ToolQA: A Dataset for LLM Question Answering with External Tools

Yuchen Zhuang[*], Yue Yu[*], Kuan Wang[*], Haotian Sun, Chao Zhang

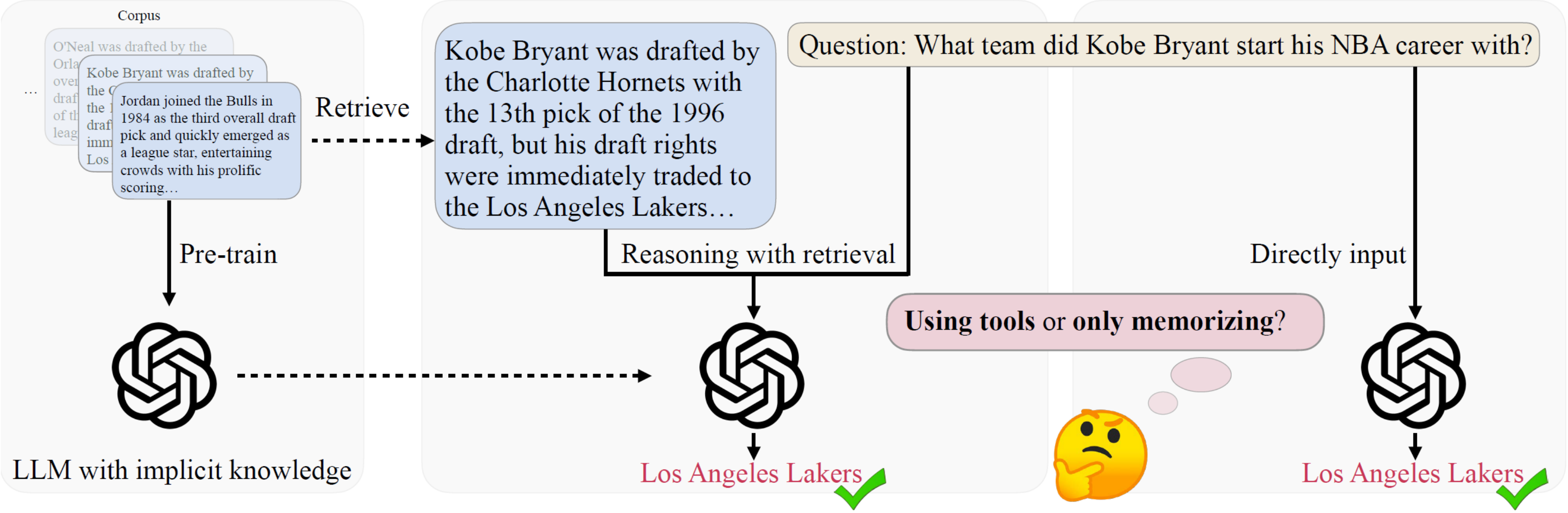
Georgia Institute of Technology

Contact: yczhuang@gatech.edu

Nov 11th, 2023

Internal Knowledge vs. External Tools

- Pre-trained on vast range of corpus, LLMs possess extensive knowledge, which may overlap with evaluation data.
- We propose ToolQA to discern whether the model is merely recalling pre-trained information or genuinely employing external tools for problem-solving.



