correct answer. Since  $q_{upd}^{s(i)}$

( $0 \leq i < n$ ) is the wrong query that current student more likely to generate, for the given question and data schema, we increase the generation probability of  $q_{cor}^s$  and reduce the generation probability of  $q_{upd}^{s(i)}$  ( $0 \leq i < n$ ). The loss of the second distillation stage is

$$\mathcal{S}(q) = \sum_{j=1}^{|q|} \log P_{\mathcal{M}}(q_{(j)} | \text{P\_Q}(Q, S); q_{(<j)}) \quad (4)$$

$$\mathcal{L}_2 = - \sum_{i=1}^n \left( S(q_{cor}^s) - S(q_{upd}^{s(i-1)}) \right),$$

where  $q_{upd}^{s(0)} = q_{err}^s$ .

## 5 Experiments

### 5.1 Dataset

Following (Zhang et al. 2025), we adopt 3 types of structured data: table with WikiSQL (Zhong, Xiong, and Socher 2017) and WTQ (Pasupat and Liang 2015), KG with WebQuestionsSP(WebQSP) (Yih et al. 2016) and MetaQA (Zhang et al. 2018), temporal KG with CronQuestions(CronQ) (Saxena, Chakrabarti, and Talukdar 2021).

Datasets	# SFT		# Raw			Other Information train/dev/test
	train	dev	train	dev	test	
<b>WTQ</b>	11,322	2,830	14.2K	-	4.3K	2,100 tables
<b>WikiSQL</b>	56.351	3,000	56.4K	8.4K	15.9K	26,531 tables
<b>MetaQA</b>	30,000	6,000	329.3K	39.1K	39.1K	134,741 triples
- 1-hop	9,000	2,000	96.1K	10.0K	9.9K	simple 1 hop
- 2-hop	10,500	2,000	119.0K	14.9K	14.9K	21 types
- 3-hop	10,500	2,000	114.2K	14.3K	14.3K	15 types
<b>WebQSP</b>	2,788	308	2,788	308	1,639	retrieved version
<b>CronQ</b>	25,000	4,000	350K	30K	30K	333.6K triples
- Simple	13,000	2,800	181.3K	21.3K	17.1K	simple types
- Complex	12,000	1,200	168.7K	8.7K	12.9K	3 complex types
- Entity	16,204	2,485	225.7K	19.4K	19.5K	query for entity
- Time	8,796	1,515	124.3K	10.6K	10.5K	query for time

Table 2: Statistics of experimental datasets.

Due to the unbalanced distribution of datasets and the high cost of using OpenAI, we employ uniform sampling for excessively large data to construct our final training dataset. For MetaQA, we sample 30,000 training and 6,000

Method	Table QA				KG QA					Temporal KG QA												
	WikiSQL		WTQ		MetaQA					WebQSP		CronQuestion										
					1-hop		2-hops		3-hops			All		Complex		Simple		Entity		Time		
SOTA perf.	89.5 <sup>†</sup>		66.7*		98.4*		99.9*		99.4*	85.7 <sup>◊</sup>		97.2 <sup>‡</sup>		95.4 <sup>‡</sup>		99.5 <sup>‡</sup>		96.1 <sup>‡</sup>		99.1 <sup>‡</sup>		
teacher(GPT4)	<b>91.1</b>	87.7	53.2	48.5	<b>98.6</b>	97.2	99.4	98.2	99.3	98.3	<b>91.7</b>	84.4	<b>98.1</b>	97.6	<b>97.3</b>	96.5	<b>99.7</b>	99.6	<b>96.9</b>	96.2	<b>99.6</b>	99.5
Llama3.1	74.5	73.4	35.1	33.9	94.6	93.5	91.1	88.7	83.4	81.9	75.2	74.9	88.1	87.8	80.9	80.7	97.8	97.5	83.2	82.9	96.9	96.8
Naive-SFT	79.3	77.1	42.2	40.7	96.1	94.5	93.6	92.7	88.1	86.2	78.2	76.7	92	90.8	86.4	85.3	99.2	99.1	88.6	87.2	98.7	96.5
PERsD	84.2	79.6	44.3	40.4	97.2	96.2	98.1	94.8	97.3	93.2	79.7	75.6	93.4	91.2	89.1	86.1	99.2	99.1	91.4	87.7	97.2	96.5
KPOD	86.1	81.8	46.5	42.7	97.1	96.8	<b>98.5</b>	96.7	97.8	95.6	82.3	77.3	95.5	92.1	92.4	87.9	99.3	98.6	94.2	88.7	98.4	97.1
SCD(ours)	<b>86.9</b>	<b>83.1</b>	<b>48.9</b>	<b>43.9</b>	<b>97.4</b>	<b>97.1</b>	98.4	<b>97.2</b>	<b>98.2</b>	<b>96.3</b>	<b>83.7</b>	<b>78.8</b>	<b>96.9</b>	<b>92.6</b>	<b>95.1</b>	<b>88.8</b>	<b>99.5</b>	<b>99.2</b>	<b>95.8</b>	<b>89.2</b>	<b>98.8</b>	<b>97.3</b>

