ROUTE: Robust Multitask Tuning and Collaboration for Text-to-SQL

Yang Qin, Chao Chen, Zhihang Fu, Ze Chen, Dezhong Peng, Peng Hu, Jieping Ye
College of Computer Science, Sichuan University

GitHub: https://github.com/alibaba/Route

Singapore EXPO
Thu Apr 24 – Mon Apr 28th, 2025

Background

Database (d)

Tables:
CREATE TABLE movies(⋯);
CREATE TABLE users(⋯);
CREATE TABLE ratings(⋯);

Rows:
'movies': [row11, row12, ⋯]
'users': [row21, row22, ⋯]
'ratings': [row31, row32, ⋯]

SQL, SQLite Database

Question (q)

Question:
What is the average number of Mubi users who love movies directed by Stanley Kubrick?

Question Hint:
average = AVG (movie_popul-arity); number of Mubi users who loves the movie refers to movie_popularity;

LLMs

Prompt

Given the following database schema and question, your task is to write a valid SQL query whose execution results can accurately answer the question.

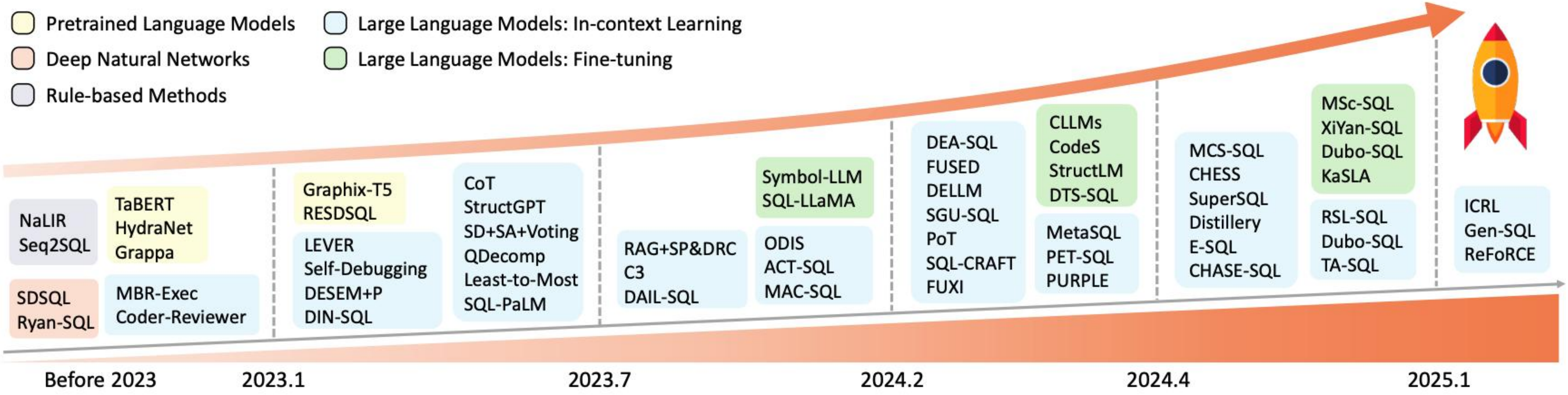
...omitted...

SELECT AVG(movie_popularity) FROM movies
WHERE director_name = 'Stanley Kubrick' ;

Results

Background

- Pre-LLM methods
  - Rule modeling, specialized neural networks, pre-trained models, and etc.
- LLM-based methods
  - Prompt Engineering
  - Fine-tuning-based methods



Ref: Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL

Motivation

- Existing prompt engineering methods are not applicable to small-sized LLMs.
- Most fine-tuning-based methods only focus on SFT of a single SQL-related task

