



MMSQL: Evaluating and Enhancing LLMs for Multi-turn Text-to-SQL with Multiple Question Types

Ziming Guo, Chao Ma, Yinggang Sun, Tiancheng Zhao, Guangyao Wang, Hai Huang

Harbin University of Science and Technology, Harbin Institute of Technology



INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS
IJCNN2025
30 JUNE - 5 JULY 2025 | **ROME, ITALY**
INTERNATIONAL NEURAL NETWORK SOCIETY



Motivation

The Need for Robust Text-to-SQL Systems

- ▶ Progress in Text-to-SQL Systems: Empowering non-technical users to query databases using natural language.
- ▶ Challenges: Handling **dynamic and uncertain real-world user queries**.
- ▶ Multi-turn conversations
- ▶ Ambiguous or unanswerable questions
- ▶ Current models' limitations leading to unreliable predictions
- ▶ Objective: Develop more robust and dependable text-to-SQL systems.

Motivation

The Need for Robust Text-to-SQL Systems

Database schema given

Airlines

uid:airline id type:number PRIMARY KEY

Airline:airline name type:text

Abbreviation:abbreviation type:text

Country:country type:text

Airports

City:city type:text

AirportCode:airport code type:text PRIMARY KEY

AirportName:airport name type:text

Country:country type:text

CountryAbbrev:country abbrev type:text

Flights

Airline:airline type:number PRIMARY KEY

FlightNo:flight number type:number

SourceAirport:source airport type:text

DestAirport:destination airport type:text

Foreign keys:

flights.'DestAirport' = airports.'AirportCode'

flights.'SourceAirport' = airports.'AirportCode'

Turn1: Unanswerable



How many flights take off **from** **airport APG** each day?



SELECT date, COUNT(*) AS flight_count
FROM flights WHERE departure_airport =
'APG' GROUP BY date



I'm sorry, the database does not include
information about flight schedules or
dates.

Turn2: Answerable



How many airlines have flights
departing from **here**?

SELECT COUNT(T1.Airline) FROM
Airlines AS T1 JOIN Flight AS T2
ON T1.uid = T2.Airline WHERE
T2.sourceAirport like "%APG%"



Turn3: Ambiguous



What is the **flight number of Delta Airlines**?



SELECT T2.FlightNo FROM airlines AS T1
JOIN flights AS T2 ON T1.uid = T2.Airline
WHERE T1.Airline = 'Delta Airlines'



Your question is ambiguous. Are you asking
for **all flight numbers operated by Delta Airlines**, or specifically **those departing from APG**? If you're asking for the former, the SQL query would be [SQL1]. If you're asking for the latter, the SQL query would be [SQL2].

Turn4: Improper



Thanks!

You're welcome! If you have any
more questions, feel free to ask.



Dataset

Introduction to MMSQL: Multi-turn Multi-type text-to-SQL test suit.

- ▶ Question Types: **Answerable / Unanswerable / Improper / Ambiguous**
- ▶ Dataset Generation: QDA-SQL to generate multi-turn Q&A pairs.
- ▶ Original data: Transformed samples from SPARC and CoSQL.
- ▶ Dataset Characteristics: Complex multi-turn and multi-type scenarios.
- ▶ Download (dataset and evaluation scripts):



Github:

- mcxiaoxiao.github.io/MMSQL



Huggingface:

- `from datasets import load_dataset`
- `ds = load_dataset("GuoZiming/MMSQL")`

	SParC	CoSQL	NoisySP	AmbiQT	AMBROSIA	MMSQL
# Dialogues	4,298	3,007	-	-	-	6493
Total # turns	12,726	15,433	15,598	3,046	4,242	38,666
Avg. # Q turns	3.0	5.2	1	1	1	6.0
Avg. Q len	10.2	11.2	-	-	-	11.4
Ans. Q type	✓	✓	✗	✗	✓	✓
Amb. Q type	✗	✓	✓	✓	✓	✓
Una. Q type	✗	✓	✓	✗	✗	✓
Imp. Q type	✗	✓	✗	✗	✗	✓



