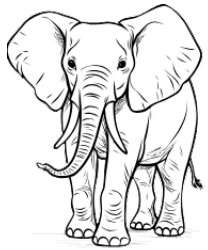


ACL 2024

Bangkok, Thailand

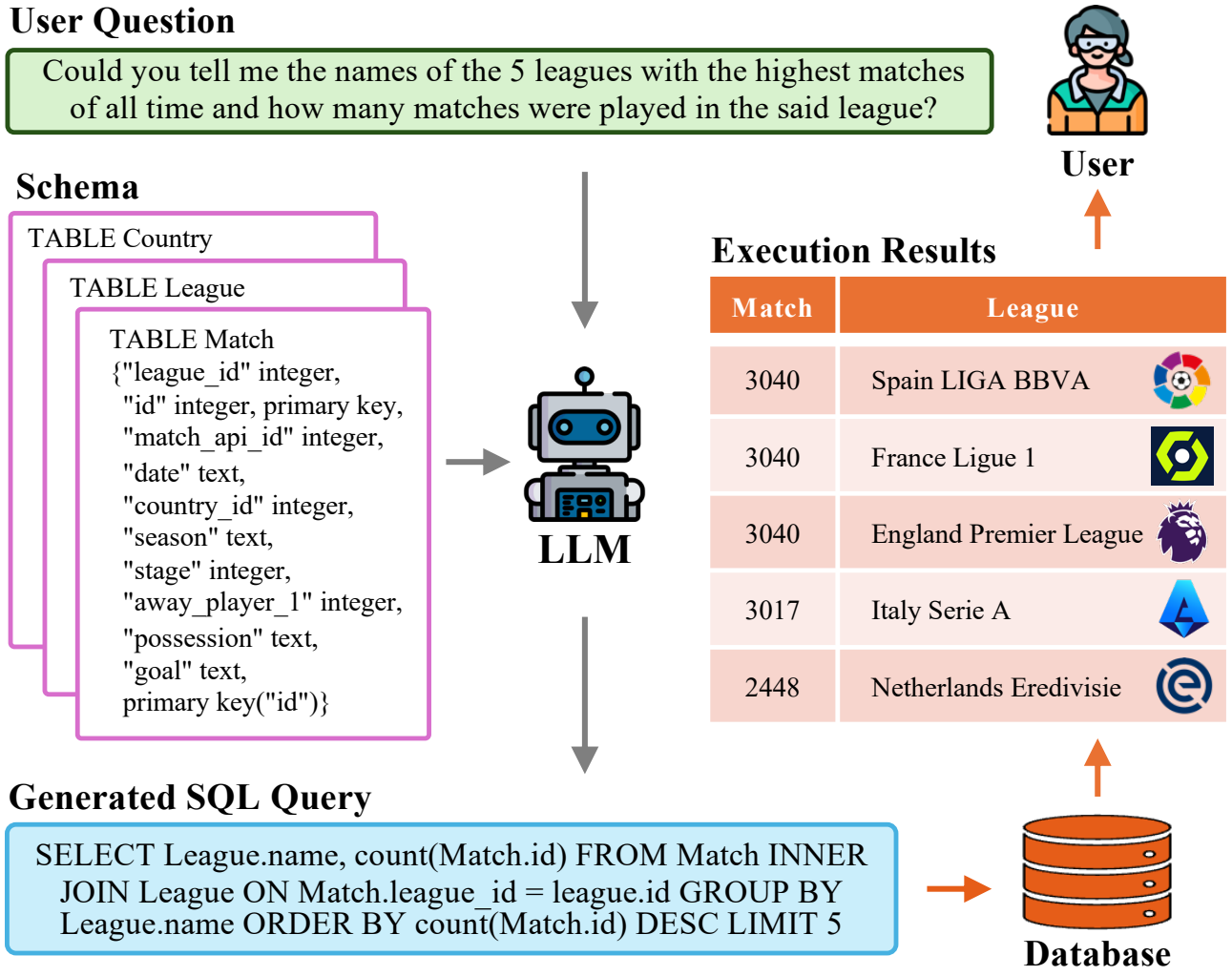
Knowledge-to-SQL: Enhancing SQL Generation with Data Expert LLM