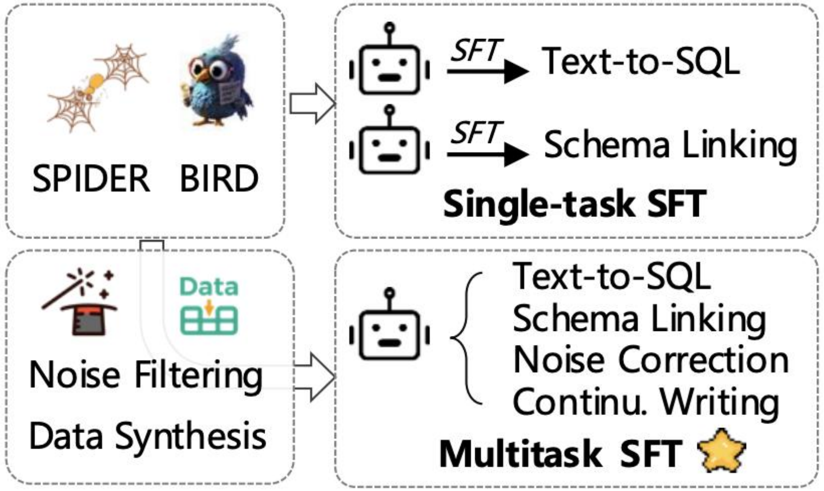
Methods	SPIDER			BIRD	
	Dev-EX	Dev-TS	Test-EX	Dev-EX	Dev-VES
Llama3-8B (Touvron et al., 2023)	69.3	58.4	69.1	32.1	31.6
Qwen2.5-7B (Yang et al., 2024a)	72.5	64.0	75.9	41.1	42.0
Qwen2.5-14B (Yang et al., 2024a)	76.9	66.3	78.4	48.4	49.2
DIN-SQL + Llama3-8B	48.7	39.3	47.4	20.4	24.6
DIN-SQL + Qwen2.5-7B	72.1	61.2	71.1	30.1	32.4
MAC-SQL + Llama3-8B	64.3	52.8	65.2	40.7	40.8
MAC-SQL + Qwen2.5-7B	71.7	61.9	72.9	46.7	49.8
Ours: MCP + Llama3-8B	75.0	63.4	72.0	42.7	44.8
Ours: MCP + Qwen2.5-7B	78.3	67.2	78.7	49.7	52.8
Ours: MCP + Qwen2.5-14B	**80.0**	**67.3**	**80.6**	**56.3**	**57.6**

Lower Transferability

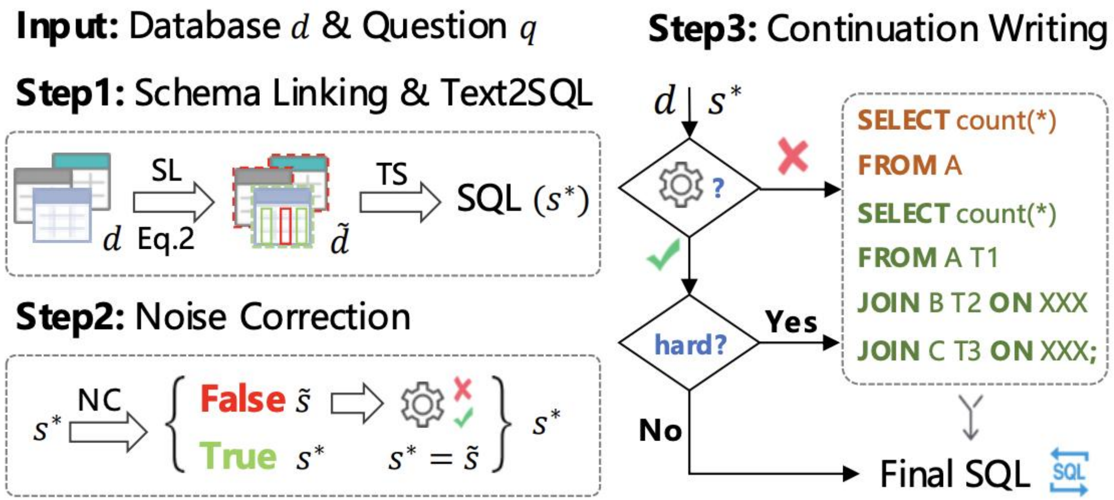
No.	Settings	TS EX	EX	SPIDER-SL Table-R/P	Column-R/P	BIRD-SL Table-R/P	Column-R/P	NC EX	EX	CW EX	EX
#1	**MSFT**	**83.6**	**53.6**	**97.38/95.71**	**98.59/96.98**	**90.87/90.22**	**96.13/90.89**	**83.4**	**53.4**	**91.1**	**73.9**
#2	MSFT w/o TS	0.1	16.2	96.58/93.94	98.40/96.32	90.79/88.26	95.95/90.34	77.4	45.5	86.5	69.6
#3	MSFT w/o SL	81.8	50.9	–	–	–	–	76.3	47.4	91.3	73.5
#4	MSFT w/o NC	82.8	51.0	96.52/94.25	99.00/96.59	90.41/88.85	96.09/90.75	–	–	91.7	73.4
#5	MSFT w/o CW	81.2	50.3	96.51/93.97	98.65/96.39	90.59/88.12	96.05/90.64	79.4	49.0	81.2	56.7
#6	SFT with TS	83.1	52.9	–	–	–	–	–	–	85.6	69.2
#7	SFT with SL	–	–	95.55/92.69	98.91/95.29	87.84/85.11	94.93/89.51	–	–	–	–
#8	SFT with NC	0.1	8.7	–	–	–	–	78.9	49.3	48.6	38.6
#9	SFT with CW	68.1	39.0	–	–	–	–	–	–	89.8	70.1
#10	Llama3 w/o SFT	69.3	32.1	88.35/76.37	94.83/91.46	83.77/75.38	89.55/86.39	72.1	38.1	80.3	57.6

