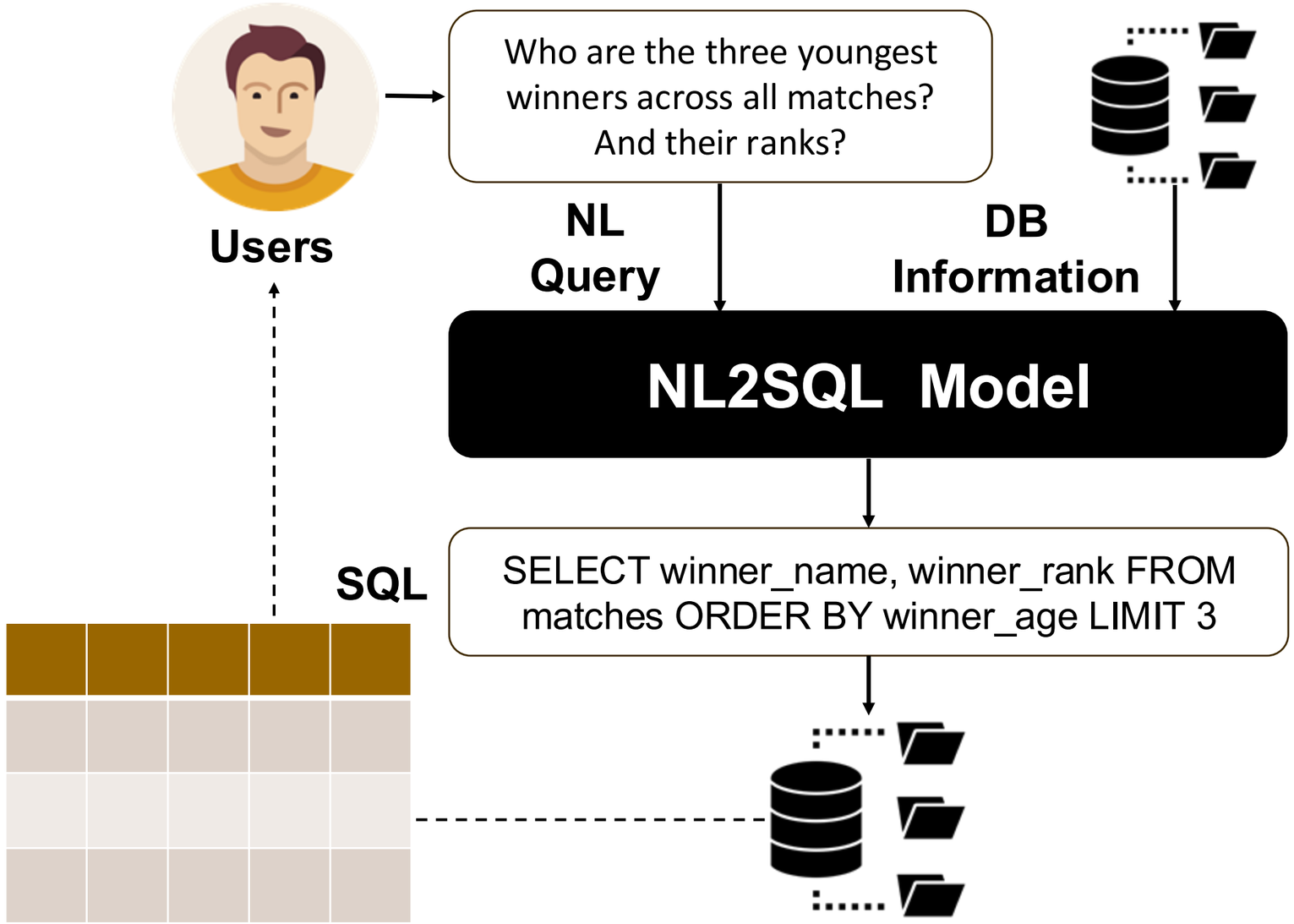
VLDB 2024 - 50th International Conference on Very Large Data Bases

The Dawn of Natural Language to SQL: Are We Fully Ready?

August 29, 2024

Boyan Li[1], *Yuyu Luo[1,2], Chengliang Chai[3], Guoliang Li[4], Nan Tang[1,2]
[1]HKUST(GZ), [2]HKUST, [3]Beijing Institute of Technology, [4]Tsinghua University

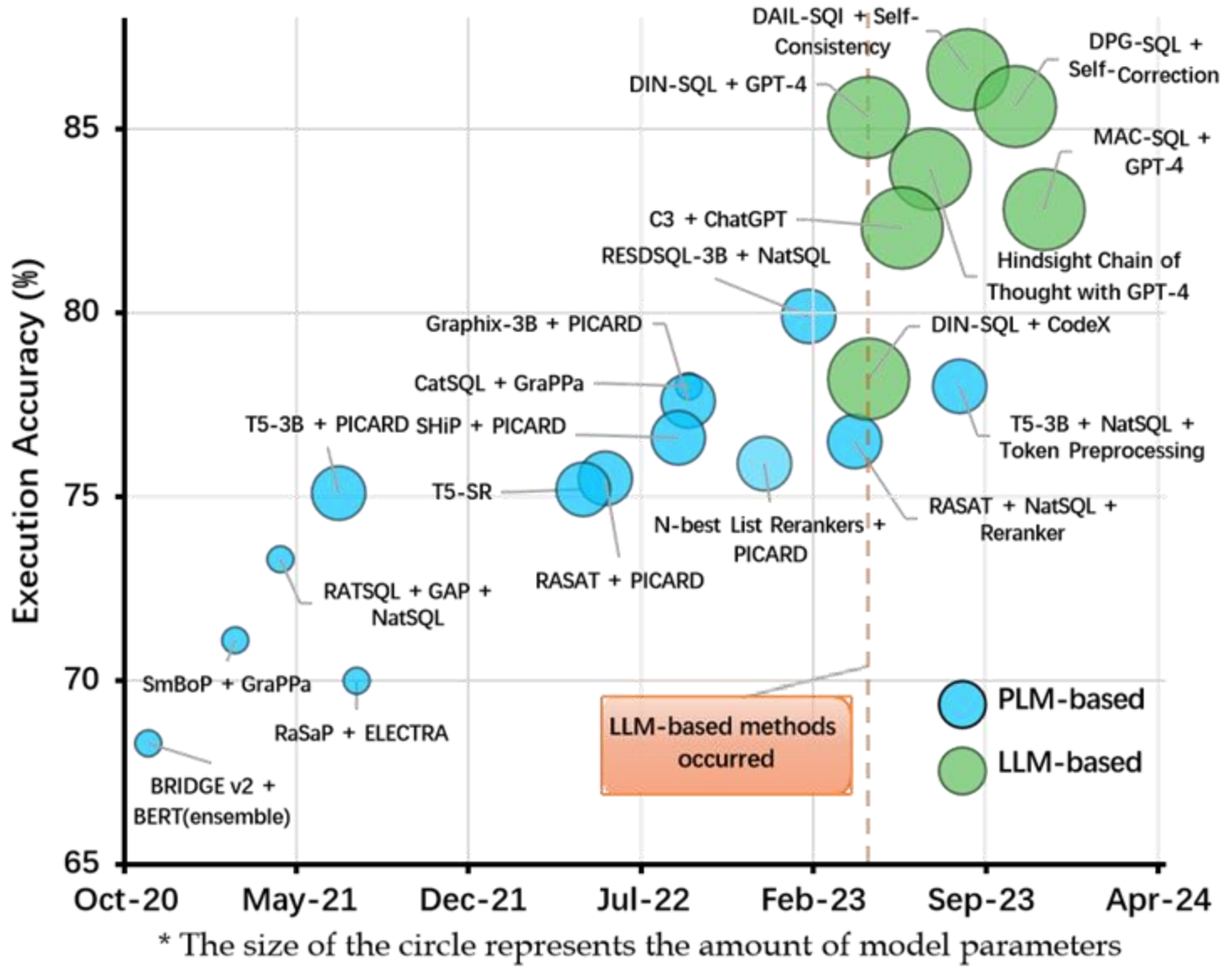
WHAT IS NATURAL LANGUAGE TO SQL (NL2SQL)?



WHERE ARE WE NOW IN NL2SQL?

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

■ Recent advanced NL2SQL methods are mainly driven by LLM.



■ *Can we conclude that **LLM-based** models are **"the best choice"** for any NL2SQL application?*

ONE MODEL DOES NOT FIT ALL!