Zijin Hong, Zheng Yuan, Hao Chen, Qinggang Zhang, Feiran Huang, Xiao Huang
Jinan University, The Hong Kong Polytechnic University

暨南大學
JINAN UNIVERSITY

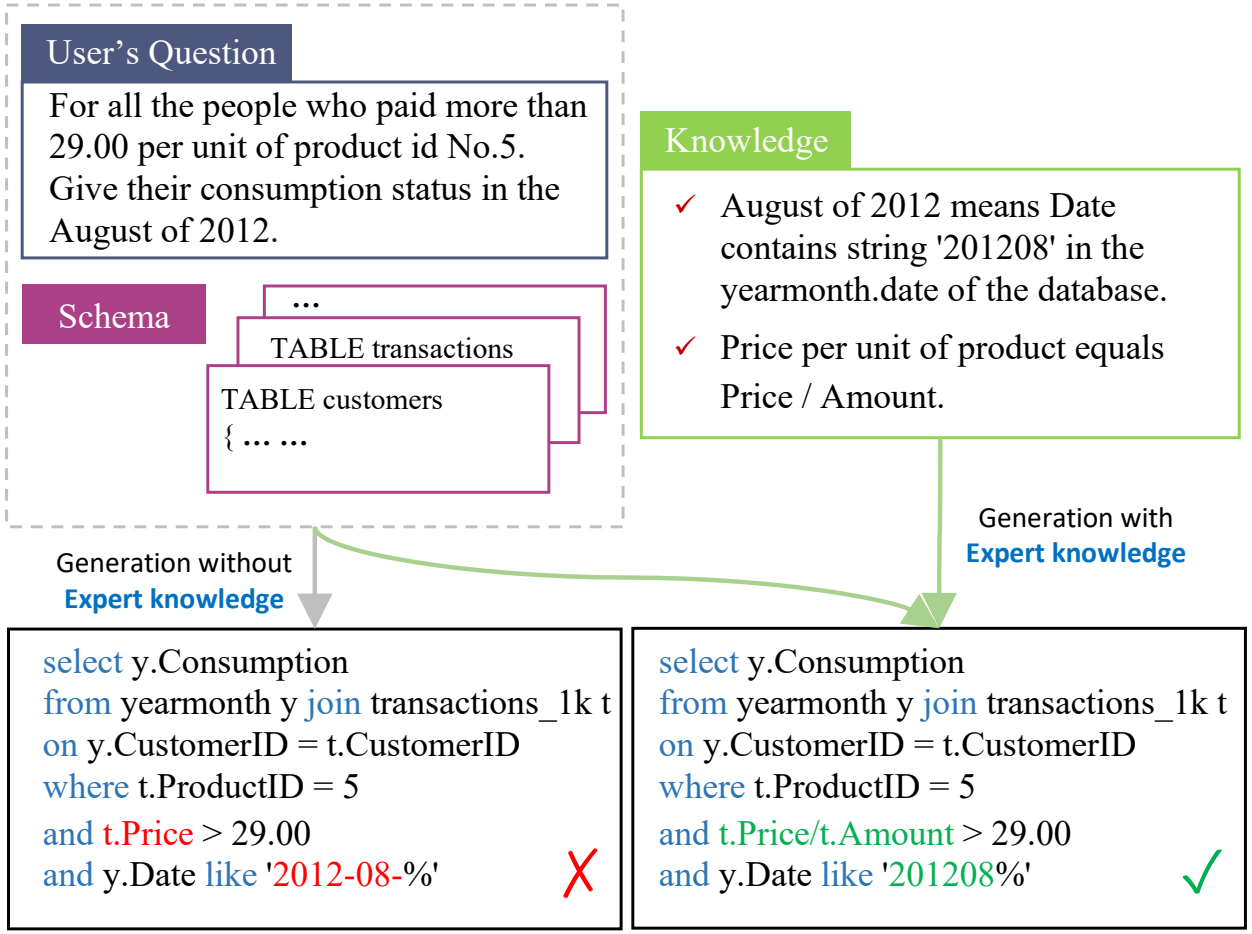
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Introduction of Text-to-SQL



- Firstly, a user proposes a question that queries some contents stored in the target database, seeking specific information based on the available data.
- Then, the schema of the database will be combined with the user question as the input for the text-to-SQL model.
- Taking the input, the model is required to understand the intent of the question and the structure of the target database to generate a corresponding SQL query.
- The generated SQL query will be executed in the target database, and the retrieved content is expected to answer the question.

