

# SQLForge: Synthesizing Reliable and Diverse Data to Enhance Text-to-SQL Reasoning in LLMs

Yu Guo<sup>1</sup>, Dong Jin<sup>2\*</sup>, Shenghao Ye<sup>1</sup>, Shuangwu Chen<sup>1\*</sup>, Jian Yang<sup>1</sup>, Xiaobin Tan<sup>1</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{yukariguo, ssh0321y}@mail.ustc.edu.cn

{kingdon, chensw, jianyang, xbtan}@ustc.edu.cn

## Abstract

Large Language models (LLMs) have demonstrated significant potential in text-to-SQL reasoning tasks, yet a substantial performance gap persists between existing open-source models and their closed-source counterparts. In this paper, we introduce **SQLForge**, a novel approach for synthesizing reliable and diverse data to enhance text-to-SQL reasoning in LLMs. We improve data reliability through SQL syntax constraints and SQL-to-question reverse translation, ensuring data logic at both structural and semantic levels. We also propose an SQL template enrichment and iterative data domain exploration mechanism to boost data diversity. Building on the augmented data, we fine-tune a variety of open-source models with different architectures and parameter sizes, resulting in a family of models termed **SQLForge-LM**. SQLForge-LM achieves the state-of-the-art performance on the widely recognized Spider and BIRD benchmarks among the open-source models. Specifically, SQLForge-LM achieves EX accuracy of 85.7% on Spider Dev and 59.8% on BIRD Dev, significantly narrowing the performance gap with closed-source methods.

## 1 Introduction

Text-to-SQL, which transforms natural language questions into SQL queries, serves as a critical bridge between non-expert users and database systems, significantly lowering the barriers to data access (Liu et al., 2024; Shi et al., 2024). Recently, LLMs have demonstrated exceptional capabilities in various NLP tasks, marking a new paradigm for text-to-SQL solutions. Existing LLM-based approaches (Pourreza and Rafiei, 2024; Gao et al., 2023) usually rely on powerful closed-source models combined with prompt engineering (Wei et al., 2022). Although these methods achieve impressive performance, they pose challenges such as high

computational costs, limited customizability, and significant data privacy concerns.

In response, open-source LLMs have gained attraction due to their lower costs, enhanced data privacy, and greater customizability, making them well-suited for resource-constrained or privacy-sensitive applications. Despite their success in various NLP tasks, open-source models still have a significant performance gap in text-to-SQL (Yang et al., 2024b). For example, the performance difference between CodeLlama-13B and GPT-4 on BIRD benchmark exceeds 20%. To narrow this gap, we need to synthesize large-scale text-to-SQL data to fine-tune open-source LLMs for enhancing their capabilities for text-to-SQL reasoning.

Existing data synthesis methods for text-to-SQL, such as AnnotatedTables (Hu et al., 2024) and MultiSQL (Li et al., 2024a), primarily rely on simple prompt engineering to expand data from existing domains directly. The data generated by these methods suffer from data domain scarcity and poor SQL structure diversity. Other methods like SENSE (Yang et al., 2024b), generate data in new domains directly based on prompting. However, the generated data exhibit poor execution rates and weak semantic alignment between question and SQL pairs, requiring significant time and labor for filtering and refinement. Specifically, according to our evaluation, when the token length is 50, the SQL statements generated by GPT-4 have a high execution failure rate of about 14%. To generate text-to-SQL data at scale, we have to overcome two critical challenges: (1) how to ensure the diversity of synthetic data to enhance model generalization, and (2) how to ensure the reliability of synthetic data without manual annotation.

To tackle these challenges, we seek to use syntactic constraints to improve augmented data’s reliability, and introduce SQL structure enrichment and domain exploration to expand diversity. Specifically, to guarantee the executability of generated

\*Corresponding authors

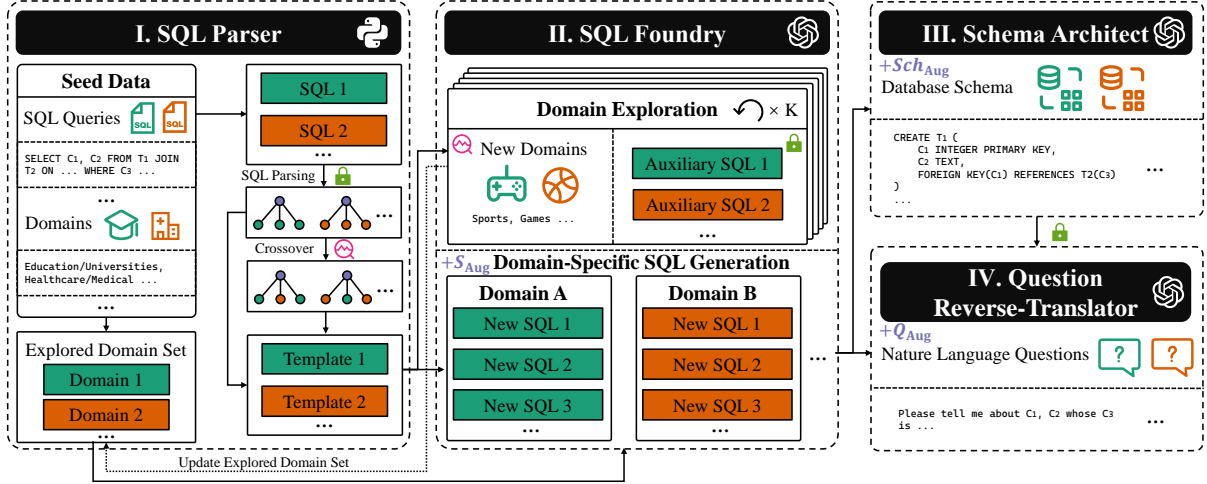




Figure 1: Overview of the proposed SQLForge framework. SQLForge comprises four key components. The SQL Parser generates and enriches SQL templates from seed SQL queries. SQL Foundry generates SQL statements across diverse domains using templates, and upon reaching a domain threshold, it focuses on generating statements within these domains. Then, Schema Architect adds detailed schemas to the generated SQL statements. Finally, Question Reverse-Translator converts the SQL queries into natural language questions aligned with their schemas. The symbol  represents reliability enhancements, while  signifies diversity enhancements.