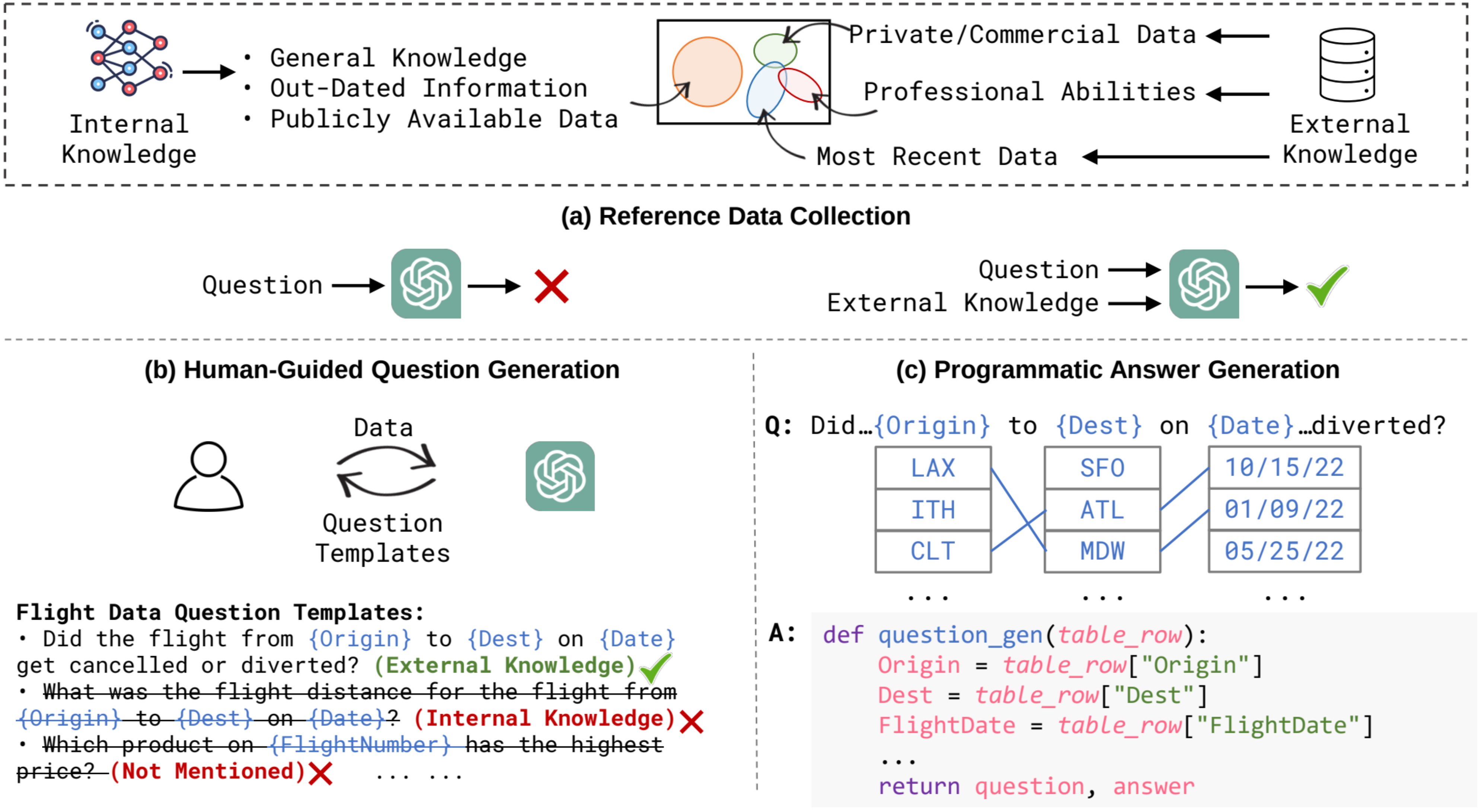
Jordan joined the Bulls in 1984 as the third overall draft pick and quickly emerged as a league star, entertaining crowds with his prolific scoring…

Kobe Bryant was drafted by the Charlotte Hornets with the 13th pick of the 1996 draft, but his draft rights were immediately traded to the Los Angeles Lakers…

Question: What team did Kobe Bryant start his NBA career with?

Using tools or **only memorizing**?

Curation of ToolQA Dataset



Internal Knowledge

- General Knowledge
- Out-Dated Information
- Publicly Available Data

Private/Commercial Data

Professional Abilities

Most Recent Data

External Knowledge

(a) Reference Data Collection

Question → ✗

Question
External Knowledge → ✓

(b) Human-Guided Question Generation

Data

Question Templates

Flight Data Question Templates:

- Did the flight from {Origin} to {Dest} on {Date} get cancelled or diverted? **(External Knowledge)** ✓
- ~~What was the flight distance for the flight from {Origin} to {Dest} on {Date}~~? **(Internal Knowledge)** ✗
- ~~Which product on {FlightNumber} has the highest price?~~ **(Not Mentioned)** ✗ ... ...