Challenges & Motivations



Challenges

- With the advancement of big data technology, database structures have become increasingly complex, often requiring domain-specific knowledge for answering a given question.
- Necessary knowledge is sometimes not explicitly contained in questions or schemas, nor is it always learned by LLMs, which negatively impacts the performance and robustness of text-to-SQL systems.

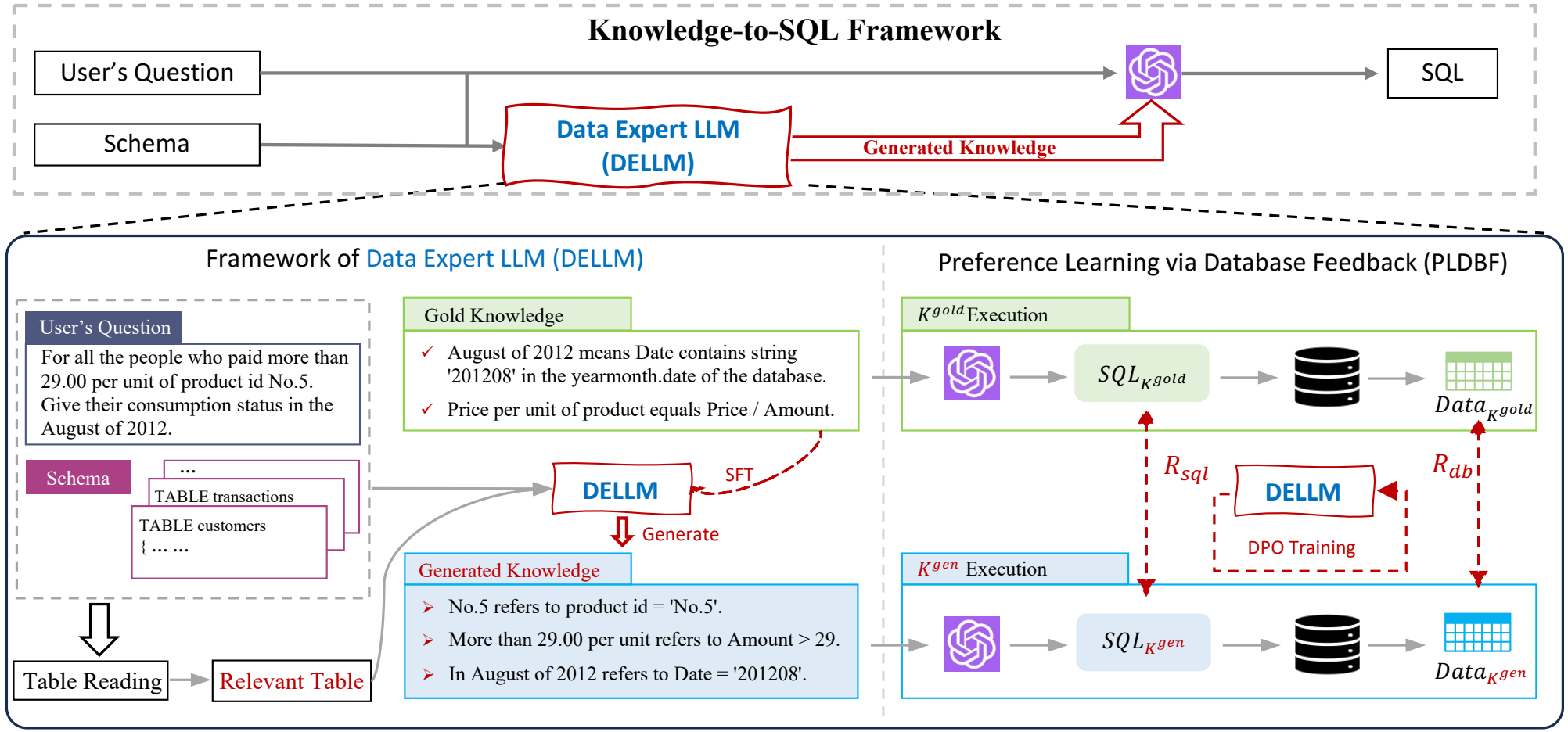
Motivations

- Providing helpful knowledge that explains components in the question with database content can enhance LLMs' understanding.
- Designing a system that can automatically generate the required knowledge (so-called "**expert knowledge**") for various questions can improve overall accuracy.

Our Contributions