Single task overfitting risk

Method



(a) Multitask Supervised Fine-Tuning



(b) Multitask Collaboration Prompting

- ➢ MSFT: Filtering potential noisy pairs in the training set and synthesizing multiple SQL-related task SFT data.
- ➢ MCP: The proposed four tasks are combined to reduce hallucinations/errors in the SQL generation process through multitask collaboration.

Method-MSFT

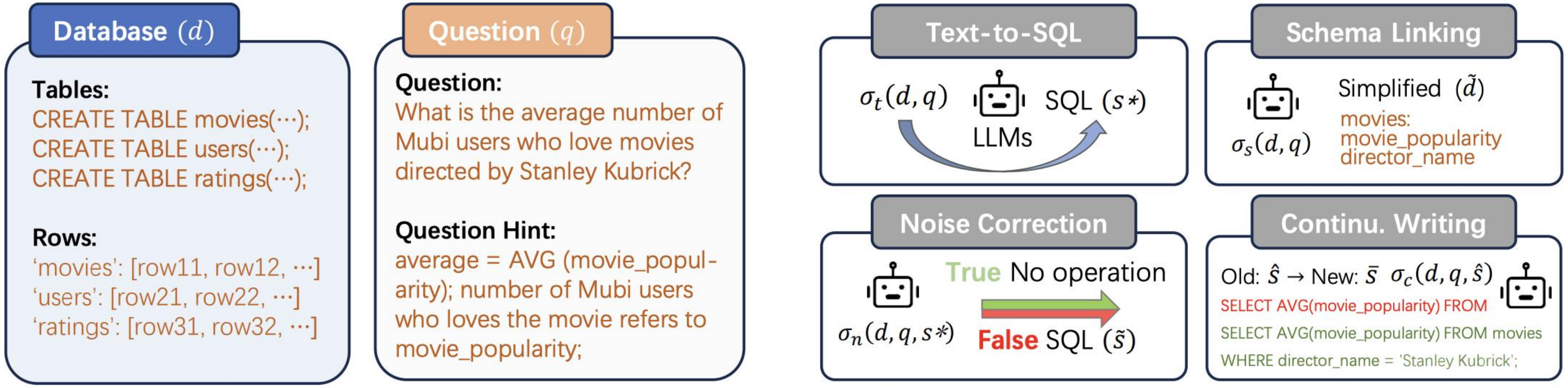
- ➢ Noisy correspondence filtering
- ➢ MSFT data synthesizing

$$\mathcal{D}_M = \mathcal{D}_t \cup \mathcal{D}_s \cup \mathcal{D}_n \cup \mathcal{D}_c$$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathcal{D}_M}\left[\sum_{t=1}^{T}\log p_{\mathcal{M}}(y_t|\boldsymbol{y}_{1:t-1},\boldsymbol{x})\right]$$