SQL statements, we use the templates derived from parsing valid SQL statements as the syntactic constraints of SQL for data synthesis. Additionally, to increase the diversity of the synthetic data, we expand the sentence patterns of these templates and propose an iterative domain exploration to generate SQL of entirely new domains. Different from previous work (Hu et al., 2023; Kobayashi et al., 2025), which performs direct SQL-to-question in existing domains, we synthesize data of entirely new domains, generate corresponding database schemas, and incorporate these schemas into the final problem translation, thereby strengthening semantic alignment and enhancing data reliability.

In light of the above idea, we propose a data synthesis framework, named **SQLForge**, to generate a large-scale, reliable, and diverse text-to-SQL dataset as shown in Fig. 1. To evaluate SQLForge’s effectiveness, we fine-tune several popular open-source pretrained models, such as CodeLlama (Roziere et al., 2023), resulting in a new family of text-to-SQL models named **SQLForge-LM**. Our experiments demonstrate that SQLForge-LM achieves high performance on well-known benchmarks like Spider (Yu et al., 2018) and BIRD (Li et al., 2024c). Specifically, it attains state-of-the-art (SOTA) performance among existing open-source model-based methods, significantly narrowing the gap with the closed-source model-based methods. Additionally, we evaluate the robust-

ness of SQLForge-LM using datasets meticulously crafted to assess its resistance against perturbations and generalization, including SYN (Gan et al., 2021a), REALISTIC (Deng et al., 2021), and DK (Gan et al., 2021b), where it demonstrates strong and consistent performance. Furthermore, we evaluate SQLForge’s data synthesis capabilities under corner cases, by employing highly intricate SQL statements, to verify its consistent generation of reliable and high-quality data in these challenging edge scenarios. We also deploy and test our data synthesis framework on open-source models, demonstrating its adaptability in different computational environments.

The main contributions of this paper are summarized as follows:

- We propose SQLForge, a framework that incorporates syntax constraints and domain exploration, to drive LLMs to generate reliable and diverse text-to-SQL data.
- We fine-tune a new family of open-source models using the data synthesized by SQLForge, i.e. SQLForge-LM, which achieves SOTA performance across multiple benchmarks among methods based on open-source models.
- Further evaluation shows that SQLForge is capable of handling extreme data generation tasks and achieving the excellent performance

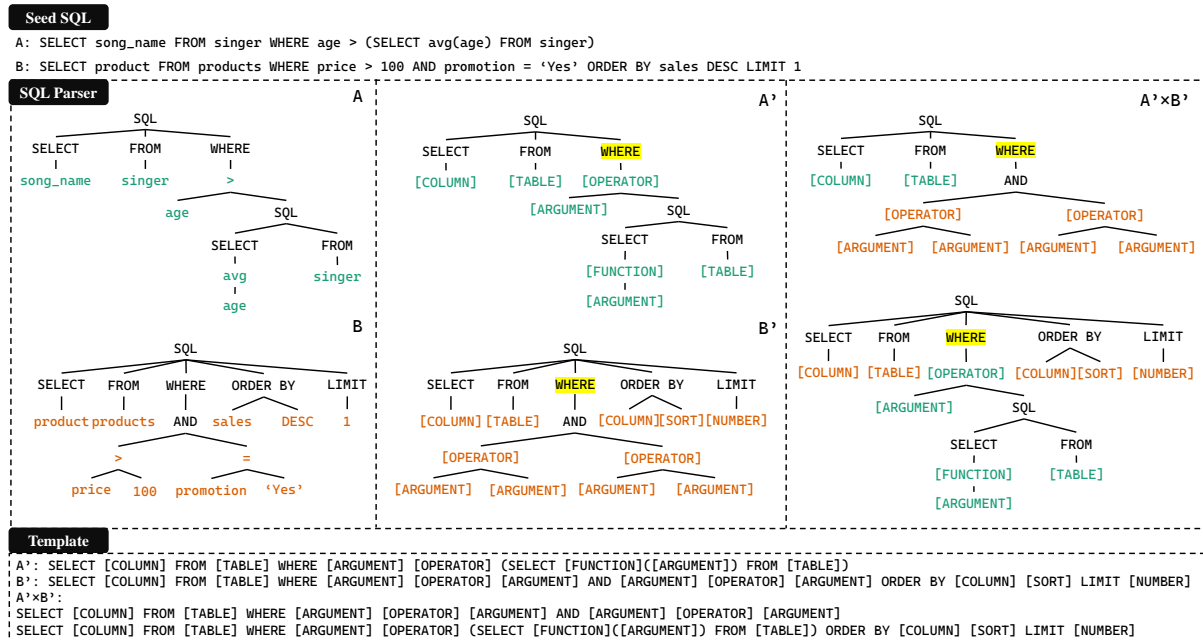


Figure 2: An example of the SQL Parser, which converts SQL into template or generates new template with crossover via AST.

even with open-source base models, significantly enhancing its practical applicability.