- We highlight the significance of expert knowledge and present the knowledge-to-SQL framework for improving SQL generation.
- We introduce a well-designed Data Expert Large Language Model (DELLM), along with customized structure, fine-tuning technique, and preference-tuning training strategies.
- We validate the effectiveness of our approach on the BIRD and Spider datasets, demonstrating that DELLM can generally enhance the performance of common LLM-based text-to-SQL implementations.

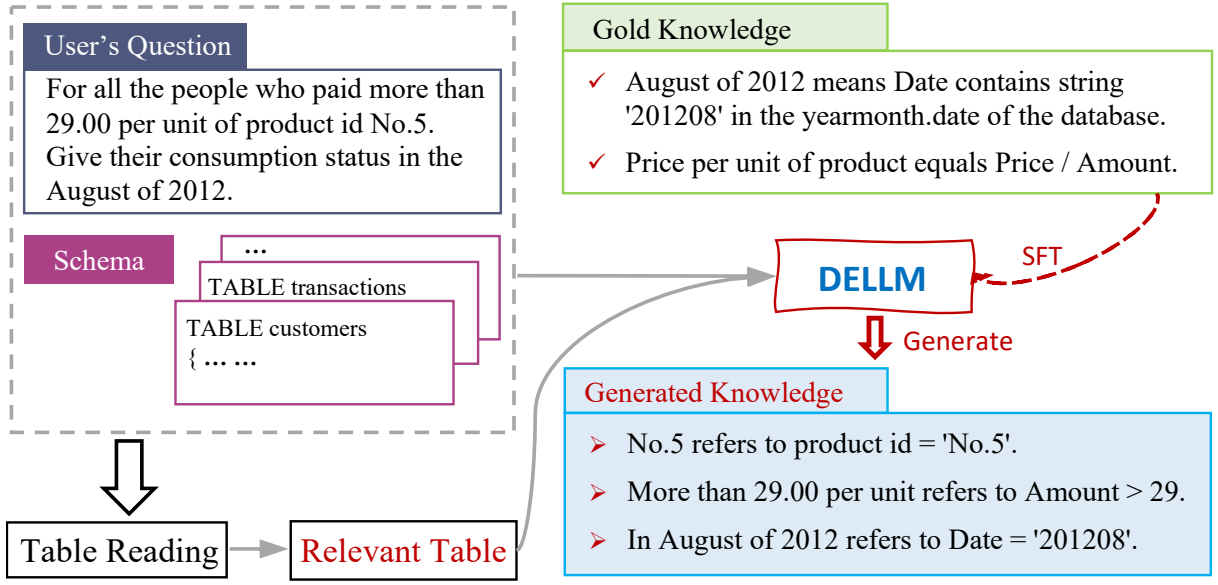
Proposed Method



- We first train a Data Expert LLM (DELLM) to generate helpful knowledge based on the provided input
- Then, the generated knowledge is combined with the user question and database schema to assist in SQL generation.