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

SPECIFIC DOMAIN

- Domain-specific database
  - Competition
  - Medical
  - Finance

MULTIPLE JOINS

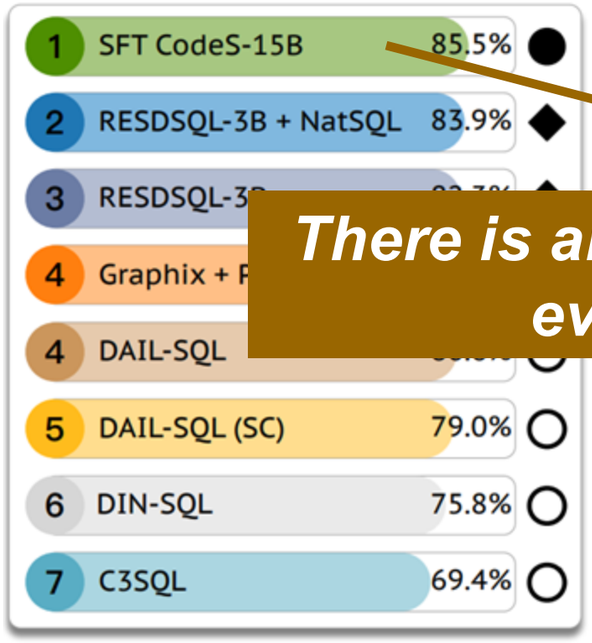
- Query with multiple JOINs

NESTED SQL QUERY

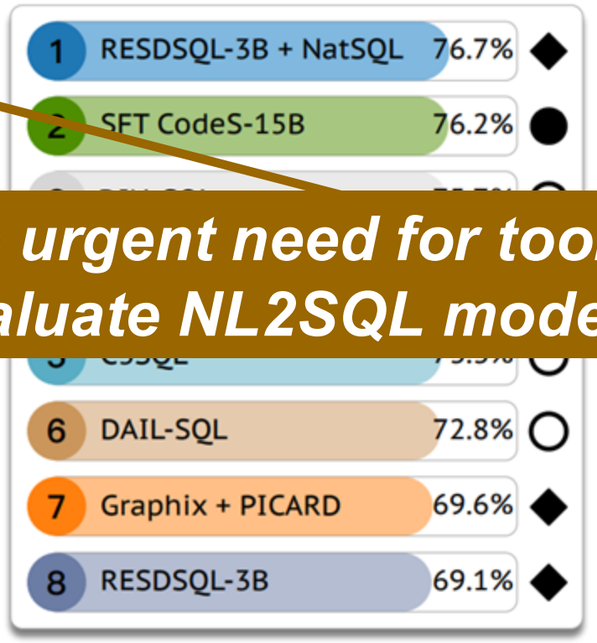
- Query with nested SQL

QUESTION VARIANCE

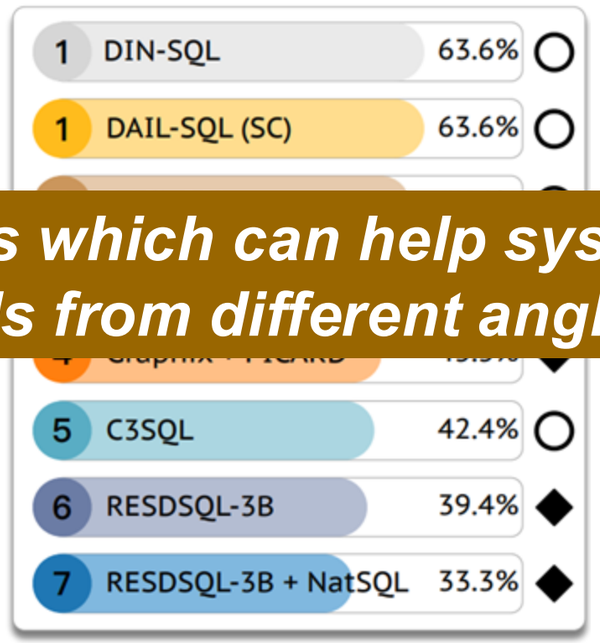
- Different user questions for the same SQL query



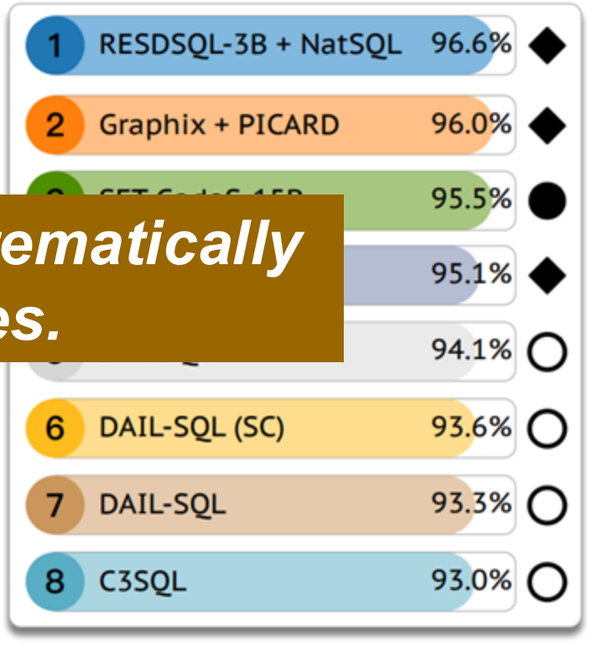
(a) Test Results in Domain of *Competition*



(b) Test Results in SQL with JOIN



(c) Test Results in Nested SQL

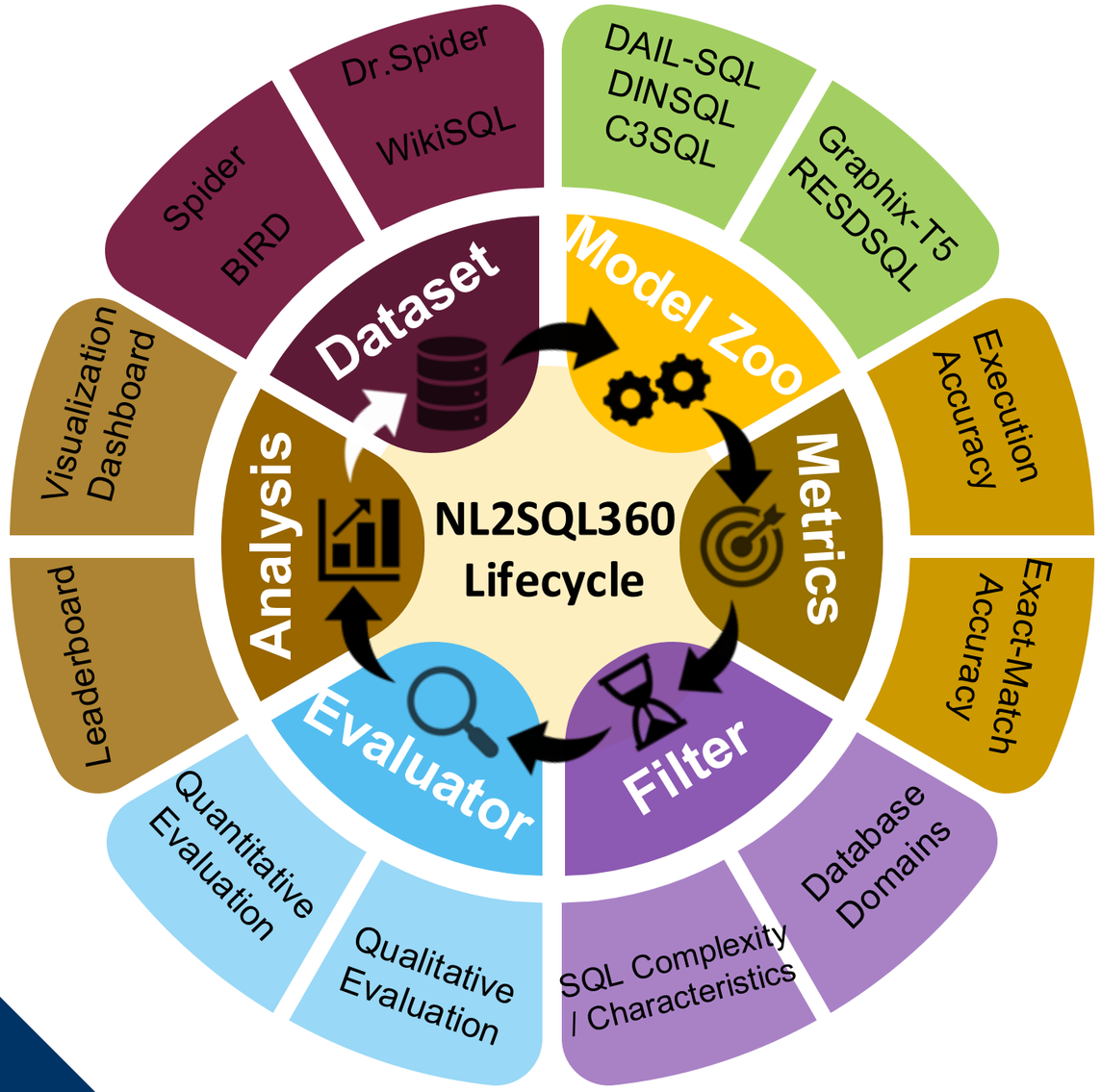


(d) Test Results in Query Variance

There is an urgent need for tools which can help systematically evaluate NL2SQL models from different angles.