## 2 SQLForge

In this section, we present SQLForge, a unified framework designed for generating reliable and diverse text-to-SQL data, which consists of four core components: *SQL Parser*, *SQL Foundry*, *Schema Architect*, and *Question Reverse-Translator*, as illustrated in Fig.1. Unlike existing data synthesis approaches (Li et al., 2024a; Yang et al., 2024b), which directly generate SQL statements from natural language questions, SQLForge reverses such synthesis process. Initially, the SQL Parser extracts SQL templates from seed data as the inherent syntactic structure. Then, the SQL Foundry leverages these templates to generate SQL statements in new domains. Afterwards, the Schema Architect supplements the database schema according to SQL. Finally, the Question Reverse-Translator transforms the SQL into corresponding natural language questions via reverse translation.

**Seed Data** A limited set of seed text-to-SQL data provides the initial corpus for data synthesis. In this paper, the seed data are sourced from the well-known Spider and BIRD datasets, comprising approximately 18K text-to-SQL samples extracted from their training sets. We define this data set as  $\mathcal{D}_{\text{Seed}} = \{(q^i, d^i, sch^i, s^i)\}_{i=1}^n$ , where  $q^i$  repre-

sents the  $i$ -th natural language question,  $d^i$  denotes its domain,  $sch^i$  specifies its database schema,  $s^i$  corresponds to its SQL statement answer, and  $n$  is the total number of samples. The domains from seed data set are added to the explored domain set  $E$ , while the SQL data, referred to as the seed SQL, are represented as  $S_{\text{Seed}}$ .

**SQL Parser** SQL Parser transforms SQL into a structured abstract syntax tree (AST) to standardize the SQL queries while preserving their inherent structure, which converts the seed SQL  $S_{\text{Seed}}$  into the template  $T$ . Different from the method (Kobayashi et al., 2025), which only replaces the columns and table names in SQL statements, the SQL Parser traverses the tree and replaces non-keyword nodes with placeholders. Each placeholder is annotated with a fill type, which describes the role and context of the original node within the query. Additionally, the parser performs crossover operations on sub-trees with the same keywords for enriching SQL structure. In this process, the SQL templates remain the intact grammatical structure of SQL statements, thereby guaranteeing that the synthetic statement is compliant with the SQL specifications throughout the following synthesis process. A working example of the SQL Parser is illustrated in Fig.2.

**SQL Foundry** SQL Foundry instantiates templates  $T$  to generate diverse SQL statements  $S_{\text{Aug}}$

across new domains, which consists of two steps. In the first step, to enrich the domain diversity of generated data, we devise an iterative exploration mechanism that dynamically synthesizes new domain names which are specified to the semantic contexts in database scenario. Specifically, SQL Foundry iteratively generates new domain name in conjunction with auxiliary SQL statement tailored to the domain. The generated domain names are ensured to be outside of the explored domain set  $E$  mentioned in the Seed Data section. By coupling domain name generation with auxiliary SQL construction, we impose a conditional entropy constraint  $H(\alpha|\beta)$  (where  $\alpha$  denotes domain name and  $\beta$  denotes auxiliary SQL) that compresses the domain naming space into a database scenario compatible distribution subspace, thereby systematically eliminating domain drift risks. Following each iteration, the explored domain set is updated with the newly created domain name.

The second step is triggered when exploring enough domains after  $K$  rounds of iteration. At this step, SQL Foundry dynamically binds an SQL template with a domain name derived from  $E$ , and generates syntactically diverse and domain-specific SQL statements  $S_{Aug}$ , which incorporates both the structural diversity and execution reliability of generated SQL statements in previously unexplored domains.

**Schema Architect** Since the newly generated SQL statements belong to entirely new domains, the original database schema is no longer applicable. Schema Architect generates the corresponding database schema expressions  $Sch_{Aug}$  for the newly generated SQL statements  $S_{Aug}$ . Given that SQL statements are highly structured and free of invalid or redundant information, extracting fields and parsing them into a database schema is highly reliable. Furthermore, constraints such as foreign keys, which are explicitly defined in the SQL statements, can be easily incorporated into the corresponding database schema. The above process ultimately produces the augmented schema  $Sch_{Aug}$ .

**Question Reverse-Translator** Question Reverse-Translator transforms the augmented SQL statements  $S_{Aug}$  into their corresponding natural language questions  $Q_{Aug}$ . The previous reverse translation approaches (Hu et al., 2023; Kobayashi et al., 2025) disregard essential database schema information, resulting in ambiguous entity references and structural misinterpretations. Be-

---

**Example of Augmented Data**

```

### Question:
Which are the top 10 campaigns with the highest total number of clicks?
### Schema:
CREATE TABLE Campaigns(
  CampaignID INTEGER PRIMARY KEY,
  CampaignName TEXT
  FOREIGN KEY(CampaignID) REFERENCES Impressions(CampaignID)
);
CREATE TABLE Impressions(
  ImpressionID INTEGER PRIMARY KEY,
  CampaignID INTEGER,
  Clicks INTEGER,
  FOREIGN KEY(CampaignID) REFERENCES Campaigns(CampaignID)
);
### SQL:
SELECT C.CampaignName FROM Campaigns AS C JOIN Impressions AS I
ON C.CampaignID = I.CampaignID GROUP BY C.CampaignName ORDER BY
SUM(I.Clicks) DESC LIMIT 10

```

---

Table 1: An example of Augmented Data, schema provided in the form of CREATE TABLE statements.

sides, their over-reliance on SQL patterns often yields syntactically valid but semantically inconsistent questions that fail to reflect authentic human query patterns. To address these issues, Question Reverse-Translator integrates the associated database schema  $Sch_{Aug}$ . This design enhances the model’s understanding of database topology and entity relationships, enabling alignment between SQL operations and their natural language expressions. The resulting natural language questions  $Q_{Aug}$  exhibit improved semantic fidelity and linguistic naturalness.