Table 3: Denotation accuracy of TableQA, Hit@1 of KG QA and temporal KG QA. *w/* denotes “inference w/ EPM”, and *w/o* denotes “inference w/o EPM”. The SOTA performance is the best-known results from TAPEX<sup>†</sup> (Liu et al. 2022), ReadI\* (Cheng et al. 2024), RoG<sup>◊</sup> (Luo et al. 2024b), and TrustUQA(GPT3.5)<sup>‡</sup> (Zhang et al. 2025). The bold results in teacher represent exceeding SOTA, and bold in students is the best of all students.

development instances according to each hop’s proportion. For CronQuestion, we sample 25,000 training instances and 4,000 development instances according to each class’s proportion. For WikiSQL, we sample 3,000 development instances. Since WTQ has no validation set, we split the raw training set of WTQ into random 80-20 splits for development following (Pasupat and Liang 2015). The statistical details of the original dataset (#Raw) and our sampled dataset for training (#SFT) are presented in Table 2.

## 5.2 Baseline

We choose GPT4 as teacher and Llama3.1-8B-Instruct (Meta 2024; Dubey et al. 2024) as student.

As our method is fundamentally a CoT distillation approach, we compare against representative CoT distillation baselines without distinguishing between code or text generation tasks. The baseline methods include: (1) **Llama3.1**, the original Llama3.1 base model; (2) **Naive-SFT**, the most commonly used SFT strategy, which trains on error-free executable queries (both directly generated and corrected by the teacher) but neglects error correction capability; (3) **PERsD** (Chen et al. 2023a), a code correction CoT distillation method, focuses on generating personalized data by teacher refining student output, with training data that excludes teacher-generated correct queries and error correction trajectories of teacher-generated wrong queries; (4) **KPOD** (Feng et al. 2024), the state-of-the-art CoT distillation method, employing progressive task difficult by decomposing complete queries into step-wise subtasks, ranging from generating subsequent steps given preceding ones (simple) to full query generation (complex).

We evaluate two inference configurations: (1) **inference w/ EPM**, where LLMs perform iterative error correction during multi-round reasoning, and (2) **inference w/o EPM**, limited to single-pass reasoning.

## 5.3 Experiment Setting

We implement our method by Llama-Factory (Zheng et al. 2024), an open-source LLM fine-tuning framework based on PyTorch. We use the LoRA (Hu et al. 2022) fine-tuning strategy and train the student on 2\*NVIDIA A100 PCIe 40GB GPUs with per device batch size as 1 and gradient accumulation steps as 8 for 3 epochs for each distillation stage.