MMSQL Test Set (149)



MMSQL Train Set (6493)

Evaluation Metrics

Metrics Overview

- ▶ **Exact Matching (EM)**: All components of the predicted SQL must match the reference query (excluding values).
- ▶ **Execution Accuracy (EX)**: Proportion of SQL queries with identical execution results.
- ▶ **Interaction Execution Accuracy (IEX)**: All SQL queries in a multi-turn interaction execute correctly.
- ▶ **Dual Assessment of Question Type Detection and Execution Accuracy (TDEX)**: Evaluates both question type classification and execution accuracy.
- ▶ **Response Quality Score (RQS)**: Measures the quality of natural language responses using an LLM-assisted rating method.

$$\text{TDEX} = \frac{1}{N} \sum_{i=1}^N \begin{cases} \varepsilon_{\text{exec}}(S_i, \hat{S}_i) & \text{(a)} \\ \delta_{\text{type}}(C_i, \hat{C}_i) & \text{(b)} \end{cases}$$

- (a) $C_i = \text{'Answerable' or 'Ambiguous'}$
(b) otherwise

Model Performance

Key Insights

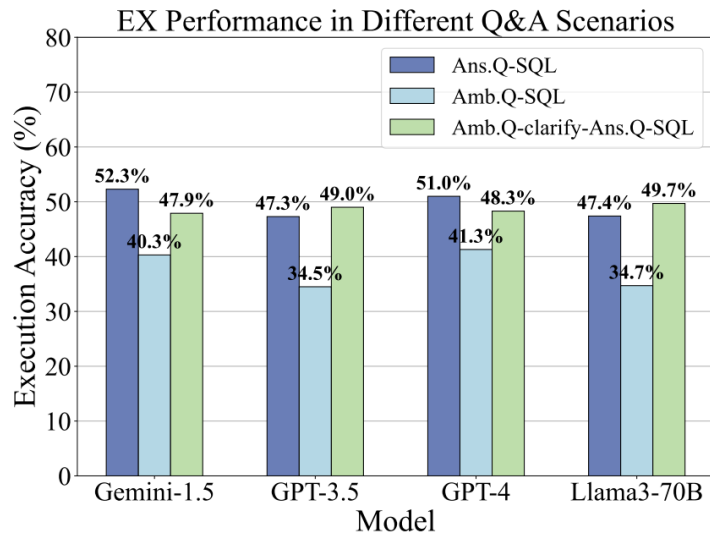
- ▶ Closed-source models (e.g., GPT-4 Turbo) have a slight performance edge.
- ▶ Open-source models (e.g., Llama3-70B) are rapidly catching up.
- ▶ All models struggle with unanswerable and ambiguous questions.

Model	Ans.		Una.		Amb.		Imp.		F1
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	
GPT-4 Turbo	<u>90.3</u>	89.8	25.9	<u>70.0</u>	56.9	38.4	100.0	<u>98.0</u>	68.2
GPT-3.5 Turbo	88.8	89.2	16.0	65.0	64.3	20.9	100.0	96.0	61.1
Gemini-1.5 Flash	85.9	<u>95.8</u>	26.3	50.0	58.3	8.1	100.0	95.4	59.3
Llama3-70B	84.9	95.2	<u>27.5</u>	55.0	80.0	9.3	100.0	92.7	59.8
Llama3-8B	88.3	93.2	21.4	60.0	83.3	23.3	100.0	96.7	<u>64.2</u>
SQLCoder-8B	84.6	99.6	77.8	35.0	0.0	0.0	100.0	<u>98.0</u>	59.7
Codellama-7B	93.9	16.5	4.3	85.0	<u>96.6</u>	66.3	56.8	100.0	46.9
Mistral-7B-v0.2	82.1	57.6	4.7	55.0	100.0	<u>50.0</u>	79.1	77.5	55.3