We deployed the described pipeline using GPT-4 as the base model, generating a total of 25K text-to-SQL data points in over 1,000 domains, collectively referred to as Augmented Data ( $Q_{Aug}, Sch_{Aug}, S_{Aug}$ ), an example of Augmented Data is presented in Tab.1. By integrating this augmented data with the Spider and BIRD training sets, we constructed a comprehensive dataset  $\mathcal{D}_{Aug}$ , which served as the training set for fine-tuning open-source pre-trained models, such as CodeLlama. For all  $(q^i, sch^i, s^i) \in \mathcal{D}_{Aug}$ , the log likelihood loss function of fine-tuning is defined as:

$$\mathbb{E}_{(q^i, sch^i, s^i) \sim \mathcal{D}_{Aug}} \left[ \sum_{l=1}^L \log p_{\theta}(s_l^i | s_{1:l-1}^i, q^i, sch^i) \right] \quad (1)$$

where  $\theta$  represents the model parameters, and  $L$  is the length of SQL statement. The resulting text-to-SQL reasoning models are named SQLForge-LM.

## 3 Experiments

### 3.1 Experimental Setup

**Datasets** We conduct evaluations on two well-known datasets: Spider (Yu et al., 2018) and BIRD (Li et al., 2024c). Additionally, we also evaluate our models on three robust datasets: SYN (Gan



et al., 2021a), REALISTIC (Deng et al., 2021), and DK (Gan et al., 2021b), to evaluate the robustness and generalization ability of our proposed method. More information about datasets is shown in Appendix B.1.

**Evaluation Metric** We adopt EX (Execution Accuracy, Yu et al., 2018) as the evaluation metric, which measures whether the SQL execution result exactly matches the execution result of provided golden SQL.

**Models** We perform LoRA (Hu et al., 2022) fine-tuning on two model families, including CodeLlama 7B and 13B (Roziere et al., 2023), Qwen2 0.5B and 7B (Yang et al., 2024a). Fine-tuning details are described in Appendix B.2.

**Compared Methods** We compare our approach against four categories of baselines:

**Closed-Source Models** CodeX (Chen et al., 2021), ChatGPT (Ouyang et al., 2022), PaLM (Anil et al., 2023), GPT-4 (Achiam et al., 2023), and Claude-2 (Anthropic, 2023).

**Open-Source Models** Qwen-2 (Yang et al., 2024a), DeepSeek-Coder (Guo et al., 2024), CodeLlama (Roziere et al., 2023), Gemma (Team et al., 2024), StarCoder (Lozhkov et al., 2024), and Llama-3 (Dubey et al., 2024).

**Prompting Closed-Source Models** CHASE-SQL (Pourreza et al., 2024), DAIL-SQL (Gao et al., 2023), DIN-SQL (Pourreza and Rafiei, 2024), PTD-SQL (Luo et al., 2024), and TA-SQL (Qu et al., 2024).

**Fine-tuning Open-Source Models** CodeS (Li et al., 2024b), PICARD (Scholak et al., 2021), RESDSQL (Li et al., 2023), and SENSE (Yang et al., 2024b).

### 3.2 Main Results

Tab.2 presents the EX accuracy of SQLForge-LMs on Spider and BIRD, the results reveal the following observations: (1) Among open-source models, SQLForge-LM achieves state-of-the-art (SOTA) EX accuracy, SQLForge-LM (CodeLlama-13B variant) is 2.2% ahead of SENSE-13B in average performance, even when SQLForge-LM uses the LoRA fine-tuning and SENSE uses full parameter fine-tuning. (2) SQLForge-LM demonstrates high performance across models with varying parameter

scales. In the case of CodeLlama, the average performance of the 7B and 13B models increased by 26.4% and 27.0% respectively, showcasing strong parametric scalability. (3) The model consistently performs well across different model families, highlighting its transferability. (4) SQLForge-LM not only surpasses many closed-source models but also reduces the gap between open-source model and prompting closed-source approaches. SQLForge-LM achieves performs comparable to most prompting closed-source approaches, although a significant performance gap remains with CHASE-SQL on BIRD. The performance reported in Tab.2 are all obtained using greedy decoding.

Additionally, as shown in Tab.3, we evaluate the performance of SQLForge-LM on the Robust datasets (SYN, REALISTIC, DK). SQLForge-LM secures a leading position across these datasets without extra training. Using the same base model, SQLForge-LM-7B and SQLForge-LM-13B outperform SENSE-7B and SENSE-13B by an average of 1.1% and 1.3%, respectively, showcasing strong resilience to perturbations and exceptional generalization capabilities.

### 3.3 Ablation Study

The following section details the ablation studies conducted in this work, all of which utilize CodeLlama-7B as the base model.