Example 1:
Q1: What is the number of inhabitants and income of geographic identifier 239?
A1: SELECT INHABITANTS_K FROM Demog WHERE GEOID = 239;
R1: SELECT INHABITANTS_K, INCOME_K FROM Demog WHERE GEOID = 239;

Example 2:
Q2: List the geographic id of places where the income is above average.
A2: SELECT AVG(INCOME_K) FROM Demog;
R2: SELECT GEOID FROM Demog WHERE INCOME_K > (SELECT AVG(INCOME_K) FROM Demog);



Method-MCP

Algorithm 1 The algorithm of MCP

Input: The database d, user question q, and LLM $\mathcal{M}$;

// Conduct schema linking.

Obtain simplified database $\tilde{d}$ via Equation (2);

// SQL generation.

Generate intermediate SQL query s^* via $\mathcal{M}(\sigma_t(\tilde{d}, q))$;

// Conduct noise correction.

Check the SQL query s^* via $\mathcal{M}(\sigma_n(d, q, s^*, e))$.

Obtain the the correct SQL $\tilde{s}$ if $\mathcal{M}$ shows that s^* is inaccurate.

if $\mathcal{M}$ shows s^* is inaccurate and $\mathrm{SQLer}(d, \tilde{s})$ is `True` **then**

    $s^* = \tilde{s}$.

end if

// Refine wrong or hard SQL queries by continuation writing.

if $\mathrm{SQLer}(d, s^*)$ is `False` or $h(s^*, d) > 2$ **then**

    Construct the truncated SQL query $\hat{s}$ based on s^*;

    Continue writing: $\bar{s} = \mathcal{M}(\sigma_c(d, q, \hat{s}))$;

    if $\mathrm{SQLer}(d, \bar{s})$ is `True` **then**

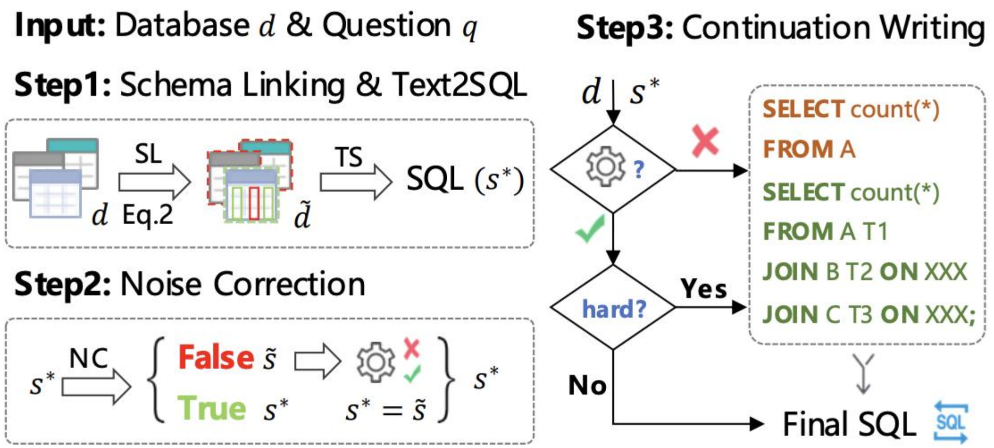
        $s^* = \bar{s}$;

    end if

end if

Output: The final SQL query s^*.

$$\tilde{d}_i = \underbrace{\mathcal{M}(\sigma_s(d_i, q_i))}_{\text{LLM for SL}} \uplus \underbrace{f(\mathcal{M}(\sigma_t(d_i, q_i)), d_i)}_{\text{Pseudo SQL}}$$



Experiments

Table 1: Performance comparison on SPIDER and BIRD benchmarks. The results of re-evaluation using the open-source code repository are marked with '†'. In groups of open-source LLMs, the best results are highlighted in **bold** and the second-best results are in underlined.