We use fp16 training with AdamW (Loshchilov and Hutter 2019) optimizer whose learning rate is 0.0001, together with a cosine learning rate scheduler with a warm-up ratio of 0.1. Same as previous work (Zhang et al. 2025), we use SentenceBERT (Reimers and Gurevych 2019) as the dense text encoder, set the number of retrieves to 15, the number of query generation demonstrations in P\_Q to 8 and use self-consistency strategy of 5 times. For multi-round error correction, we set the number of error correction demonstrations in P\_C to 8 and the max correction time to 3. Detailed specifications of EPM implementation and error type provenance are in Appendix B.

## 5.4 Result Analysis

Table 3 shows the results on different datasets. Teacher “inference w/ EPM” improves significantly than “inference w/o EPM” and outperforms SOTA on most datasets, demonstrating the effectiveness of EPM. Although our student model does not achieve SOTA, it outperforms other distillation baseline methods on almost all datasets. The results of the raw Llama3.1 are significantly worse than others. For Llama3.1 and Naive-SFT, “inference w/ EPM” shows limited improvement compared to “inference w/o EPM”, indicating that small-scale LLMs can’t obtain good error correction capability without specific fine-tuning.

## 5.5 Model analysis

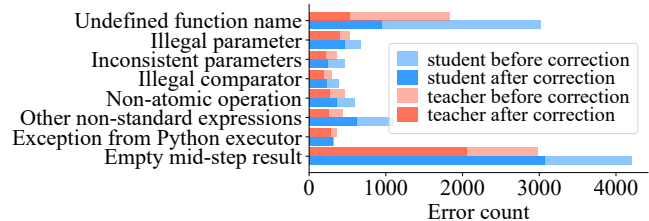


Figure 4: Error count before/after correction on all test data. Bar values are the percentage of errors corrected.

**Correction of Different Error Types** Figure 4 shows the most common error type is the empty mid-step result of execution errors, followed by undefined function names in

	WikiSQL	WTQ	MetaQA				WebQSP	Cronquestion				
			avg.	1-hop	2-hops	3-hops		All	Com	Sim	Ent	Tim
SCD	<b>86.9</b>	<b>48.9</b>	<b>98.0</b>	<b>97.4</b>	<b>98.4</b>	<b>98.2</b>	<b>83.7</b>	<b>96.9</b>	<b>95.1</b>	<b>99.5</b>	<b>95.8</b>	<b>98.8</b>
- <i>self-distillation</i>	86.1	46.2	97.1	97.1	97.2	97.1	82.2	95.3	93.6	99.2	92.6	98.5
- <i>error-msg in EPM</i>	83.7	44.8	97.2	97.2	97.7	96.8	79.5	93.8	90.9	99.1	90.9	97.7

Table 4: Ablation Study. - *self-distillation* means to train student with only teacher-distillation stage. - *error-msg in EPM* means EPM only signals error occurrence but does not provide details on error type and specific messages.

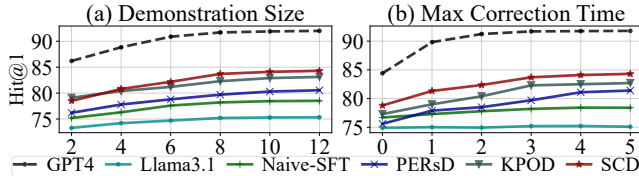


Figure 5: Hyperparameter analysis on WebQSP.

parsing errors. The average correction rate of parsing errors reaches 40.2% and 45.5% for teacher and student, while on execution errors it’s only 27.6% and 15.7% for teacher and student. Overall, parsing errors are easier to correct than execution errors, likely because parsing errors are often superficial and identifiable, whereas the root causes of execution errors are harder to pinpoint.

**Ablation Study** Table 4 shows that without the second self-distillation stage (- *self-distillation*) or without specific error messages during the multiple rounds of error correction (- *error-msg in EPM*), the performance of the student model significantly declines in all datasets. Especially on more complex problems such as 3-hops of MetaQA and complex class of Cronquestion, the performance is notably diminished if LLMs do not obtain the error details during error correction.