**Analysis of Data Composition** We investigate the impact of different data components on model performance by training the model with various combinations of datasets, as shown in Tab.4. When using only the Spider training set, the model’s performance on Spider improves significantly (15.0%). Similarly, using the BIRD training set leads to notable improvements on BIRD (25.6%). These results are consistent with the inherent characteristics of the datasets. Compared to Spider, BIRD is a more complex dataset from a distinct domain. Notably, data from disjoint domains, such as BIRD and Spider, can still complement each other (6.7% on Spider, 13.1% on BIRD), indicating that diverse domain data can stimulate the model’s implicit domain adaptability. Because our augmented data not only cover both simple and complex SQL generation tasks but also expands the domain scope substantially. When adding augmented data, the model’s performance is significantly improved on both datasets (22.5% on Spider, 33.4% on BIRD), demonstrating the effectiveness of our data aug-

Model	Base	Size	# Calls	Spider		BIRD Dev	Average
				Dev	Test		
<i>Closed-Source Models</i>							
CodeX	-	175B	✗	71.8	-	34.4	-
ChatGPT	-	-	✗	72.3	-	37.2	-
PaLM-2	-	-	✗	-	-	27.4	-
GPT-4	-	-	✗	72.9	-	46.4	-
Claude-2	-	-	✗	-	-	42.7	-
<i>Open-Source Models</i>							
Qwen-2	-	0.5B	✗	44.8	43.1	14.5	34.1
DeepSeek-Coder	-	1.3B	✗	59.7	58.4	22.5	46.9
CodeLlama	-	7B	✗	61.9	62.6	23.5	49.3
Gemma	-	7B	✗	50.1	50.4	21.2	40.6
StarCoder	-	7B	✗	61.5	62.1	21.0	48.2
Qwen-2	-	7B	✗	52.6	55.7	20.1	42.8
CodeLlama	-	13B	✗	63.5	64.4	23.9	50.6
Llama-3	-	70B	✗	67.4	68.1	30.6	55.4
<i>Prompting Closed-Source Models</i>							
CHASE-SQL	Gemini-1.5	-	✓	-	<b>87.6</b>	<b>73.0</b>	-
DAIL-SQL	GPT-4	-	✓	83.5	86.6	54.8	75.0
DIN-SQL	GPT-4	-	✓	82.9	85.3	50.7	73.0
PTD-SQL	GPT-4	-	✓	<b>85.7</b>	-	57.0	-
TA-SQL	GPT-4	-	✓	<u>85.0</u>	-	56.2	-
<i>Fine-tuning Open-Source Models</i>							
PICARD	T5	3B	✗	79.3	75.1	-	-
RESDSQL	RoBERTa	3B	✓	84.1	79.9	-	-
SENSE	CodeLlama	13B	✗	84.1	86.6	55.5	75.4
CodeS	StarCoder	15B	✗	84.9	-	58.5	-
<i>Ours</i>							
SQLForge-LM	Qwen2	0.5B	✗	76.1	75.3	36.9	62.8
SQLForge-LM	Qwen2	7B	✗	82.5	82.9	54.1	73.2
SQLForge-LM	CodeLlama	7B	✗	84.4	85.8	56.9	<u>75.7</u>
SQLForge-LM	CodeLlama	13B	✗	<b>85.7</b>	<u>87.4</u>	<u>59.8</u>	<b>77.6</b>

Table 2: Performance comparison on Spider and BIRD. # Calls indicates whether the method needs to call the model multiple times.

Model	Base	# Calls	SYN	REALISTIC	DK	Average
RESDSQL	RoBERTa-3B	✓	76.9	81.9	66.0	74.9
SENSE	CodeLlama-7B	✗	72.6	82.7	77.9	77.7
SENSE	CodeLlama-13B	✗	<u>77.6</u>	<u>84.1</u>	<u>80.2</u>	<u>80.6</u>
CodeS	StarCoder-15B	✗	77.0	83.1	70.7	76.9
SQLForge-LM	CodeLlama-7B	✗	74.6	83.3	78.9	78.8
SQLForge-LM	CodeLlama-13B	✗	<b>78.9</b>	<b>84.8</b>	<b>82.1</b>	<b>81.9</b>

Table 3: Performance comparison on robust datasets. # Calls indicates whether the method needs to call the model multiple times.

mentation strategy.

**Analysis of Scaling of Augmented Data** We analyze the scaling performance of the augmented data.

Specifically, we train the model using  $0\times$ ,  $1/8\times$ ,  $1/4\times$ ,  $1/2\times$  and  $1\times$  augmented data. The results are shown in Fig.3. As the amount of augmented

Data			Spider Dev	BIRD Dev	Average
Spider	BIRD	Augmented Data			
✓	✗	✗	76.9	36.6	56.8
✗	✓	✗	68.6	49.1	58.9
✓	✓	✗	79.8	50.5	65.2
✓	✓	✓	84.4	56.9	70.7

Table 4: Effect of different data composition with CodeLlama-7B as the base model.

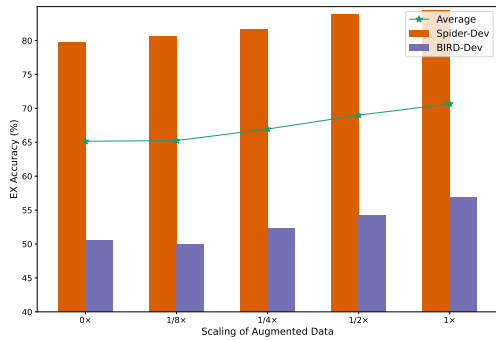


Figure 3: Effect of different scaling of augmented data with CodeLlama-7B as the base model.

data increases, the model’s performance consistently improves, demonstrating excellent scaling performance of our augmented data.

**Analysis of Data Augmentation methods** We compare our data augmentation approach with direct data augmentation (Yang et al., 2024b) and evaluate the impact of our component designs. Direct data augmentation generates new data following the sequence from schema to question, and ultimately to SQL. Additionally, we assess the effects of auxiliary SQL and schema enhancement. As shown in Tab.5, our method outperforms existing data synthesis processes, and both auxiliary SQL and schema enhancements contribute to performance improvements.