Model Performance

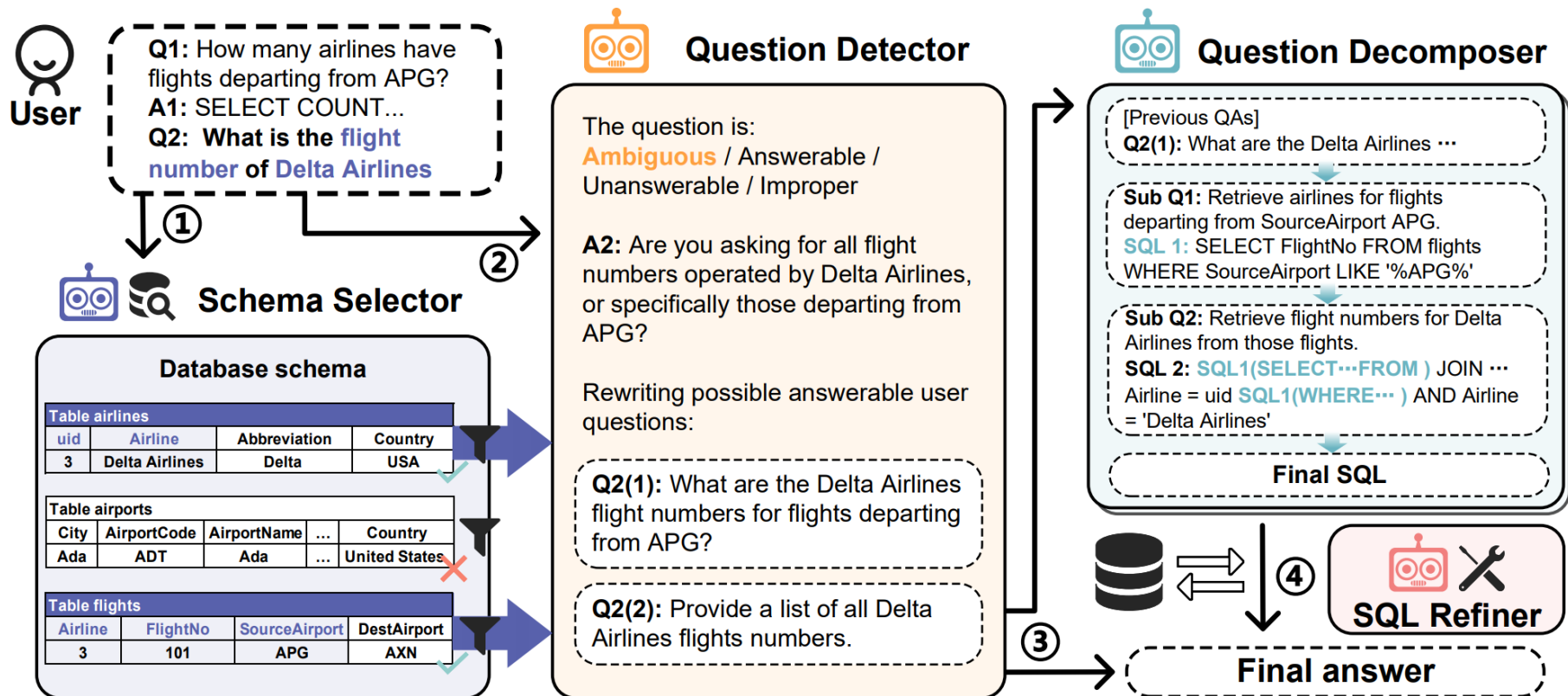
Key Insights

- ▶ Closed-source models (e.g., GPT-4 Turbo) have a slight performance edge.
- ▶ Open-source models (e.g., Llama3-70B) are rapidly catching up.
- ▶ All models struggle with unanswerable and ambiguous questions.
- ▶ **Clarification processes in multi-turn interactions improve query precision.**