**Hyperparameter Analysis** Figure 5 shows that increasing the number of error correction demonstrations and correction times slightly improves the performance of large-scale LLMs and small-scale LLMs fine-tuned with PERsD, KPOD, and SCD. However, it has little effect on the original Llama3.1 or models fine-tuned with Naive-SFT, which lack special training for error correction. Figure 5(b) also highlights that large-scale LLMs have stronger error correction abilities than small-scale LLMs: GPT4 requires a maximum of 2 rounds of error correction to correct most errors, while small-scale LLMs of PERsD, KPOD, and SCD need 3 rounds or more. Balancing cost and performance, we set the final correction demonstration number to 8 and the maximum correction times to 3 for all experiments.

## 5.6 QA on Unseen Dataset

We evaluate small-scale LLMs on the unseen QA dataset TabFact (Chen et al. 2020) and show results in Table 5. Our small-scale LLM, significantly smaller than StructGPT (a unified structured data QA framework based on GPT3.5), achieves comparable results, demonstrating our approach’s strong generalization. Compared to other knowledge distil-

SOTA perf. of unified QA		87.6*	
		<i>inference w/ EPM</i>	<i>inference w/o EPM</i>
Llama3.1		64.0	60.8
Naive-SFT		75.4	69.3
PERsD		79.8	74.7
KPOD		79.2	75.1
SCD		83.4	78.8

Table 5: Accuracy on TabFact. SOTA performance of unified models is from StructGPT(GPT3.5)\* (Jiang et al. 2023a).

lation methods, our small-scale LLM exhibits superior generalization on new data.

## 6 Limitations

We observe four failure patterns where small-scale LLM fails to produce correct answers: (1) Correction threshold exceeded: The maximum correction iterations are reached, typically for questions that are too difficult for current model. (2) False negative corrections: Valid queries returning “null/none” values are misidentified as “empty mid-step output” execution error, causing EPM to erroneously modify correct queries. (3) Base model capability gap: For queries requiring “LLM function” of TrustUQA (Zhang et al. 2025), the 8B student model underperforms GPT4 due to inherent capacity limitations. Based on the same query, large-scale LLMs can output the correct answer but small-scale LLMs cannot. (4) Alignment errors: Queries are successfully parsed and executed yet yield incorrect answers, primarily due to relation or entity mapping errors during the parameter-value alignment step (Zhang et al. 2025).

## 7 Conclusion

In this work, we propose Self-Correction Distillation (SCD), a novel CoT-distillation method for enhancing structured data QA capabilities in small-scale LLM. SCD integrates two key components: (1) an error prompt mechanism (EPM) that provides detailed error location and correction guidance during inference, and (2) a two-stage distillation strategy that effectively transfers both query-generation and error-correction competencies from large-scale LLM to small-scale LLM. Experimental results demonstrate that SCD achieves state-of-the-art performance on 8B parameter LLM with superior generalization capabilities. And GPT4 equipped with EPM surpasses SOTA benchmarks across most datasets. In the future, we focus on addressing the limitations outlined in Section 6, including false negative cases where “null/none” values trigger execution error, and relation and entity alignment errors.

## Acknowledgments

This work is funded by National Natural Science Foundation of China (NSFC62306276/NSFCU23B2055), Yongjiang Talent Introduction Programme (2022A-238-G), and Fundamental Research Funds for the Central Universities (226-2023-00138). This work was supported by Ant Group.