Training Details of Data Expert LLM

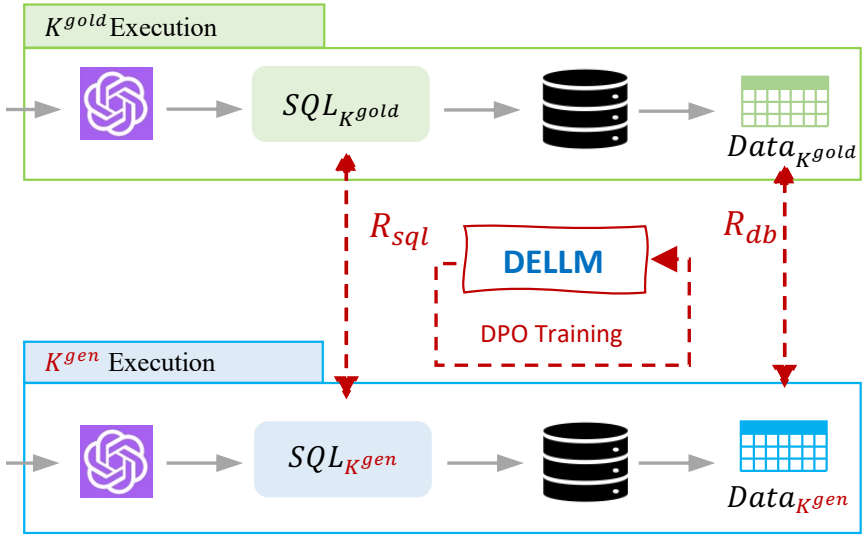
➢Supervised Fine-tuning



- To enhance the model understanding, the input of SFT training incorporates the relevant table T for the user question Q in the corresponding database S.
- Given human-annotated knowledge K^{gold} as the training label, the objective function of SFT is:

$$\mathcal{L}_{SFT} = -\log \Pr(K^{gold} \mid Q, S, T) = -\sum K^{gold} \log(K^{gen})$$

➢Preference Learning



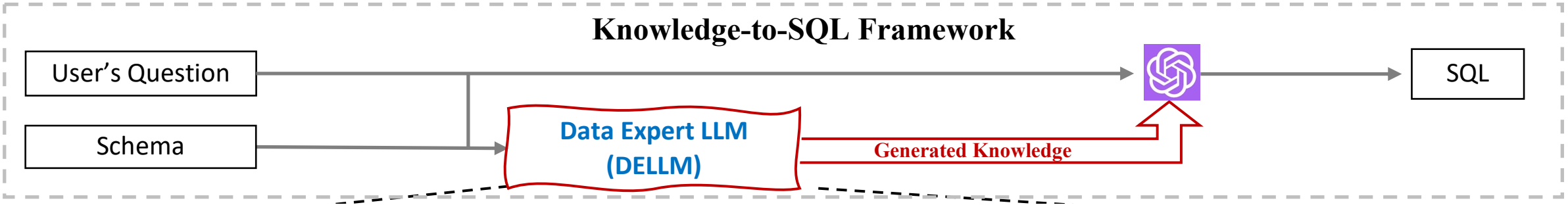
- We obtain the preference knowledge pairs K_w, K_l through indicator χ_{db} that evaluates the execution results and χ_{sql} that measures the contribution of the knowledge, with executed results V^{gold} and V^{gen} :

$$\mathcal{P}^{db}_{\{K_w,K_l\}} = \{K^{gold}, K^{gen} \mid \chi_{db}(V^{gold}, V^{gen}) = 0\}$$
$$\mathcal{P}^{sql}_{\{K_w,K_l\}} = \{K^{gold}, K^{gen} \mid \chi_{sql}(K^{gold}, Y) = 1, \chi_{sql}(K^{gen}, Y) = 0\}$$

- Then, the DPO training is conducted as further preference learning refinement:

$$\mathcal{L}_{\mathrm{PL}}(\pi^{DPO}; \pi^{SFT}) = -\mathbb{E}_{\pi}[\log \sigma(\beta R(K_w) - \beta R(K_l))]$$

Knowledge-to-SQL



- Given a PL-refined DELLM, the user question and the database schema will be formulated as the input, the DELLM with output an expert knowledge regrading the question and corresponding database.
- The generated knowledge will be appended to the question and schema to prompt an off-the-shelf LLMs (e.g. GPT-4, Claude-2) to generate SQL query.
- With the help of the knowledge, the generated SQL will be more accurate.