NL2SQL360: A MULTI-ANGLE EVALUATION FRAMEWORK

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)



- Dataset Filter
  - Case 1: SQL Complexity
  - Case 2: SQL Characteristics
  - Case 3: NL Query Variance
  - Case 4: Database Domain
  - Case 5: Economy
  - Case 6: Efficiency

NL2SQL360 Github Repository: https://github.com/HKUSTDial/NL2SQL360

NL2SQL360: A MULTI-ANGLE EVALUATION FRAMEWORK

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

NL2SQL360 Github Repository: https://github.com/HKUSTDial/NL2SQL360

NL2SQL360: A MULTI-ANGLE EVALUATION FRAMEWORK

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

NL2SQL360 Github Repository: https://github.com/HKUSTDial/NL2SQL360

EXPERIMENTS WITH NL2SQL360

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- We utilized **NL2SQL360** to evaluate **7** PLM-based and **13** LLM-based NL2SQL methods on **2** widely-used benchmarks, varing **15** settings and deriving a set of new findings.

DATASETS

- Spider Dataset
- BIRD Dataset

BASELINES

- PLM-based:
  - RESDSQL-3B, RESDSQL-3B + NatSQL, Graphix + PICARD
- LLM-based:
  - DAILSQL, DAILSQL(SC), DINSQL, C3SQL, SFT CodeS-1B/3B/7B/15B

EVALUATION METRICS

- Execution Accuracy (EX)
- Exact Match Accuracy (EM)
- Valid Efficiency Score (VES)
- Query Variance Testing (QVT)

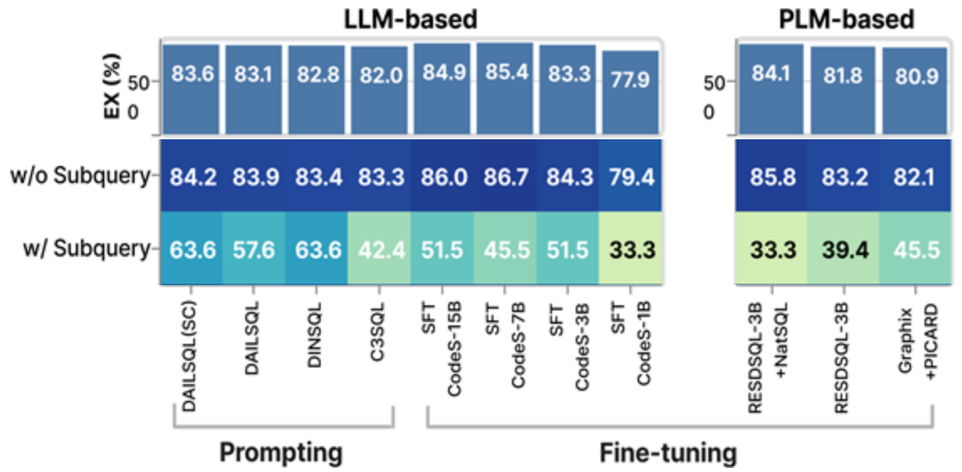
EVALUATION ANGLES

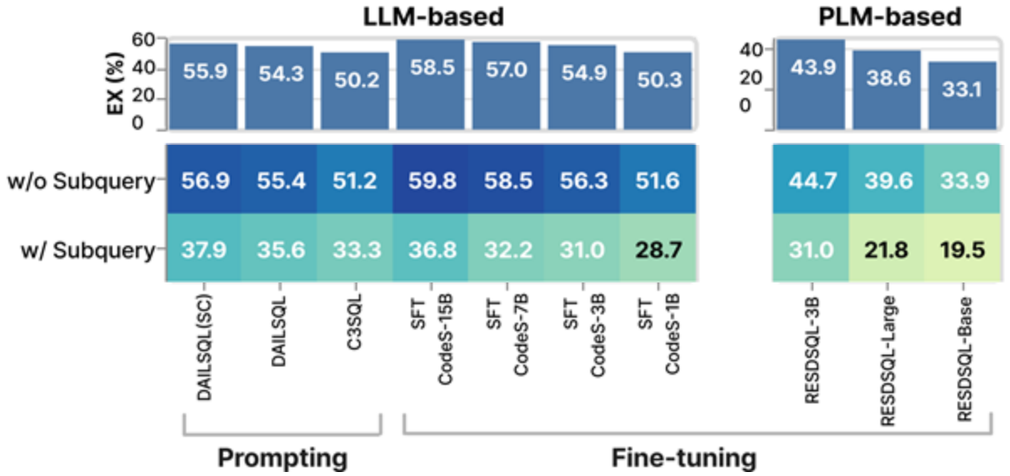
- SQL Characteristics
- Question Variance
- Database Domain Adaption
- Pre-training Corpus
- Efficiency
- Economy (Cost)

ACCURACY VS. SQL CHARACTERISTICS

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Motivation: Real-world applications often require generating SQL quries involving advanced operations like subquries.





ACCURACY VS. SUBQUERY SUBSET

- In overall dataset
  - SFT CodeS-15B > DAILSQL(SC) > DAILSQL > DINSQL
- However, in subquery scenario
  - DAILSQL(SC) = DINSQL > DAILSQL > SFT CodeS-15B

In scenarios involving subqueries, LLM-based methods outperform PLM-based methods overall, with methods using GPT-4 showing particularly better performance.

ACCURACY VS. QUESTION VARIANCE

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Motivation: For the same SQL query, different users may pose different NL questions. The ability of handling question variance is crucial for NL2SQL applications.

■ To this end, we proposed new metric called Question Variance Testing (QVT):

$$QVT = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\sum_{j=1}^{m_i}\mathbb{I}(\mathcal{F}(N_{ij}) = Q_i)}{m_i}\right)$$

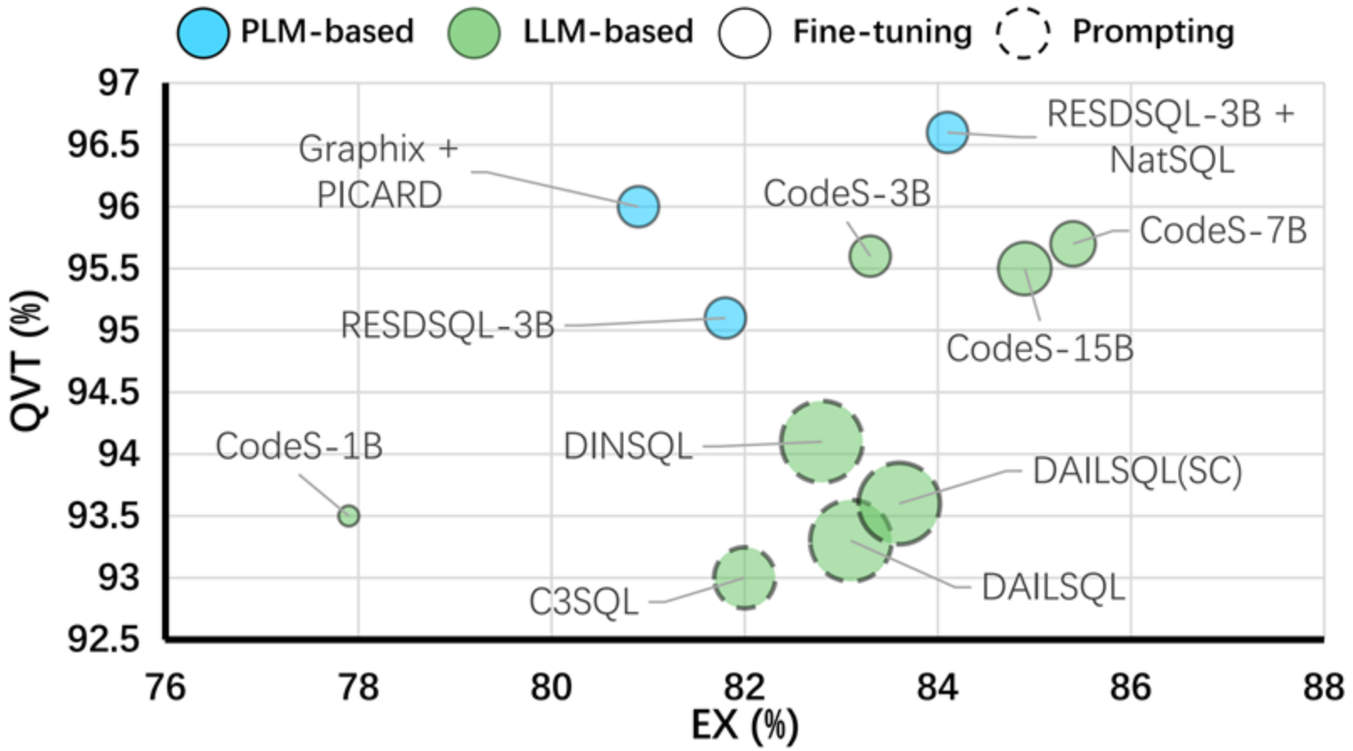
– M is the total number of SQL quries in the test set.
– m_i is the number of natural language query variations corresponding to the SQL query Q_i.
– $\mathcal{F}(N_{ij})$ represents the SQL query generated by the NL2SQL model for the j-th natural language query variation of Q_i.
– $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the query results inside are equal, and 0 otherwise.