Methods	SPIDER			BIRD	
	Dev-EX	Dev-TS	Test-EX	Dev-EX	Dev-VES
Prompting with GPT-4					
GPT-4 (Achiam et al., 2023)	72.9	64.9	-	46.4	49.8
DIN-SQL + GPT-4 (Pourreza & Rafiei, 2024a)	82.8	74.2	85.3	50.7	58.8
DAIL-SQL + GPT-4 (Gao et al., 2024a)	83.5	76.2	86.6	54.8	56.1
MAC-SQL + GPT-4 (Wang et al., 2023)	86.8	-	82.8	59.4	66.2
MCS-SQL + GPT-4 (Lee et al., 2024)	89.5	-	89.6	63.4	64.8
Prompting with Open-Source LLMs					
Mistral-7b (Jiang et al., 2023)	56.8	47.3	60.1	22.5	27.8
Llama3-8B (Touvron et al., 2023)	69.3	58.4	69.1	32.1	31.6
Qwen2.5-7B (Yang et al., 2024a)	72.5	64.0	75.9	41.1	42.0
Qwen2.5-14B (Yang et al., 2024a)	76.9	66.3	78.4	48.4	49.2
DIN-SQL + Llama3-8B	48.7	39.3	47.4	20.4	24.6
DIN-SQL + Qwen2.5-7B	72.1	61.2	71.1	30.1	32.4
MAC-SQL + Llama3-8B	64.3	52.8	65.2	40.7	40.8
MAC-SQL + Qwen2.5-7B	71.7	61.9	72.9	46.7	49.8
Ours: MCP + Llama3-8B	75.0	63.4	72.0	42.7	44.8
Ours: MCP + Qwen2.5-7B	<u>78.3</u>	<u>67.2</u>	<u>78.7</u>	<u>49.7</u>	<u>52.8</u>
Ours: MCP + Qwen2.5-14B	**80.0**	**67.3**	**80.6**	**56.3**	**57.6**
Fine-Tuning with Open-Source LLMs					
Llama3-8B + SFT (Touvron et al., 2023)	82.4	76.2	83.1	53.1	59.0
Qwen2.5-7B + SFT (Yang et al., 2024a)	80.9	75.6	82.8	51.4	53.1
DTS-SQL-7B (Pourreza & Rafiei, 2024b)	82.7†	78.4†	82.8†	55.8	<u>60.3</u>
CODES-7B + SFT (Li et al., 2024b)	85.4	80.3	-	57.2	58.8
CODES-15B + SFT (Li et al., 2024b)	84.9	79.4	-	<u>58.5</u>	56.7
SENSE-7B (Yang et al., 2024b)	83.2	<u>81.7</u>	83.5	51.8	-
SENSE-13B (Yang et al., 2024b)	84.1	**83.5**	<u>86.6</u>	55.5	-
Ours: ROUTE + Llama3-8B	<u>86.0</u>	80.3	83.9	57.3	60.1
Ours: ROUTE + Qwen2.5-7B	83.6	77.5	83.7	55.9	57.4
Ours: ROUTE + Qwen2.5-14B	**87.3**	80.9	**87.1**	**60.9**	**65.2**

Conclusion

- Although the GPT-4-based methods are effective, their effectiveness is reduced when transferred to small-sized LLMs.
- Compared with fine-tuning based methods, our ROUTE leads in most indicators.
- From Table 2, our solution shows strong potential in cross-domain performance.

Table 2: Performance on SPIDER-variant benchmarks. The best results are highlighted in **bold**.