Experiments

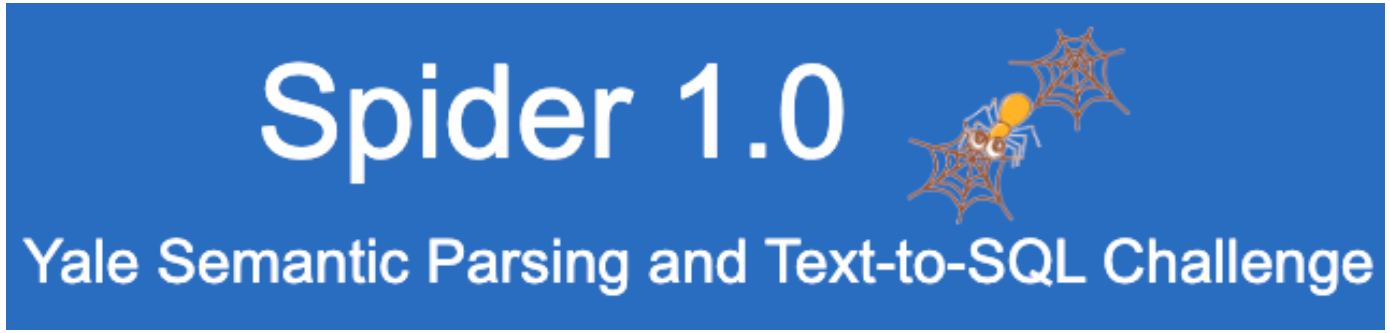
- **Main Result**: We compare the originally predicted SQL with the predicted SQL incorporating the generated expert knowledge.
- **Ablation Study**: By predicting SQL using the knowledge generated by variations of DELLM, we discuss the influence of different modules.
- **Comprehensive Evaluation:** We discuss the effectiveness of DELLM in different difficulty levels and make a comparison with human annotations.
- **Performance on Partial Training Data:** We verify the robustness of our proposed method on data scarcity.

Benchmarks

➢BIRD

J Li et al. Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs, NeurIPS 2023

➢Spider



T Yu et al. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task, EMNLP 2018

Main Results & Ablation Study

Main Results

	Models	EX		VES	
		w/o knowledge	w/ DELLM	w/o knowledge	w/ DELLM
BIRD	T5-3B	10.37	16.68 (+6.31)	13.62	20.84 (+7.22)
	GPT-3.5-Turbo	27.64	33.31 (+5.67)	28.64	36.12 (+7.48)
	GPT-4	33.25	37.94 (+4.69)	35.92	42.15 (+6.23)
	Claude-2	30.05	35.53 (+5.48)	32.97	39.71 (+6.74)
	GPT-3.5-Turbo + CoT	27.25	32.79 (+5.54)	29.16	35.51 (+6.35)
	DAIL-SQL + GPT-4	40.89	45.81 (+4.92)	45.13	51.59 (+6.46)
	MAC-SQL + GPT-4	43.65	48.92 (+5.27)	48.07	54.78 (+6.71)
Spider	GPT-3.5-Turbo	67.89	69.60 (+1.71)	68.33	70.16 (+1.83)
	GPT-4	70.02	71.68 (+1.66)	71.03	72.82 (+1.79)

Table 2: Experimental results for text-to-SQL on different benchmarks with and without knowledge generated by our proposed DELLM. The number in the bracket denotes the improvement in execution accuracy (EX) and valid efficiency score (VES) brought by DELLM's knowledge compared to the baseline performance without knowledge.

- The knowledge generated by DELLM can assist SQL generation for different LLMs.
- This knowledge is also helpful for advanced prompting techniques, assisting the SQL generation process by providing necessary knowledge.

Ablation Study

Models	EX	VES
GPT-4 + DELLM	37.94	42.15
w/o *table reading*	37.23 (-0.71)	41.30 (-0.85)
w/o *db feedback*	36.25 (-1.69)	40.46 (-2.07)
w/o *sql feedback*	36.91 (-1.03)	41.12 (-1.55)

Table 3: Ablation study on variations of DELLM.