■ For NL2SQL models, higher QVT means higher stability of handing question variance.

ACCURACY VS. QUESTION VARIANCE

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- We evaluate QVT for multple NL2SQL methods on Spider development dataset.



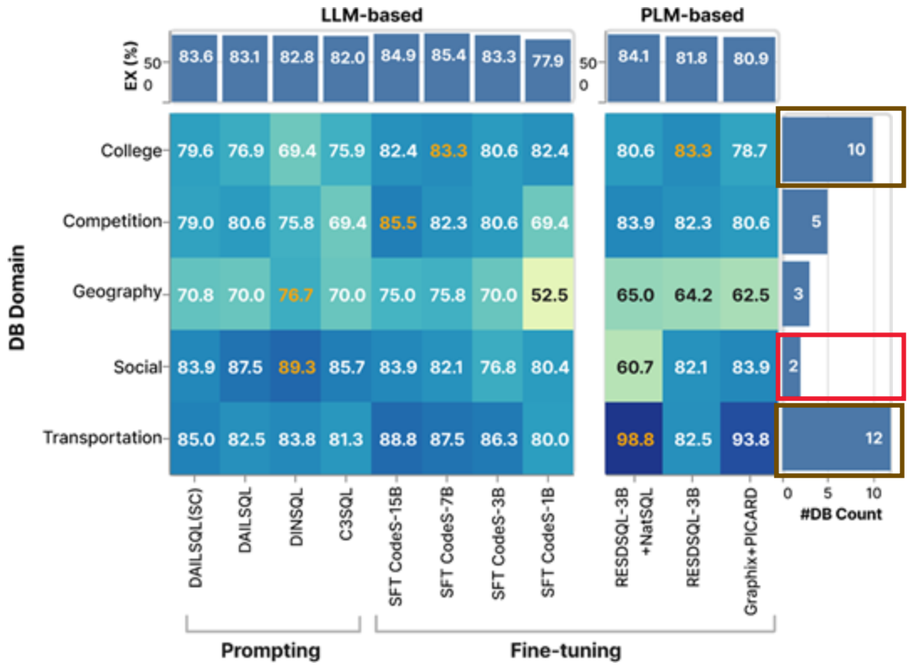
Our finding:

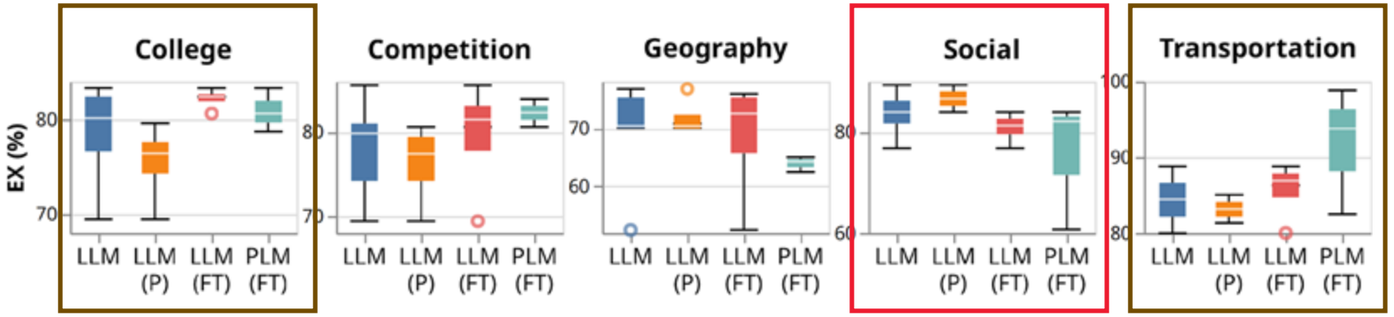
- Fine-tuning the model with task-specific datasets may help stabilize its performance against NL variations.

ACCURACY VS. DATABASE DOMAIN

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Motivation: In practical NL2SQL applications, scenarios usually involve domain-specific databases, each with unique schema designs and terminologies.





- LLM (P): Prompt-based LLMs
- LLM (FT): Fine-tuned LLMs
- PLM (FT): Fine-tuned PLMs

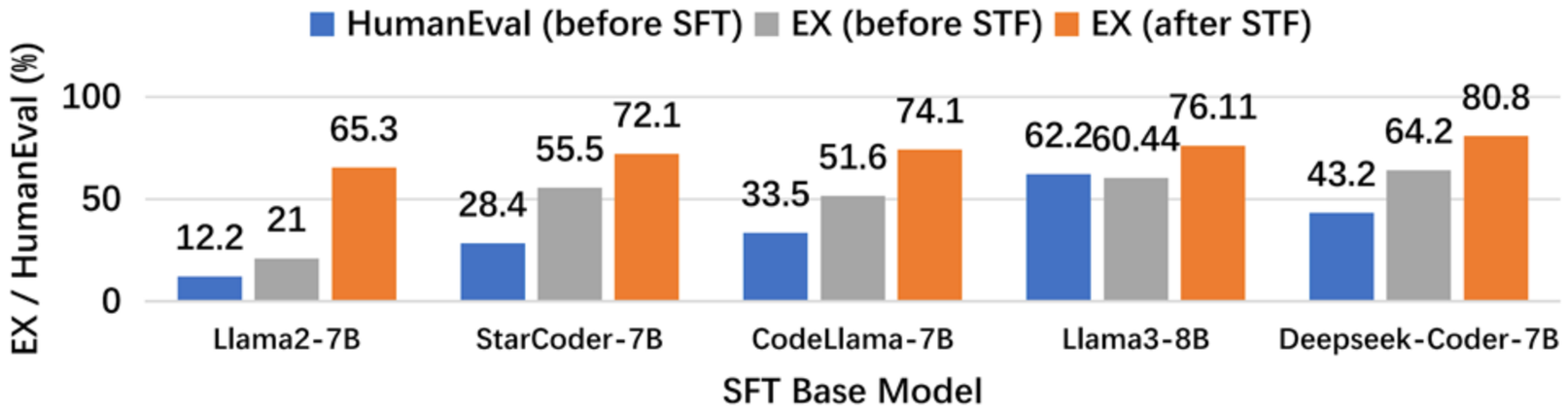
Our findings:

- Different methods exhibit varying biases towards different domains.
- However, in-domain training data during fine-tuning is crucial for model performance in specific domains.

ACCURACY VS. PRE-TRAINING CORPUS

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Motivation: A large number of LLMs with different pre-training corpus have emerged recently, such as Llama, StarCoder, etc. *Which open-source LLMs are best suited for SFT in the NL2SQL task?*
- We evaluate different LLMs which have different pre-training corpus or procedures, and evaluate them in Spider Dataset.



Our finding:

- After SFT on open-source LLMs for the NL2SQL task, we found a positive correlation between the model's inherent coding ability and its performance in NL2SQL task.

RECAP: ONE MODEL DOES NOT FIT ALL!