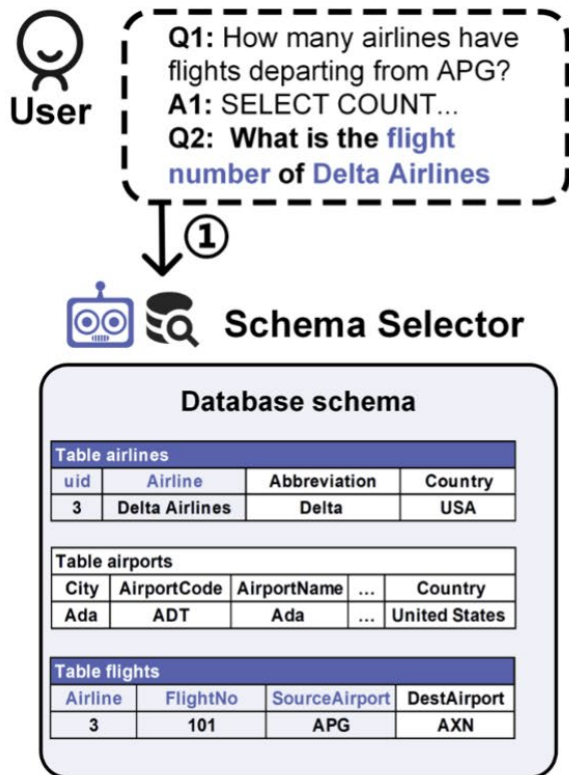
Multi-Agent Framework

Framework Overview



Multi-Agent Framework

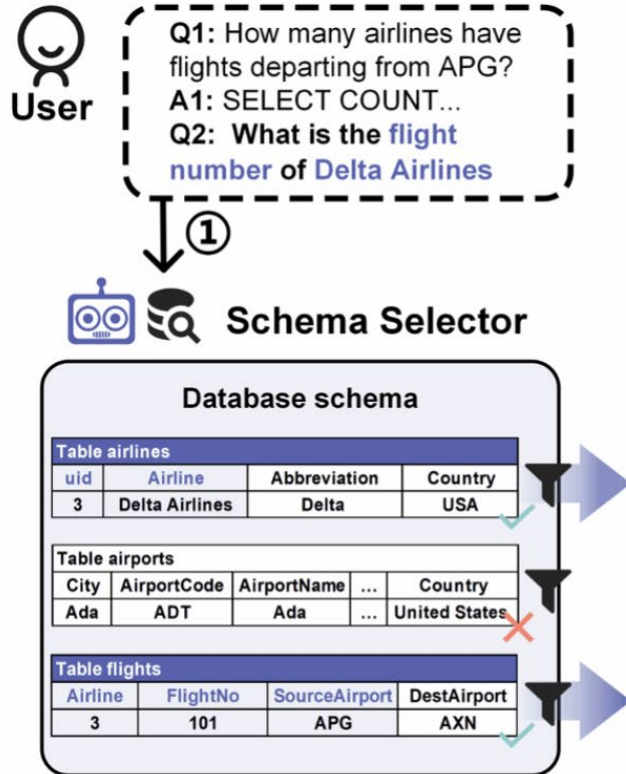
Framework Overview