**Analysis of Augmented Data Diversity** We utilize the T5 model (Raffel et al., 2020) to generate embedding vectors from the Spider training set, BIRD training set, and augmented data, which are then projected into a two-dimensional space using t-SNE. As shown in Fig.4, our augmented data effectively fills and expands the distribution within the semantic space. Additional details on the augmented data can be found in Fig.5.

### 3.4 Robustness Study

We evaluate the robustness of SQLForge by examining its scalability and adaptability. Scalability

Augmentation Method	Spider Dev	BIRD Dev	Average
<b>Direct Data Augmentation</b>	82.9	52.5	67.7 <sub>(-3.0)</sub>
<b>SQLForge w/o Auxiliary SQL</b>	84.2	55.8	70.0 <sub>(-0.7)</sub>
<b>SQLForge w/o Schema</b>	83.8	53.9	68.9 <sub>(-1.8)</sub>
<b>SQLForge (Ours)</b>	84.4	56.9	70.7

Table 5: Comparison of different data augmentation methods. **SQLForge w/o Auxiliary SQL** presents don’t generate auxiliary SQL in Domain Exploration stage. **SQLForge w/o Schema** presents don’t use schema to Enhance inputs.

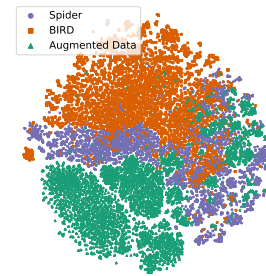


Figure 4: 2-D t-SNE illustrates the distribution of seed and augmented data in the semantic space.

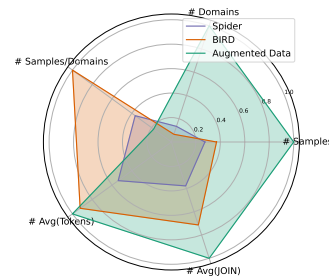


Figure 5: Detailed comparison between seed data and augmented data.

measures SQLForge’s performance with larger or more complex inputs, while adaptability measures its stability on open-source models. This analysis demonstrates SQLForge’s potential in both complex real-world scenarios and high privacy-sensitive scenarios.

**Scalability** We employ SQLForge and direct data augmentation approach to produce text-to-SQL data of varying complexities. The complexity of SQL is quantified by token length. We use GPT-4 to generate 1K data samples for each token length category (50, 100, 150, 200), and subsequently assess the executable rate of the generated data. The results are illustrated in Fig.6. As the complexity of the generated data progressively increases, the executable rate of the data produced through direct data augmentation experiences a significant decline,

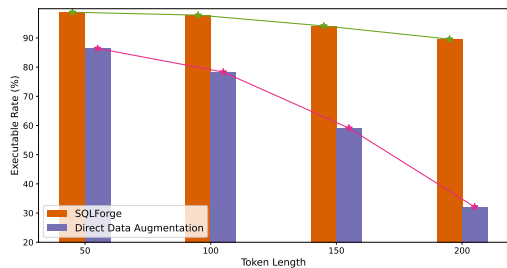


Figure 6: Executable rate of augmented data under different complexity.

whereas the rate for data generated by SQLForge exhibits only a marginal decrease.

**Adaptability** We select Llama-3-70B, Qwen-2-72B as the open-source base models and ChatGPT, GPT-4 as the closed-source base model. Utilizing both SQLForge and direct data augmentation, we generate text-to-SQL data comprising 1K samples (The token length of SQL is 50). The results are presented in Fig.7. When transitioning to the open-source model, SQLForge maintains stable performance, whereas direct data augmentation exhibits significant fluctuations as the model’s capabilities diminish.

## 4 Related Work

**Text-to-SQL Reasoning Methods** In the field of text-to-SQL, early approaches like IRNET (Guo et al., 2019) relied on attention-based models to learn intermediate representations. Subsequently, the focus shifted to fine-tuning pre-trained models, with works such as RESDSQL (Li et al., 2023), RAT-SQL (Wang et al., 2020), and PICARD (Scholak et al., 2021). The landscape of text-to-SQL research has been further revolutionized by the advent of large language models (LLMs), which have introduced powerful zero-shot and in-context learning capabilities. This advancement has led to the widespread adoption of prompt-based methods using closed-source models. Notable contributions in this area include DAIL-SQL (Gao et al., 2023), which integrated various prompt engineering techniques to enhance performance, DIN-SQL (Pourreza and Rafiei, 2024), which decomposed complex tasks into manageable sub-tasks, PTD-SQL (Luo et al., 2024), which employed query group partitioning to enable LLMs to focus on specific problem types, TA-SQL (Qu et al., 2024), which utilized task alignment to mitigate hallucinations at each stage of the reasoning process, CHASE-SQL

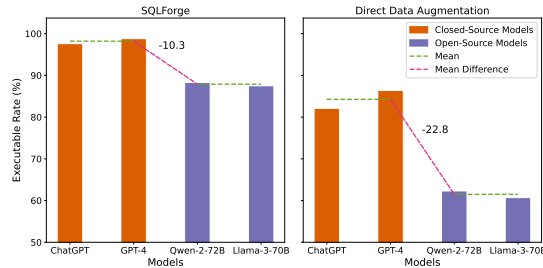


Figure 7: Executable rate of augmented data for open-source and closed-source models.