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

SPECIFIC DOMAIN

- Domain-specific database
  - Competition
  - Medical
  - Finance

MULTIPLE JOINS

- Query with multiple JOINs

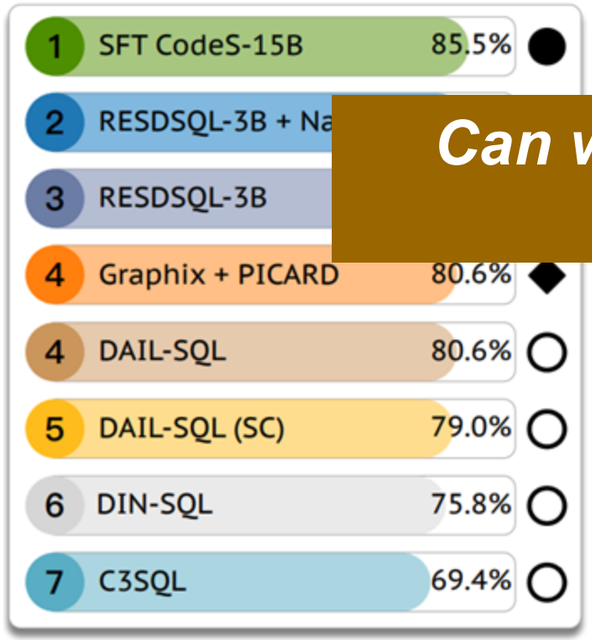
NESTED SQL QUERY

- Query with nested SQL

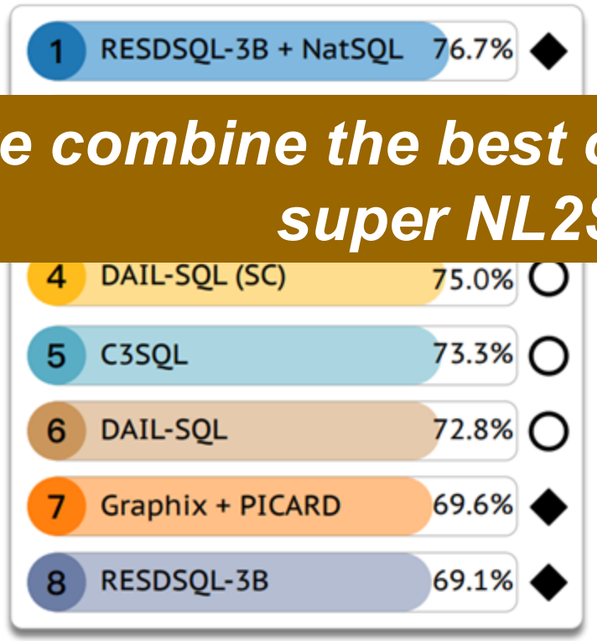
QUESTION VARIANCE

- Different user questions for the same SQL query

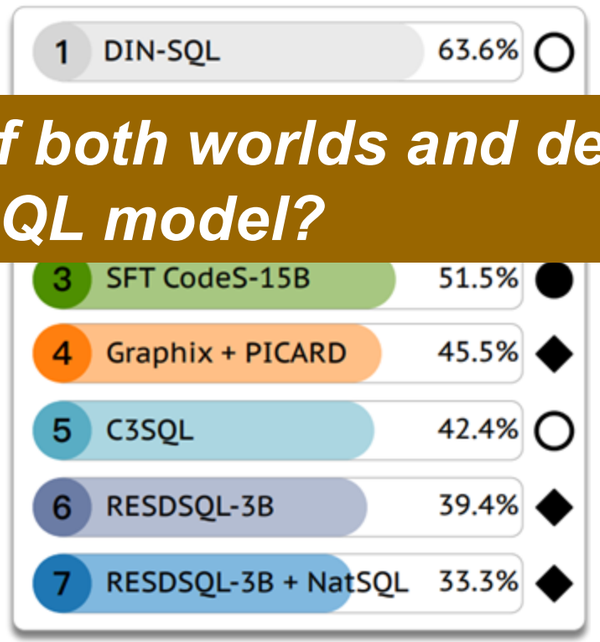
Can we combine the best of both worlds and design a super NL2SQL model?



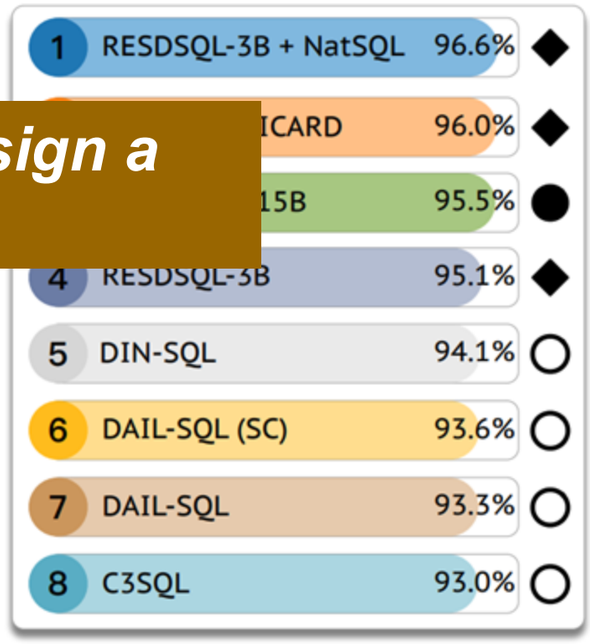
(a) Test Results in Domain of *Competition*



(b) Test Results in SQL with JOIN



(c) Test Results in Nested SQL



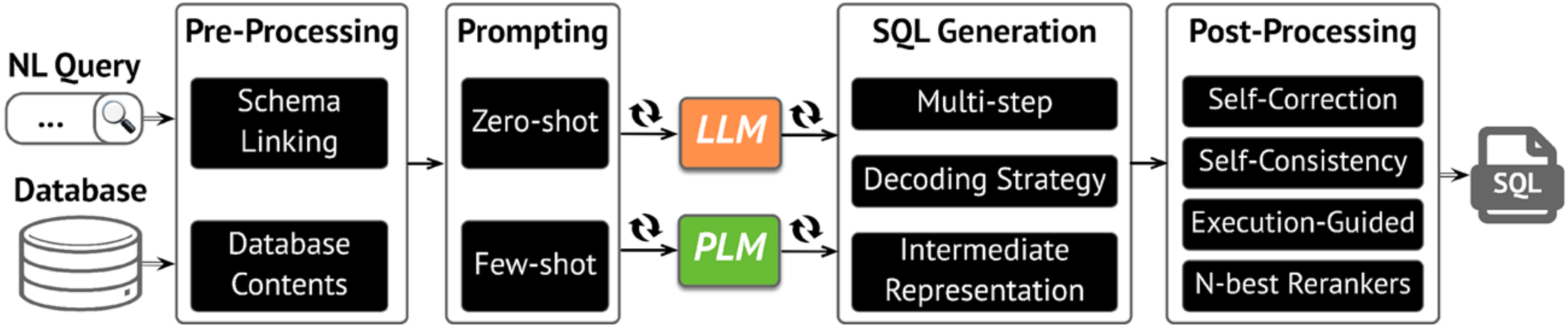
(d) Test Results in Query Variance

WHAT IS NEXT FOR NL2SQL?

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Now, we have found that different NL2SQL models exhibit distinct advantages in specific scenarios. Next, *how can we systematically design future NL2SQL models?*

- We systematically categorized and analyzed recent NL2SQL modules based on LLMs and PLMs, highlighting their commonalities and distinct features.
- We proposed **NL2SQL Design Space** in a modular style.