Methods	SYN EX	SYN TS	Realistic EX	Realistic TS	DK EX	Avg.
Llama3-8B	60.3	47.1	68.5	50.8	58.3	57.0
+ SFT	75.3	68.7	76.8	69.7	72.0	72.5
+ SFT + MCP	76.1	69.4	78.0	70.7	73.5	73.5
+ MSFT	72.1	65.1	77.0	68.1	72.3	70.9
+ **ROUTE**	**77.4**	**70.2**	**80.9**	**72.6**	**74.6**	**75.1**

Experiments

Table 3: The ablation results (EX) on SPIDER and BIRD development sets.

No.	SFT	MSFT	MCP	NF	SPIDER	BIRD
#1	✓	✓	✓	✓	86.0	57.3
#2	✓	✓		✓	83.6	53.6
#3	✓	✓	✓		84.5	57.4
#4	✓	✓			83.3	53.1
#5	✓				82.4	53.1
#6	✓		✓	✓	83.5	56.1
#7	✓			✓	83.1	52.9
#8	✓		✓		83.8	56.0
#9			✓		75.0	42.7
#10					69.3	32.1

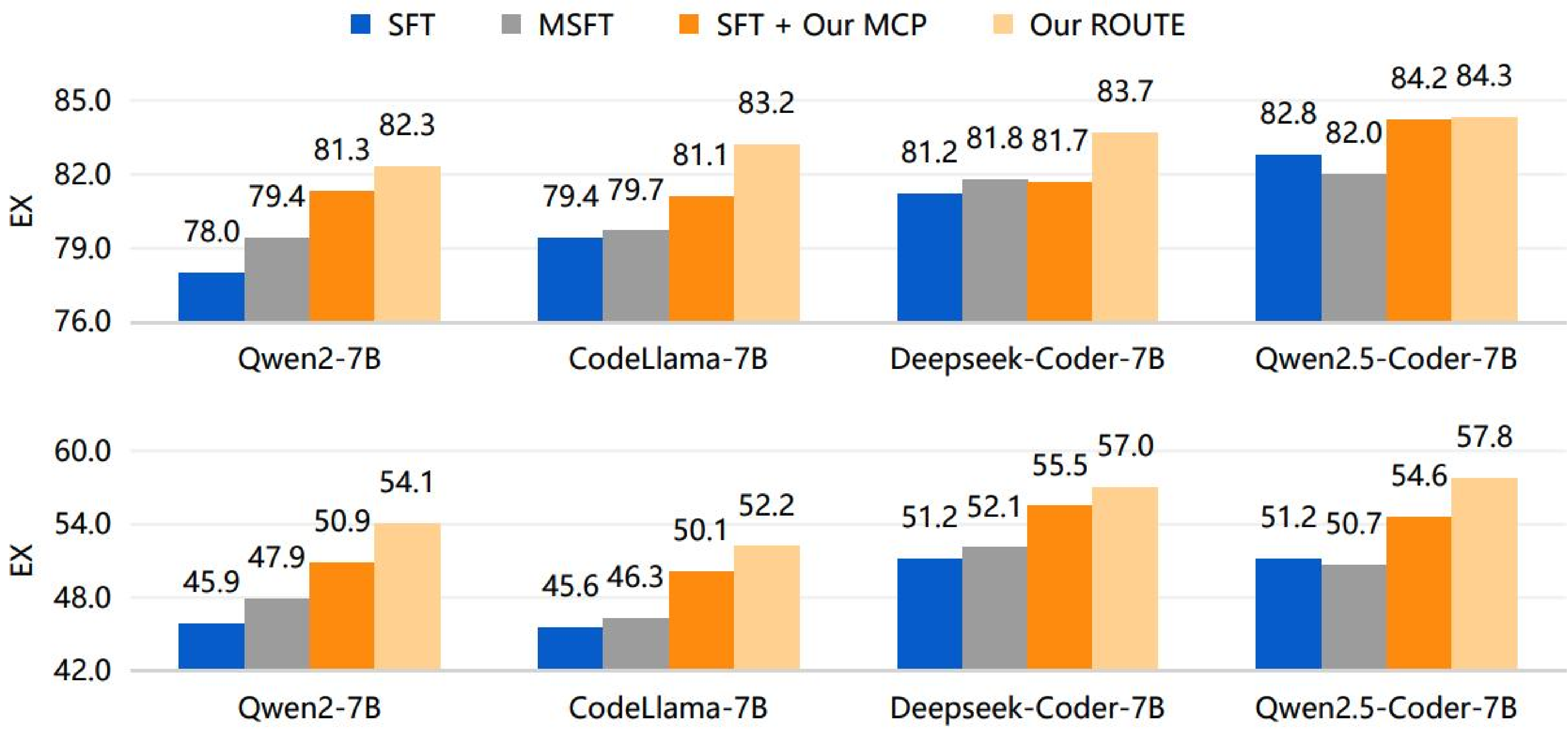
Table 4: The ablation results (EX) on multi-task collaboration prompting.