(Pourreza et al., 2024), which leveraged LLMs’ intrinsic knowledge to generate diverse and high-quality SQL candidates using different LLM generators. Additionally, some works have focused on training or fine-tuning LLMs specific to text-to-SQL realm, such as SENSE (Yang et al., 2024b) and CodeS (Li et al., 2024b). Our work fine-tunes open-source LLMs through reliable and diverse data augmentation to enhance text-to-SQL reasoning.

**Data Augmentation for Text-to-SQL** To increase the quantity of text-to-SQL data, many existing approaches (Hu et al., 2024; Li et al., 2024a) typically augment data directly based on existing databases. Li et al. (2024b) generates data for specific domain using bi-directional augmentation with manual annotation. Yang et al. (2024b) synthesizes data for new domains through direct data augmentation. In contrast, our work reverses the data augmentation process. By taking advantage of the syntax constraints inherent in the template and employing an iterative domain exploration mechanism, we generate a large volume of reliable and diverse SQL data automatically.

## 5 Conclusion

In this study, we present SQLForge, a comprehensive framework designed to synthesize high-quality data for enhancing text-to-SQL reasoning. Experiment demonstrates that SQLForge-LM achieves SOTA performance in open-source model-based methods on well-known benchmarks, significantly narrowing the performance gap between open-source and closed-source models. Additionally, the framework exhibits robustness and transferability, offering valuable insights into the development of text-to-SQL models and inspiration for other reasoning tasks.



## Limitations

Computational resource limitations restrict our ability to fine-tune larger models, leaving the performance implications of data synthesis largely unexplored in this context. Additionally, a comparative analysis between full parameter fine-tuning and LoRA fine-tuning is necessary to establish best practices. While our data generation framework has demonstrated strong performance with closed-source models like GPT-4 and shows adaptability with open-source models such as Llama-3-70B, considerations of cost efficiency and privacy drive the need for further exploration of text-to-SQL data generation with open-source models. Unfortunately, resource constraints prevent us from investigating the data generation performance of fine-tuned open-source models.

## Ethics Statement

Our work aims to provide a low-cost solution for text-to-SQL scenarios by enhancing open-source models using synthetic data. However, like any language model, it may generate unrealistic or even harmful content. We strongly encourage users to review the outputs when utilizing SQLForge and SQLForge-LM. Additionally, our research leverages open-source models such as CodeLlama and Qwen-2, as well as software frameworks like PyTorch and HuggingFace. We adhere to the policies and licenses of these resources and acknowledge their significant contributions to our work.

## Acknowledgement

This work was financially supported by National Key Research and Development Program of China, No.2024YDLN0004.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Anthropic. 2023. Introducing claude. <https://www.anthropic.com/news/introducing-claude>.

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Xiang Deng, Ahmed Hassan, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. 2021. Structure-grounded pretraining for text-to-sql. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1337–1350.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R Woodward, Jinxia Xie, and Pengsheng Huang. 2021a. Towards robustness of text-to-sql models against synonym substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2505–2515.
- Yujian Gan, Xinyun Chen, and Matthew Purver. 2021b. Exploring underexplored limitations of cross-domain text-to-sql generalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8926–8931.
- Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. Text-to-sql empowered by large language models: A benchmark evaluation. *arXiv preprint arXiv:2308.15363*.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.
- Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4524–4535.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Yaojie Hu, Ilias Fountalis, Jin Tian, and Nikolaos Vasiloglou. 2024. Annotatedtables: A large tabular dataset with language model annotations. *arXiv preprint arXiv:2406.16349*.

- Yiqun Hu, Yiyun Zhao, Jiarong Jiang, Wuwei Lan, Henghui Zhu, Anuj Chauhan, Alexander Hanbo Li, Lin Pan, Jun Wang, Chung-Wei Hang, et al. 2023. Importance of synthesizing high-quality data for text-to-sql parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1327–1343.
- Hideo Kobayashi, Wuwei Lan, Peng Shi, Shuaichen Chang, Jiang Guo, Henghui Zhu, Zhiguo Wang, and Patrick Ng. 2025. You only read once (YORO): Learning to internalize database knowledge for text-to-SQL. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1889–1901.
- Chunhui Li, Yifan Wang, Zhen Wu, Zhen Yu, Fei Zhao, Shujian Huang, and Xinyu Dai. 2024a. Multisql: A schema-integrated context-dependent text2sql dataset with diverse sql operations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13857–13867.
- Haoyang Li, Jing Zhang, Cuiping Li, and Hong Chen. 2023. Resdsq: Decoupling schema linking and skeleton parsing for text-to-sql. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13067–13075.
- Haoyang Li, Jing Zhang, Hanbing Liu, Ju Fan, Xiaokang Zhang, Jun Zhu, Renjie Wei, Hongyan Pan, Cuiping Li, and Hong Chen. 2024b. Codes: Towards building open-source language models for text-to-sql. *Proceedings of the ACM on Management of Data*, 2(3):1–28.
- Jinyang Li, Binyuan Hui, Ge Qu, Jiayi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. 2024c. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36.
- Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2024. A survey of nl2sql with large language models: Where are we, and where are we going? *arXiv preprint arXiv:2408.05109*.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.
- Ruilin Luo, Liyuan Wang, Binghui Lin, Zicheng Lin, and Yujiu Yang. 2024. Ptd-sql: Partitioning and targeted drilling with llms in text-to-sql. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3767–3799.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan O Arik. 2024. Chase-sql: Multi-path reasoning and preference optimized candidate selection in text-to-sql. *arXiv preprint arXiv:2410.01943*.
- Mohammadreza Pourreza and Davood Rafiei. 2024. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *Advances in Neural Information Processing Systems*, 36.
- Ge Qu, Jinyang Li, Bowen Li, Bowen Qin, Nan Huo, Chenhao Ma, and Reynold Cheng. 2024. Before generation, align it! a novel and effective strategy for mitigating hallucinations in text-to-SQL generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5456–5471.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. Picard: Parsing incrementally for constrained auto-regressive decoding from language models. *arXiv preprint arXiv:2109.05093*.
- Liang Shi, Zhengju Tang, Nan Zhang, Xiaotong Zhang, and Zhi Yang. 2024. A survey on employing large language models for text-to-sql tasks. *arXiv preprint arXiv:2407.15186*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiayi Yang, Binyuan Hui, Min Yang, Jian Yang, Junyang Lin, and Chang Zhou. 2024b. Synthesizing text-to-sql data from weak and strong llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 7864–7875.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921.