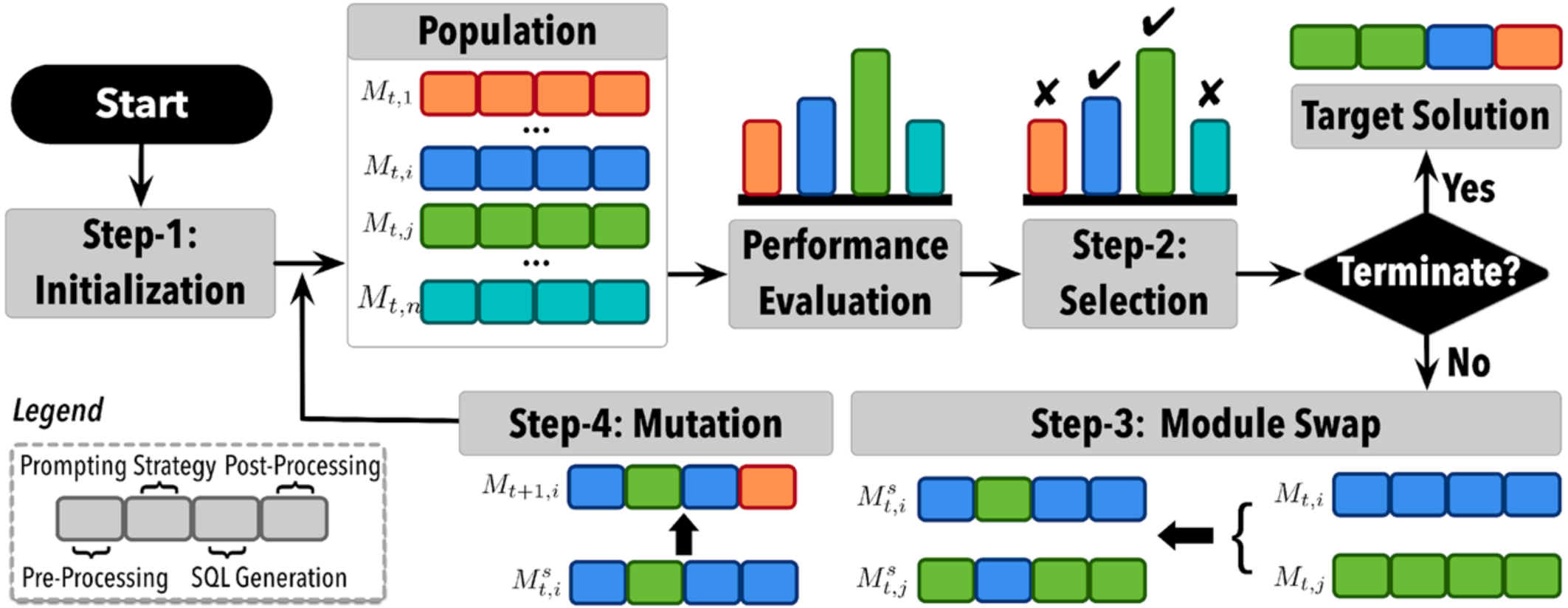
NL2SQL DESIGN SPACE

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

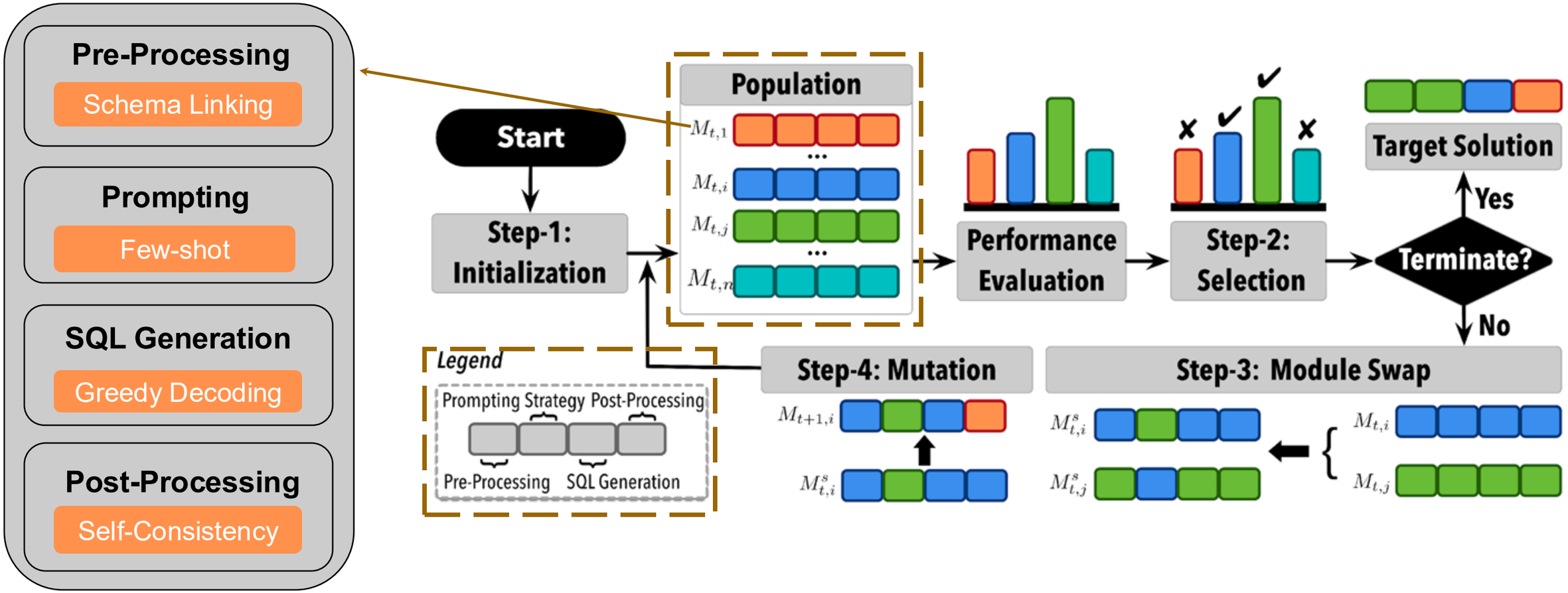
Types		Methods	Backbone Models	Example Selection (Few-shot)	Schema Linking	DB Content	SQL Generation Strategy			Post-processing Strategy
							Multi-Step	Intermediate Representation	Decoding Strategy	
LLM-based	Prompting	DIN-SQL [38]	GPT-4	Manual	✓	✗	Classification Decomposition	NatSQL	Greedy Search	Self-Correction
		DAIL-SQL [11] (with Self-Consistency)	GPT-4	Similarity-based	✗	✗	✗	✗	Greedy Search	Self-Consistency
		MAC-SQL [52]	GPT-4	N/A	✓	✗	Sub-question	✗	Greedy Search	Refiner
										Self-Consistency
										Execution-Guided SQL Selector
	Fine-tuning									Execution-Guided SQL Selector
PLM-based		RESDSQL + NatSQL [25]	T5	N/A	✓	✓	Skeleton Parsing	NatSQL	Beam Search	Execution-Guided SQL Selector
		Graphix + PICARD [25]	T5	N/A	✓	✓	✗	✗	PICARD	✗
		N-best Rerankers + PICARD [59]	T5	N/A	✓	✓	✗	✗	PICARD	N-best Rerankers
		T5 + NatSQL + Token Preprocessing [42]	T5	N/A	✓	✓	✗	NatSQL	Greedy Search	✗
		RASAT + PICARD [39]	T5	N/A	✓	✓	✗	✗	PICARD	✗
		SHiP + PICARD [16]	T5	N/A	✗	✓	✗	✗	PICARD	✗
		T5 + PICARD [45]	T5	N/A	✗	✓	✗	✗	PICARD	✗
		RATSQL + GAP + NatSQL [10]	BART	N/A	✓	✓	✗	NatSQL	✗	✗
		BRIDGE v2 [28]	BERT	N/A	✗	✓	✗	✗	Schema-Consistency Guided Decoding	✗

Can we design an automatic algorithm to explore the NL2SQL design space and achieve stronger performance?

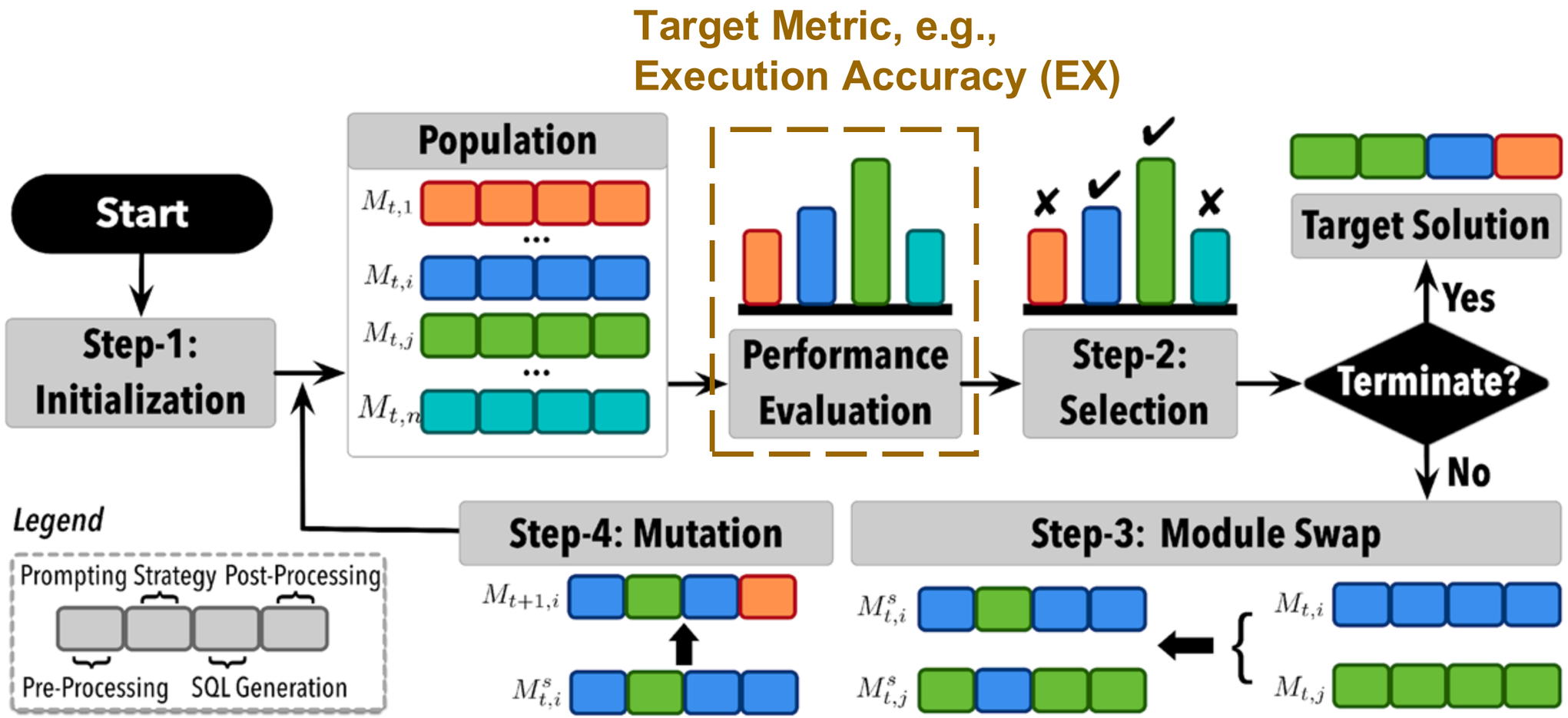
NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