Select relevant tables and fields

Multi-Agent Framework

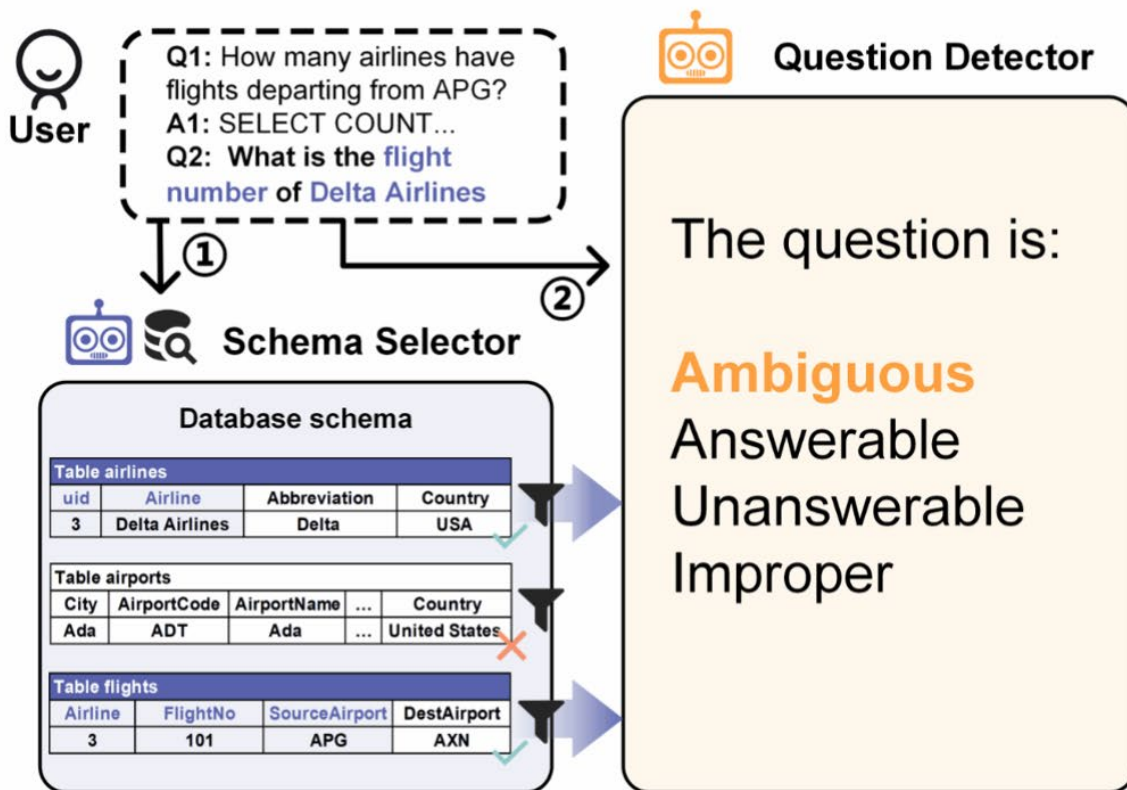
Framework Overview



Only relevant information will be used for subsequent reasoning.