## References

- Bao, P.; Shao, Z.; Yang, W.; Ng, B. P.; Er, M. H.; and Kot, A. C. 2024. Omnipotent Distillation with LLMs for Weakly-Supervised Natural Language Video Localization: When Divergence Meets Consistency. In *AAAI*, 747–755.
- Chen, H.; Saha, A.; Hoi, S.; and Joty, S. 2023a. Personalized Distillation: Empowering Open-Sourced LLMs with Adaptive Learning for Code Generation. In *EMNLP*, 6737–6749.
- Chen, H.; Wu, S.; Quan, X.; Wang, R.; Yan, M.; and Zhang, J. 2023b. MCC-KD: Multi-CoT Consistent Knowledge Distillation. In *EMNLP (Findings)*, 6805–6820.
- Chen, W.; Wang, H.; Chen, J.; Zhang, Y.; Wang, H.; Li, S.; Zhou, X.; and Wang, W. Y. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *ICLR*.
- Chen, X.; Lin, M.; Schärli, N.; and Zhou, D. 2024. Teaching Large Language Models to Self-Debug. In *ICLR*.
- Cheng, S.; Zhuang, Z.; Xu, Y.; Yang, F.; Zhang, C.; Qin, X.; Huang, X.; Chen, L.; Lin, Q.; Zhang, D.; Rajmohan, S.; and Zhang, Q. 2024. Call Me When Necessary: LLMs can Efficiently and Faithfully Reason over Structured Environments. In *ACL (Findings)*, 4275–4295.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; Goyal, A.; Hartshorn, A.; Yang, A.; Mitra, A.; Sravankumar, A.; Korenev, A.; Hinsvark, A.; Rao, A.; Zhang, A.; Rodriguez, A.; Gregerson, A.; Spataru, A.; Rozière, B.; Biron, B.; Tang, B.; Chern, B.; Caucheteux, C.; Nayak, C.; Bi, C.; Marra, C.; McConnell, C.; Keller, C.; Touret, C.; Wu, C.; Wong, C.; Ferrer, C. C.; Nikolaidis, C.; Allonsius, D.; Song, D.; Pintz, D.; Livshits, D.; Esiobu, D.; Choudhary, D.; Mahajan, D.; and et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.
- Feng, K.; Li, C.; Zhang, X.; Zhou, J.; Yuan, Y.; and Wang, G. 2024. Keypoint-based Progressive Chain-of-Thought Distillation for LLMs. In *ICML*.
- Feng, S.; Ren, X.; Li, K.; and Sun, X. 2021. Multi-View Feature Representation for Dialogue Generation with Bidirectional Distillation. In *AAAI*, 12812–12820.
- Fu, Y.; Peng, H.; Ou, L.; Sabharwal, A.; and Khot, T. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 10421–10430.
- Gu, Z.; Fan, J.; Tang, N.; Zhang, S.; Zhang, Y.; Chen, Z.; Cao, L.; Li, G.; Madden, S.; and Du, X. 2023. Interleaving Pre-Trained Language Models and Large Language Models for Zero-Shot NL2SQL Generation. *CoRR*, abs/2306.08891.
- Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.
- Ho, N.; Schmid, L.; and Yun, S. 2023. Large Language Models Are Reasoning Teachers. In *ACL (1)*, 14852–14882.