Models	EX	VES
GPT-4	33.25 ± 0.61	35.92 ± 1.03
+ DELLM$_{PPO}$	35.28 ± 1.18	40.03 ± 1.81
+ DELLM$_{DPO}$	37.94 ± 0.57	42.15 ± 0.95

Table 4: Ablation study on the utilized PL algorithms.

- All modules are important for the proposed framework.
- The PPO algorithm also works for our preference learning, though with less stability.

Further Analysis

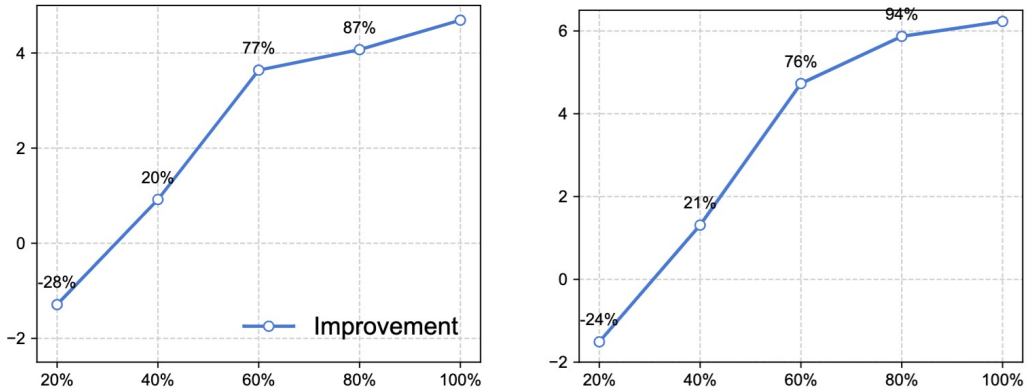
➢Comprehensive Evaluation

Model	Simp.	Mod.	Chall.	All
GPT-3.5-Turbo	35.58	14.60	17.61	27.64
GPT-3.5-Turbo + D	43.09	18.30	17.61	33.31
GPT-3.5-Turbo + E	50.27	31.81	20.42	41.98
GPT-4	41.05	21.13	21.13	33.25
GPT-4 + D	47.16	24.18	21.83	37.94
GPT-4 + E	54.01	36.38	31.69	46.67

Table 5: The execution accuracy (EX) for questions with different difficulty. Simp. denotes Simple, Mod. denotes Moderata, and Chall denotes Challenging.

- ❑ DELLM mainly works on simple and moderate questions.
- ❑ The performance of DELLM still has a significant gap compared to human experts.

➢Performance on Partial Training Data



(a) Execution Accuracy (b) Valid Efficiency Score

Figure 3: Improvement to GPT-4 on different metrics with DELLM on different ratios of training data.

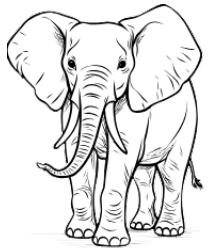
- ❑ DELLM maintains its effectiveness with partial training data.
- ❑ The budget for annotating the knowledge for SQL generation can be reconsidered with the use of DELLM.

Conclusion

- **Significance of Knowledge Generation:** Our study demonstrates the crucial role of knowledge generation in improving the performance of LLMs for text-to-SQL tasks, effectively bridging the gap between user questions and database schemas.

- **Novel Framework:** We propose a novel framework that utilizes database content, execution feedback, and comparisons with ground-truth SQL, addressing existing challenges and enhancing the model's understanding and accuracy.

- **Experimental Results:** Extensive experiments on the BIRD and Spider datasets show substantial improvements in execution accuracy and efficiency scores for models like GPT-4, highlighting the framework's efficacy in advancing text-to-SQL research and driving innovations in natural language processing and data mining.



ACL 2024

Bangkok, Thailand

Thanks for your listening!

暨南大學

JINAN UNIVERSITY

THE HONG KONG POLYTECHNIC UNIVERSITY

香港理工大學