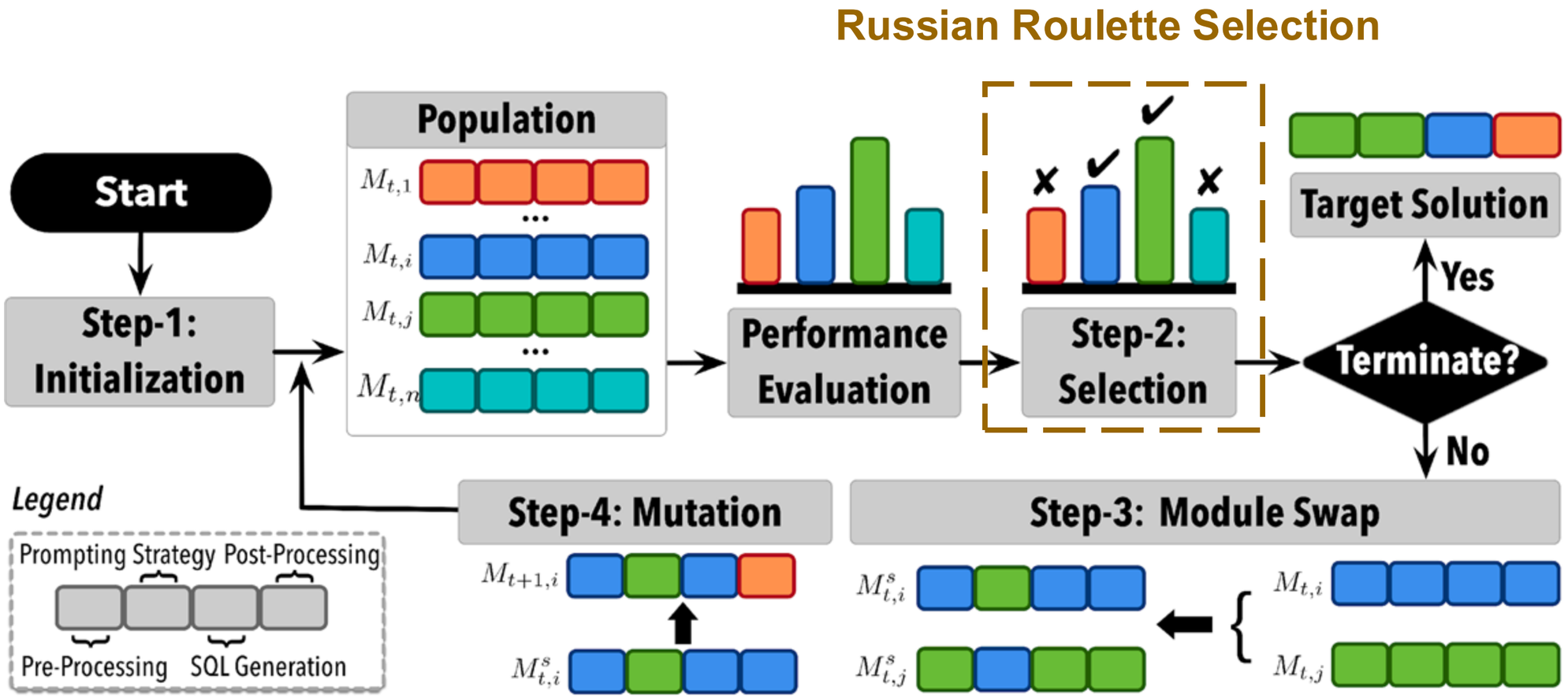
NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



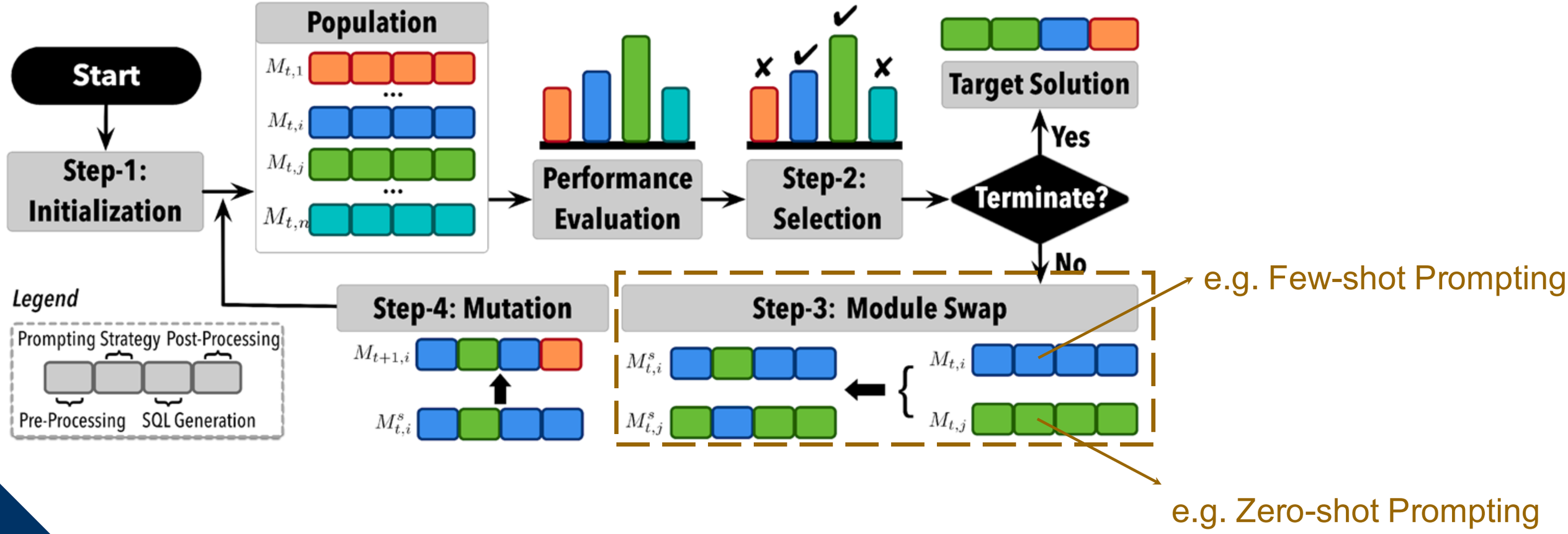
NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



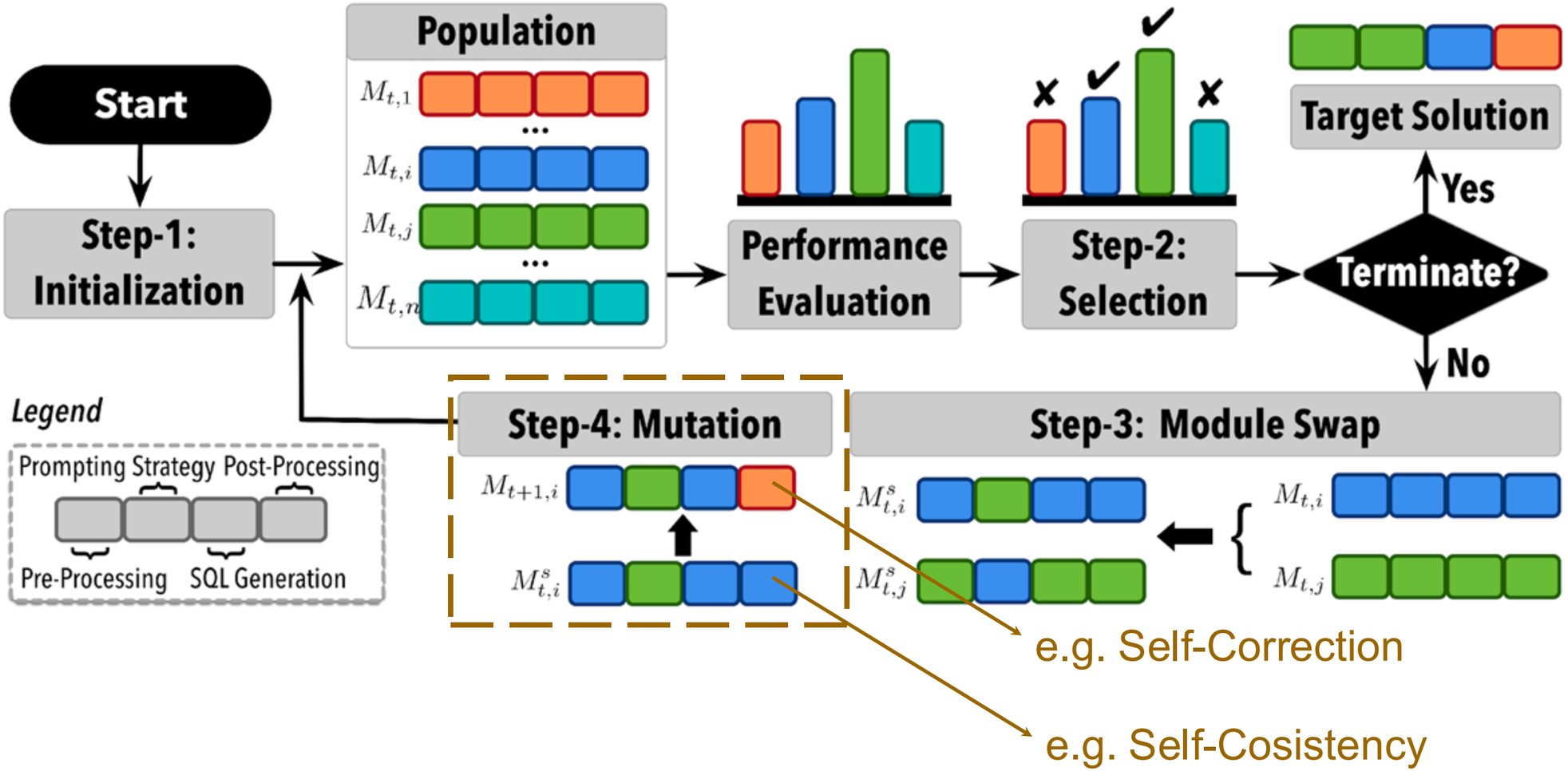
NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