No.	SL	NC	CW	SPIDER	BIRD
#1	✓	✓	✓	86.0	57.3
#2	✓			85.8	56.0
#3		✓		83.9	54.7
#4			✓	83.8	54.0
#5				83.6	53.6
#6	✓	✓	✓	75.0	42.7
#7	✓			73.3	36.8
#8		✓		72.1	38.1
#9			✓	71.3	36.4
#10				69.3	32.1

Conclusion

- The experimental results show that each proposed technology has a significant effect on the accuracy of SQL generation.
- Our MCP shows generalizability for both fine-tuned and non-fine-tuned models.
- Each task is indispensable. Through multi-tasking collaboration, performance is optimized.

Experiments



Experiments

Table 6: The performance (EX) of various-sized open-source LLMs.

Models	≈7B SPIDER	≈7B BIRD	≈70B SPIDER	≈70B BIRD
Llama3	69.3	32.1	77.9	46.9
Llama3 + MCP	75.0	42.7	79.0	51.8
Qwen2.5	72.5	41.1	81.7	53.3
Qwen2.5 + MCP	78.3	49.7	82.3	57.1

Methods	SPIDER			BIRD	
	Dev-EX	Dev-TS	Test-EX	Dev-EX	Dev-VES
Llama3-8B (Touvron et al., 2023)	69.3	58.4	69.1	32.1	31.6
Qwen2.5-7B (Yang et al., 2024a)	72.5	64.0	75.9	41.1	42.0
Qwen2.5-14B (Yang et al., 2024a)	76.9	66.3	78.4	48.4	49.2
DIN-SQL + Llama3-8B	48.7	39.3	47.4	20.4	24.6
DIN-SQL + Qwen2.5-7B	72.1	61.2	71.1	30.1	32.4
MAC-SQL + Llama3-8B	64.3	52.8	65.2	40.7	40.8
MAC-SQL + Qwen2.5-7B	71.7	61.9	72.9	46.7	49.8
Ours: MCP + Llama3-8B	75.0	63.4	72.0	42.7	44.8
Ours: MCP + Qwen2.5-7B	78.3	67.2	78.7	49.7	52.8
Ours: MCP + Qwen2.5-14B	**80.0**	**67.3**	**80.6**	**56.3**	**57.6**

Conclusion

➢ The existing Prompt method shows low versatility and is only applicable to large-sized LLMs with strong instruction comprehension capabilities.

➢ Our MCP is not only effective in small-size LLMs, but also in large-size LLMs.

Conclusion

- In this paper, we study and propose a robust multitask tuning and collaboration method named ROUTE to stimulate the potential of open-source LLM in Text2SQL, narrowing the gap with existing solutions based on closed-source LLMs, such as GPT-4.
- Our method minimizes the risk of hallucination in SQL generation by explicitly learning multiple SQL-related tasks and conducting multitask collaboration.
- We apply our approach to recent LLMs to demonstrate its effectiveness and superiority on multiple benchmarks. The results show that our method has satisfactory transferability and achieves promising execution accuracy on Text2SQL.

Thanks for your attention!

College of Computer Science
Sichuan University