Multi-Agent Framework

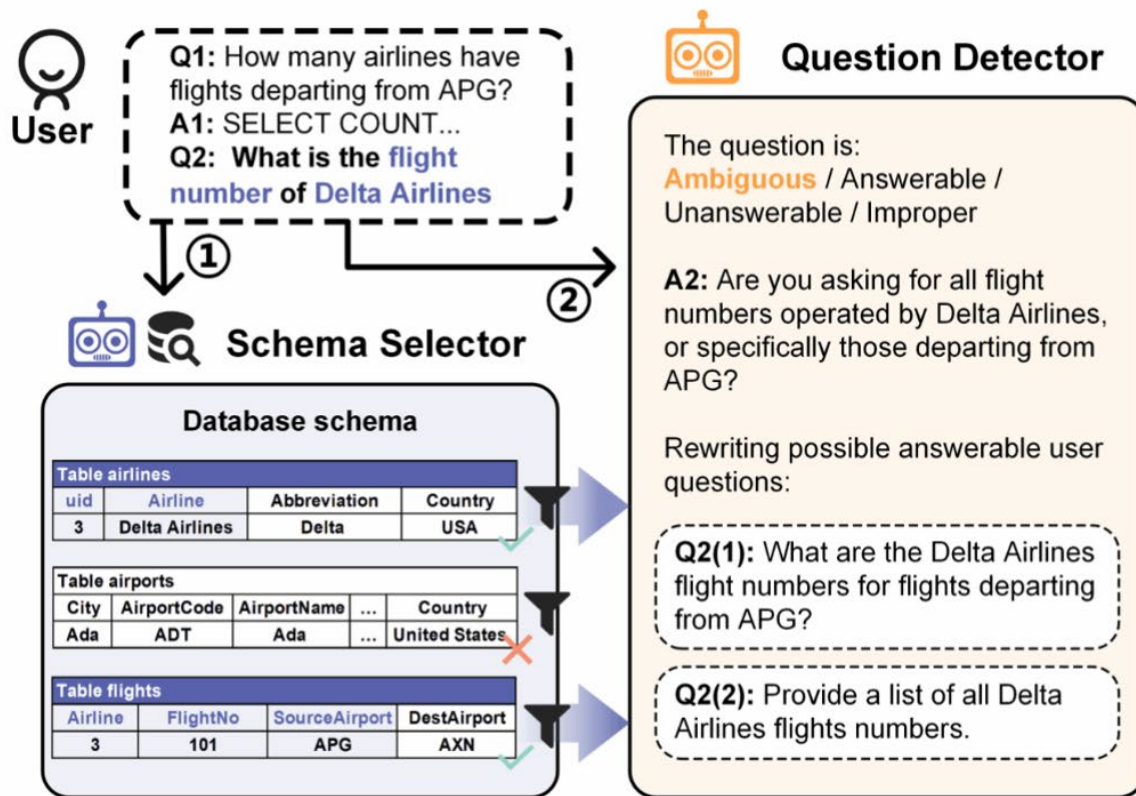
Framework Overview



Identify the question type and decide on subsequent strategy.