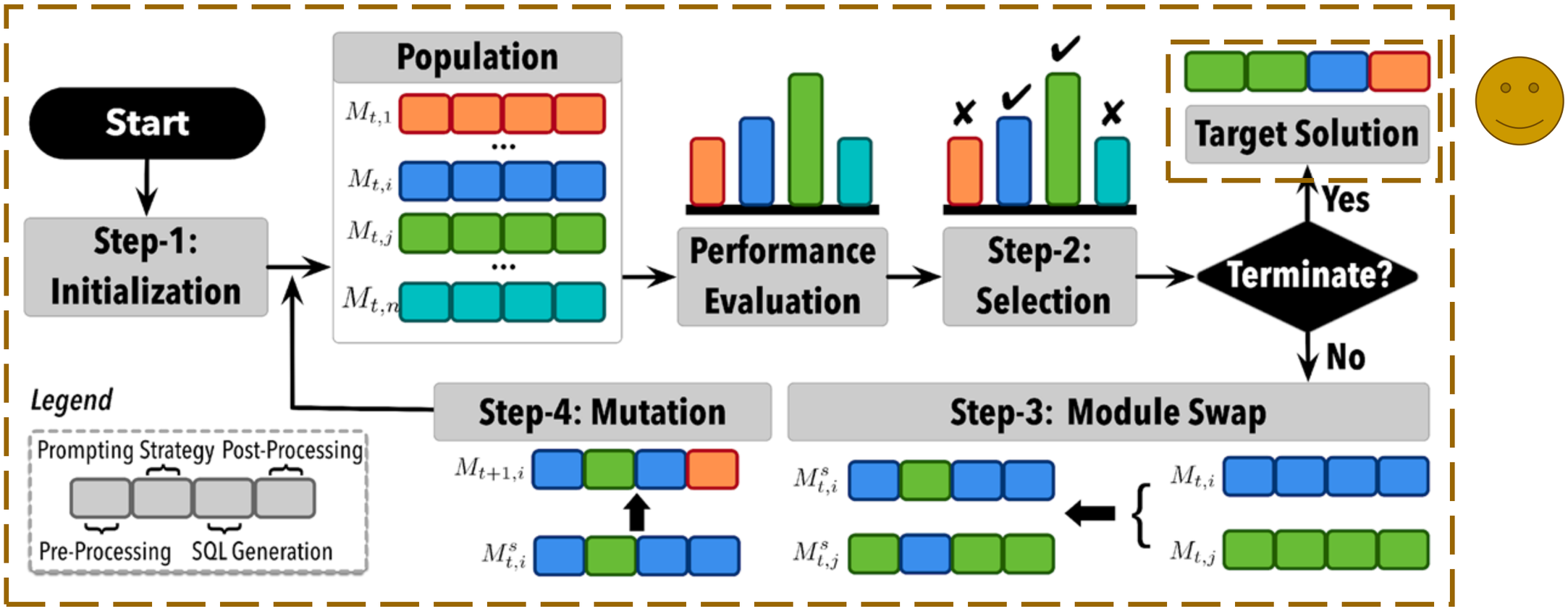
NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



NL2SQL360-AAS:
NL2SQL AUTOMATED ARCHITECTURE SEARCH ALGORITHM



SUPERSQL

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- To validate the effectiveness of the NL2SQL360-AAS algorithm, we conduct a case study.

- Dataset: Spider
- Target Metric: Execution Accuracy (EX)
- Model Backend: GPT-3.5-Turbo
- Predefined Search Space:

NL2SQL Pipeline	Module Candidates
Pre-processing	Schema Linking Module (Source: RESDSQL), DB Content Module (Source: BRIDGE v2)
Prompting	Similarity-based Few-shot Module (Source: DAILSQL)
SQL Generation	Greedy Search Decoding (Limited by OpenAI)
Post-Processing	Self-Correction Module (Source: DINSQL), Self-Consistency Module (Source: C3SQL) Execution-Guided SQL Selector Module (Source: RESDSQL)

SUPERSQL

香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- From the final generation produced by NL2SQL360-AAS, we select the individual with the highest EX metric as our final NL2SQL solution, i.e., **SuperSQL**. Note that we replace the backbone model with GPT-4 for more powerful performance.

■ Take a deep look at prompt improvement compared to DAILSQL.

```
/* Given the following database schema: */
CREATE TABLE airports (
    City text,
    AirportCode text primary key,
    AirportName text
)

CREATE TABLE airlines (
    uid int primary key,
    Airline text
)

/* Answer the following: What are airport names at City
'Aberdeen'?*/
SELECT AirportName FROM airports WHERE City = "Aberdeen";
```

DAILSQL Prompt

```
/* Given the following database schema (with linked
value list after column definition): */
CREATE TABLE airports (
    City text,
    AirportCode text primary key,
    AirportName text -- ["Aberdeen"]
)

CREATE TABLE airlines (
    uid int primary key,
    Airline text
)

/* Answer the following: What are airport names at City
'Aberdeen'?*/
SELECT AirportName FROM airports WHERE City = "Aberdeen";
```

SuperSQL Prompt

SUPERSQL

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Effectiveness of SuperSQL.

Model	Spider-Dev(%)	Spider-Test(%)	BIRD-Dev(%)	BIRD-Test(%)
BRIDGE v2	70.3	-	-	-
RESDSQL-3B	81.8	-	43.9	-
C3SQL	82.0	82.3	50.2	-
DINSQL	82.8	85.3	50.7	55.9
DAILSQL(SC)	83.6	86.6	55.9	57.4
SuperSQL	**87.0** **(3.4↑)** **(17.7↑)**	**87.0** **(0.4↑)** **(4.7↑)**	**58.5** **(2.6↑)** **(14.6↑)**	**62.7** **(5.3↑)** **(6.8↑)**

Table 5: Accuracy vs. LLM Economy on Spider/BIRD Dev Set.

Methods	LLMs	Avg. Tokens / Query		Avg. Cost / Query		EX(%)		EX / Avg. Cost	
		Spider	BIRD	Spider	BIRD	Spider	BIRD	Spider	BIRD
C3SQL	GPT-3.5	5702	5890	0.0103	0.0104	82.0	50.2	7961	4825
DINSQL	GPT-4	9571	-	0.2988	-	82.8	-	277	-
DAILSQL	GPT-4	930	1559	0.0288	0.0486	83.1	54.3	2885	1117
DAILSQL(SC)	GPT-4	1063	1886	0.0377	0.0683	83.6	55.9	2218	819
SuperSQL	**GPT-4**	**942**	**1412**	**0.0354**	**0.0555**	**87.0**	**58.5**	**2458**	**1053**

CONCLUSION

- We proposed a multi-angle testbed, named **NL2SQL360**, for evaluating NL2SQL methods from different perspectives.
- We utilized our NL2SQL360 to evaluate **13** LLM-based and **7** PLM-based NL2SQL methods on **2** widely-used benchmarks, varying **15** settings and deriving **a set of new findings**.
- We employed our NL2SQL360 to analyze the **NL2SQL Design Space** for NL2SQL solutions and automatically search (**NL2SQL360-AAS**) for one of the best solutions, named **SuperSQL**.

THANK YOU & QUESTIONS?

香港科技大学（广州）
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)