- Hsieh, C.; Li, C.; Yeh, C.; Nakhost, H.; Fujii, Y.; Ratner, A.; Krishna, R.; Lee, C.; and Pfister, T. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In *ACL (Findings)*, 8003–8017.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*.
- Huang, J.; Chen, X.; Mishra, S.; Zheng, H. S.; Yu, A. W.; Song, X.; and Zhou, D. 2024a. Large Language Models Cannot Self-Correct Reasoning Yet. In *ICLR*.
- Huang, X.; Cheng, S.; Huang, S.; Shen, J.; Xu, Y.; Zhang, C.; and Qu, Y. 2024b. QueryAgent: A Reliable and Efficient Reasoning Framework with Environmental Feedback based Self-Correction. In *ACL (1)*, 5014–5035.
- Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, X.; and Wen, J. 2023a. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. In *EMNLP*, 9237–9251.
- Jiang, J.; Zhou, K.; Zhao, W. X.; Li, Y.; and Wen, J. 2023b. ReasoningLM: Enabling Structural Subgraph Reasoning in Pre-trained Language Models for Question Answering over Knowledge Graph. In *EMNLP*, 3721–3735.
- Jiang, J.; Zhou, K.; Zhao, X.; and Wen, J. 2023c. UniKGQA: Unified Retrieval and Reasoning for Solving Multi-hop Question Answering Over Knowledge Graph. In *ICLR*.
- Jiang, N.; Li, X.; Wang, S.; Zhou, Q.; Hossain, S. B.; Ray, B.; Kumar, V.; Ma, X.; and Deoras, A. 2024. LeDex: Training LLMs to Better Self-Debug and Explain Code. In *NeurIPS*.
- Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, 4163–4174.
- Li, L. H.; Hessel, J.; Yu, Y.; Ren, X.; Chang, K.; and Choi, Y. 2023a. Symbolic Chain-of-Thought Distillation: Small Models Can Also “Think” Step-by-Step. In *ACL (1)*, 2665–2679.
- Li, T.; Ma, X.; Zhuang, A.; Gu, Y.; Su, Y.; and Chen, W. 2023b. Few-shot In-context Learning on Knowledge Base Question Answering. In *ACL*, 6966–6980.
- Li, Y.; Yuan, P.; Feng, S.; Pan, B.; Sun, B.; Wang, X.; Wang, H.; and Li, K. 2024. Turning Dust into Gold: Distilling Complex Reasoning Capabilities from LLMs by Leveraging Negative Data. In *AAAI*, 18591–18599.
- Liu, Q.; Chen, B.; Guo, J.; Ziyadi, M.; Lin, Z.; Chen, W.; and Lou, J. 2022. TAPEX: Table Pre-training via Learning a Neural SQL Executor. In *ICLR*.
- Liu, S.; Semnani, S. J.; Triedman, H.; Xu, J.; Zhao, I. D.; and Lam, M. S. 2024. SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions. In *EMNLP (Findings)*, 15977–16001.
- Liu, Y.; Liang, D.; Li, M.; Giunchiglia, F.; Li, X.; Wang, S.; Wu, W.; Huang, L.; Feng, X.; and Guan, R. 2023. Local