## A Prompt Design

### A.1 Prompt of SQL Foundry

Tab.6 presents the prompt format used in Domain Exploration, Tab.7 presents the prompt format used in Domain-Specific SQL Generation.

### A.2 Prompt of Schema Architect

Tab.8 presents the prompt format used in Schema Architect.

### A.3 Prompt of Question Reverse-Translator

Tab.9 presents the prompt format used in Question Reverse-Translator.

## B Experiment Details

### B.1 Datasets Details

Spider consists of 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables covering 138 different domains, while BIRD contains over 12,751 unique question-SQL pairs, 95 big databases with a total size of 33.4 GB. It also covers more than 37 professional domains, such as blockchain, hockey, healthcare and education, etc, SYN contains 7,000 training and 1,034 development examples with replacing simple string-matched problem tags or pattern names with their synonyms, REALISTIC contains 508 text-to-SQL pairs from Spider, replaced mentioned schema items in questions to make them closer to realworld scenarios, DK contains 535 text-to-SQL pairs drawn from the Spider development set, where 270 pairs are the same as the original Spider samples, while the rest 265 pairs are modified to incorporate the domain knowledge.

### B.2 Fine-tuning Details

Our experiments are conducted using the HuggingFace library and leverage 6 NVIDIA A100 40GB GPUs. We employ the AdamW optimizer with a learning rate of  $2e^{-4}$  and a cosine learning rate scheduler, the warm-up phase accounts for 1% of the total training steps. For the LoRA adaptation, we set the rank  $r$  to 128 and the scaling factor  $\alpha$  to 256, applying LoRA modules to all optional target modules in the models.

---

**Prompt of Domain Exploration**

---

Your task is to generate a new domain and SQLite based on the provided template.

Requirements:

1. Domain: Avoid domains listed below and ensure diversity by exploring new or under-represented domains.
  - Existing domains: {domain\_explored}
2. Template: {template}
  - The "[ ]" placeholders should be filled independently and meaningfully based on their context. These placeholders do not need to be identical, even if they appear multiple times in the template.
  - The content of each placeholder should align with the intent of the query and the domain.
3. Guidelines:
  - Strictly adhere to the template structure.
  - Ensure the SQLite statement is meaningful.
  - Limit the domain name to {domain\_count} word(s).
  - Refrain from adding unrelated content or remarks.

(examples goes here...)

---

Table 6: Prompt of Domain Exploration, "{domain\_explored}" is replaced with the explored domain set, "{template}" is replaced with the template from  $T$ , "{domain\_count}" is chosen from {1,2,3}.

---

**Prompt of Domain-Specific SQL Generation**

---

Your task is to generate a SQLite query based on the following template and domain.

Requirements:

1. Domain: {domain}
2. Template: {template}
  - The "[ ]" placeholders should be filled independently and meaningfully based on their context. These placeholders do not need to be identical, even if they appear multiple times in the template.
  - The content of each placeholder should align with the intent of the query and the domain.
3. Guidelines:
  - Strictly adhere to the template structure.
  - Ensure the SQLite query is meaningful, logically coherent, and specific to the domain.
  - Provide only the SQLite query without any additional explanations or comments.

(examples goes here...)

---

Table 7: Prompt of Domain-Specific SQL Generation, "{domain}" is chosen from the explored domain set, "{template}" is replaced with the template from  $T$ .

---

**Prompt of Schema Architect**

---

Your task is to generate a schema (CREATE TABLE statements) based on the given SQLite and domain.

Requirements:

1. Domain: {domain}
2. SQLite: {sql}
3. Guidelines:
  - Strictly adhere to the column names and types as implied in the SQLite statement.
  - Ensure that foreign keys are correctly represented without using ALTER TABLE.
  - Follow the output format strictly, including table names, column names, types, and foreign key constraints.
  - Do not add any unrelated content or explanations.

(examples goes here...)

---

Table 8: Prompt of Schema Architect, "{sql}" is replaced with the SQL statement from  $S_{Aug}$ , "{domain}" is its domain.

---

**Prompt of Question Reverse-Translator**

---

Your task is to generate a human-readable, natural language question based on the following SQLite query and the schema.

SQLite: {sql}

schema: {schema}

Guidelines:

- Focus on the query's intent, not just repeating table or column names.
- Keep the question clear, concise, and intuitive, following natural language patterns.
- For aggregations (e.g., COUNT, SUM), joins, or filters, ask about the result or insight, not the structure.
- Ensure the question is contextually relevant to the schema and uses table/column names meaningfully.
- Avoid unnecessary content or comments.

(examples goes here...)

---

Table 9: Prompt of Question Reverse-Translator, "{sql}" is replaced with the SQL statement from  $S_{Aug}$ , "{schema}" is its schema.