Multi-Agent Framework

Framework Overview

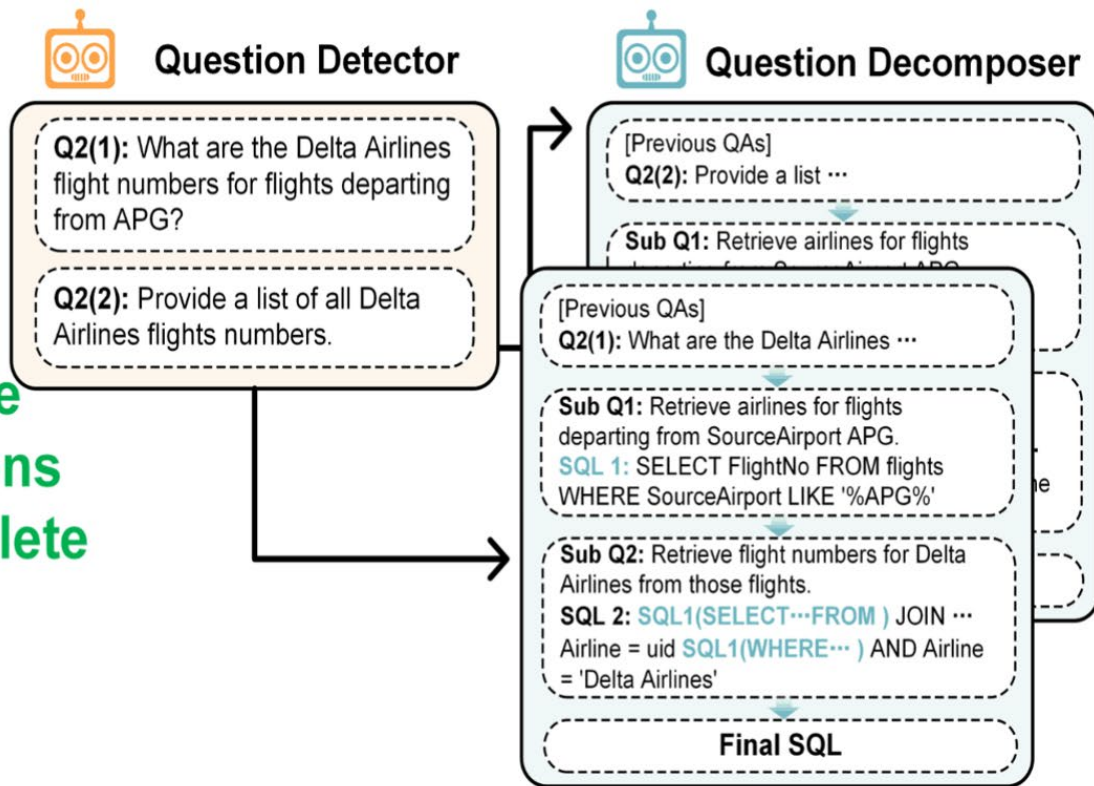


Identify ambiguities and rewrite possible answerable user questions

Multi-Agent Framework

Framework Overview

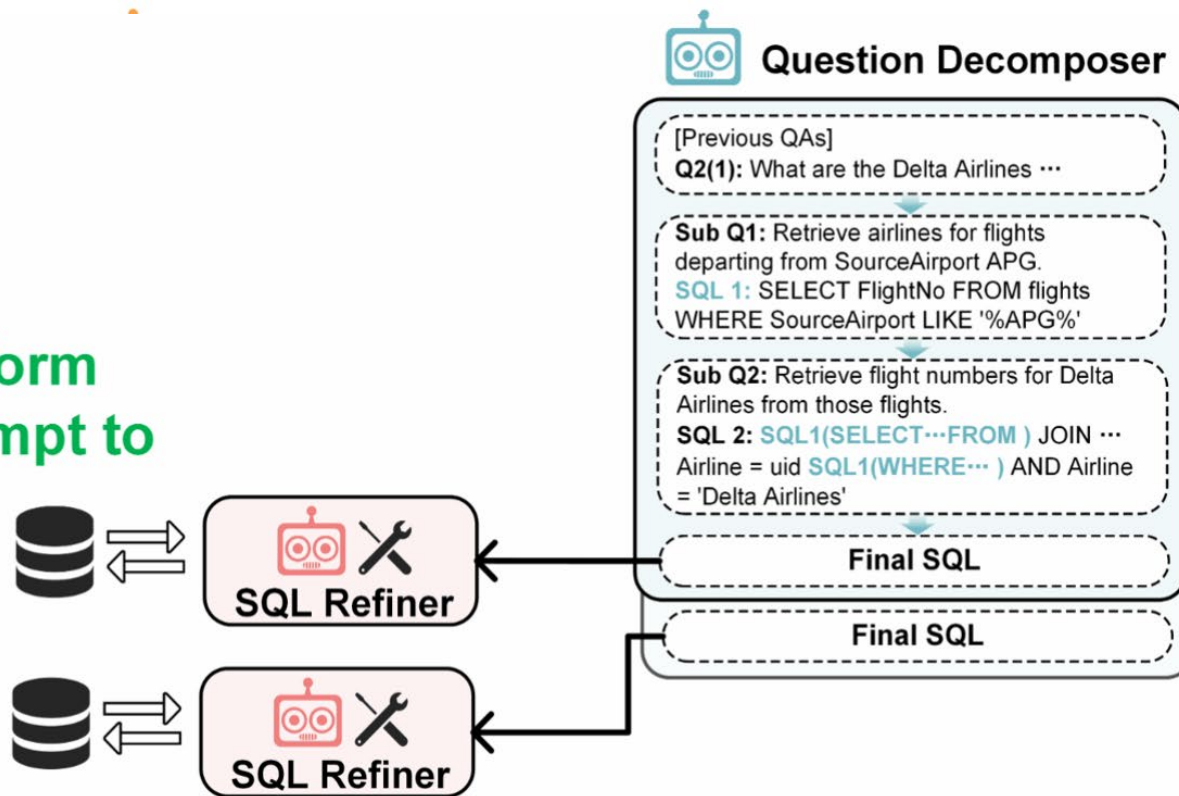
For answerable questions, decompose them into sub-questions and provide the complete SQL step by step.