- and Global: Temporal Question Answering via Information Fusion. In *IJCAI*, 5141–5149.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR (Poster)*.
- Luo, H.; E, H.; Tang, Z.; Peng, S.; Guo, Y.; Zhang, W.; Ma, C.; Dong, G.; Song, M.; Lin, W.; Zhu, Y.; and Luu, A. T. 2024a. ChatKBQA: A Generate-then-Retrieve Framework for Knowledge Base Question Answering with Fine-tuned Large Language Models. In *ACL (Findings)*, 2039–2056.
- Luo, L.; Li, Y.; Haffari, G.; and Pan, S. 2024b. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *ICLR*.
- Meta. 2024. Introducing Llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>. Accessed: 2024-10-16.
- Nie, Z.; Zhang, R.; Wang, Z.; and Liu, X. 2024. Code-Style In-Context Learning for Knowledge-Based Question Answering. In *AAAI*, 18833–18841.
- Pan, L.; Saxon, M.; Xu, W.; Nathani, D.; Wang, X.; and Wang, W. Y. 2023. Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies. *CoRR*, abs/2308.03188.
- Pasupat, P.; and Liang, P. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *ACL (1)*, 1470–1480.
- Pourreza, M.; and Rafiei, D. 2023. DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction. In *NeurIPS*.
- Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP/IJCNLP (1)*, 3980–3990.
- Saxena, A.; Chakrabarti, S.; and Talukdar, P. P. 2021. Question Answering Over Temporal Knowledge Graphs. In *ACL/IJCNLP (1)*.
- Saxena, A.; Tripathi, A.; and Talukdar, P. P. 2020. Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings. In *ACL*.
- Shang, C.; Wang, G.; Qi, P.; and Huang, J. 2022. Improving Time Sensitivity for Question Answering over Temporal Knowledge Graphs. In *ACL (1)*.
- Shojaee, P.; Jain, A.; Tipirneni, S.; and Reddy, C. K. 2023. Execution-based Code Generation using Deep Reinforcement Learning. *Trans. Mach. Learn. Res.*, 2023.
- Sun, H.; Dhingra, B.; Zaheer, M.; Mazaitis, K.; Salakhutdinov, R.; and Cohen, W. W. 2018. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. In *EMNLP*.
- Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L. M.; Shum, H.; and Guo, J. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *ICLR*.
- Tang, R.; Lu, Y.; Liu, L.; Mou, L.; Vechtomova, O.; and Lin, J. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *CoRR*, abs/1903.12136.
- Wang, D.; and Li, L. 2023. Learning from Mistakes via Cooperative Study Assistant for Large Language Models. In *EMNLP*, 10667–10685.
- Wang, P.; Wang, Z.; Li, Z.; Gao, Y.; Yin, B.; and Ren, X. 2023. SCOTT: Self-Consistent Chain-of-Thought Distillation. In *ACL (1)*, 5546–5558.
- Xie, T.; Wu, C. H.; Shi, P.; Zhong, R.; Scholak, T.; Yasunaga, M.; Wu, C.; and et al. 2022. UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models. In *EMNLP*.
- Xu, Y.; He, S.; Chen, J.; Wang, Z.; Song, Y.; Tong, H.; Liu, G.; Zhao, J.; and Liu, K. 2024. Generate-on-Graph: Treat LLM as both Agent and KG for Incomplete Knowledge Graph Question Answering. In *EMNLP*, 18410–18430.
- Yang, B.; Tang, C.; Zhao, K.; Xiao, C.; and Lin, C. 2024. Effective Distillation of Table-based Reasoning Ability from LLMs. In *LREC/COLING*, 5538–5550.
- Yih, W.; Richardson, M.; Meek, C.; Chang, M.; and Suh, J. 2016. The Value of Semantic Parse Labeling for Knowledge Base Question Answering. In *ACL (2)*.
- Zha, L.; Zhou, J.; Li, L.; Wang, R.; Huang, Q.; Yang, S.; Yuan, J.; Su, C.; Li, X.; Su, A.; Zhang, T.; Zhou, C.; Shou, K.; Wang, M.; Zhu, W.; Lu, G.; Ye, C.; Ye, Y.; Ye, W.; Zhang, Y.; Deng, X.; Xu, J.; Wang, H.; Chen, G.; and Zhao, J. 2023. TableGPT: Towards Unifying Tables, Nature Language and Commands into One GPT. *CoRR*, abs/2307.08674.
- Zhang, H.; Wang, X.; Ao, X.; and He, Q. 2024. Distillation with Explanations from Large Language Models. In *LREC/COLING*, 5018–5028.
- Zhang, W.; Jin, L.; Zhu, Y.; Chen, J.; Huang, Z.; Wang, J.; Hua, Y.; Liang, L.; and Chen, H. 2025. TrustUQA: A Trustful Framework for Unified Structured Data Question Answering. In *AAAI*, 25931–25939.
- Zhang, Y.; Dai, H.; Kozareva, Z.; Smola, A. J.; and Song, L. 2018. Variational Reasoning for Question Answering With Knowledge Graph. In *AAAI*, 6069–6076.
- Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; Feng, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *ACL*.
- Zhong, L.; Wang, Z.; and Shang, J. 2024. Debug like a Human: A Large Language Model Debugger via Verifying Runtime Execution Step by Step. In *ACL (Findings)*, 851–870.
- Zhong, V.; Xiong, C.; and Socher, R. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *CoRR*, abs/1709.00103.
- Zhu, Y.; Zhang, W.; Chen, M.; Chen, H.; Cheng, X.; Zhang, W.; and Chen, H. 2022. DualDE: Dually Distilling Knowledge Graph Embedding for Faster and Cheaper Reasoning. In *WSDM*, 1516–1524.