Multi-Agent Framework

Framework Overview

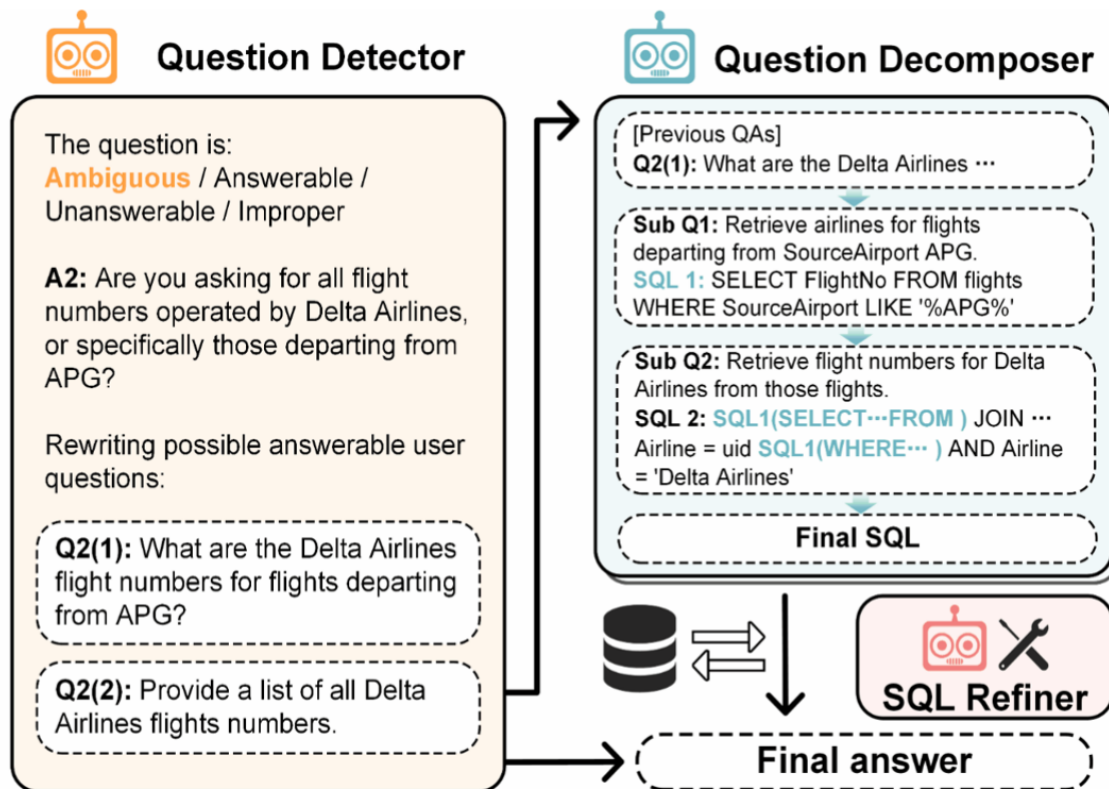
For each SQL, perform validation and attempt to correct it.



Multi-Agent Framework

Framework Overview

Combine and obtain
the final answer. 🎉



Summary of Findings

Evaluating and Enhancing LLMs for Multi-turn Text-to-SQL with Multiple Question Types

Ziming Guo^{ID*}, Chao Ma^{*}, Yinggang Sun[†], Tiancheng Zhao^{*}, Guangyao Wang^{ID*}, Hai Huang^{*}

*School of Computer Science, Harbin University of Science and Technology

Harbin, China 150040

Email: 2204050108@stu.hrbust.edu.cn, machao8396@163.com,

1783467143@qq.com, 2204050124@stu.hrbust.edu.cn, hust_hh@vip.163.com

[†] Faculty of Computing, Harbin Institute of Technology, Harbin, China 150001

Email: 23b903085@stu.hit.edu.cn



Real-world challenges&insight

Datasets

Multi-Agent Framework

Q&A

Engage with Us

- ▶ **Any questions?** orlosziming@outlook.com
- ▶ **Want to contribute?** Submit issues or PRs to our repo



github.com/mcxiaoxiao/MMSQL