(c) Programmatic Answer Generation

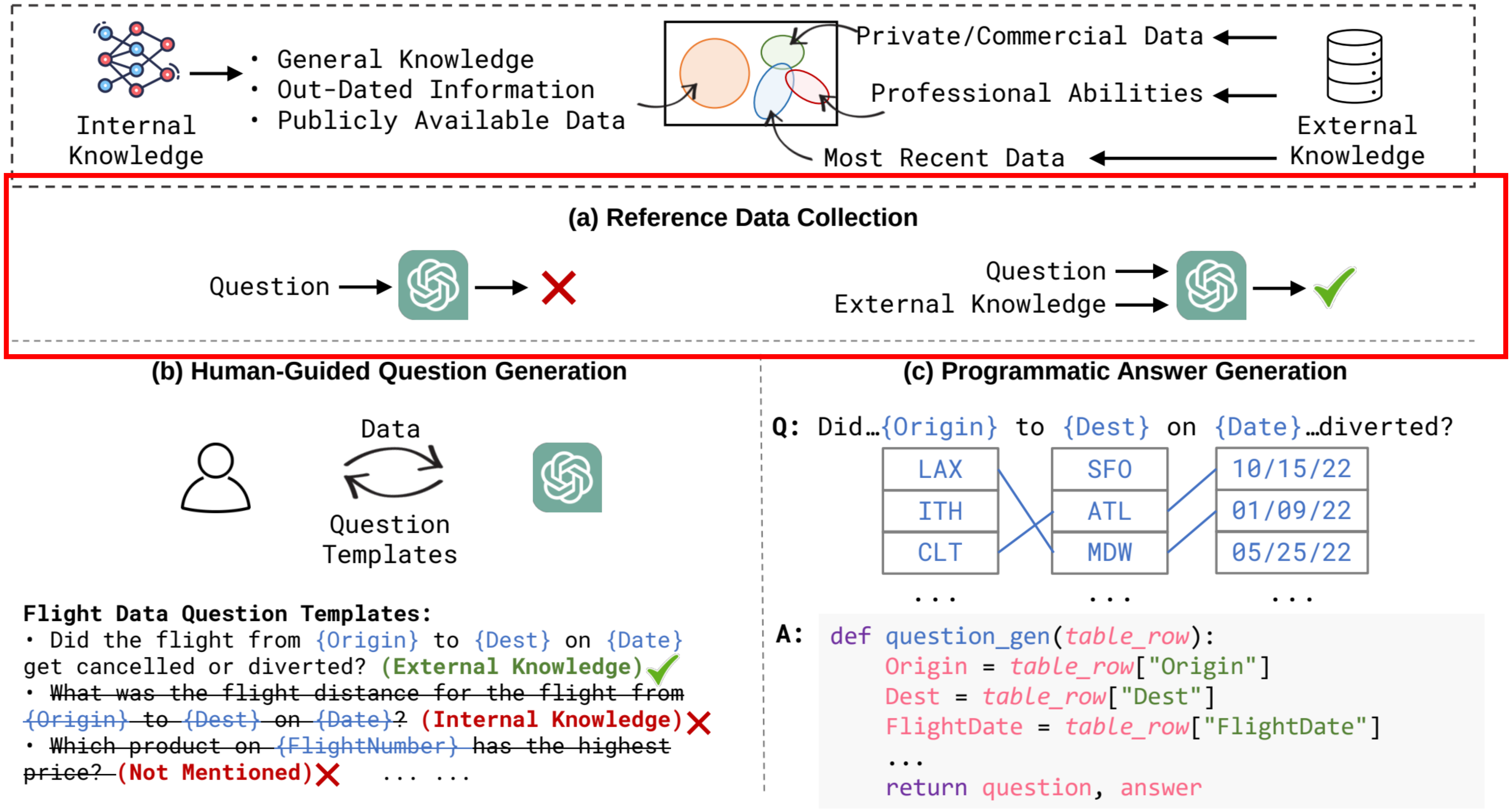
Q: Did…{Origin} to {Dest} on {Date}…diverted?

LAX	SFO	10/15/22
ITH	ATL	01/09/22
CLT	MDW	05/25/22
...	...	...

A:

```
def question_gen(table_row):
    Origin = table_row["Origin"]
    Dest = table_row["Dest"]
    FlightDate = table_row["FlightDate"]
    ...
    return question, answer
```

Curation of ToolQA Dataset



Flight Data Question Templates:

- Did the flight from {Origin} to {Dest} on {Date} get cancelled or diverted? **(External Knowledge)** ✓
- ~~What was the flight distance for the flight from {Origin} to {Dest} on {Date}?~~ **(Internal Knowledge)** ✗
- ~~Which product on {FlightNumber} has the highest price?~~ **(Not Mentioned)** ✗ ... ...

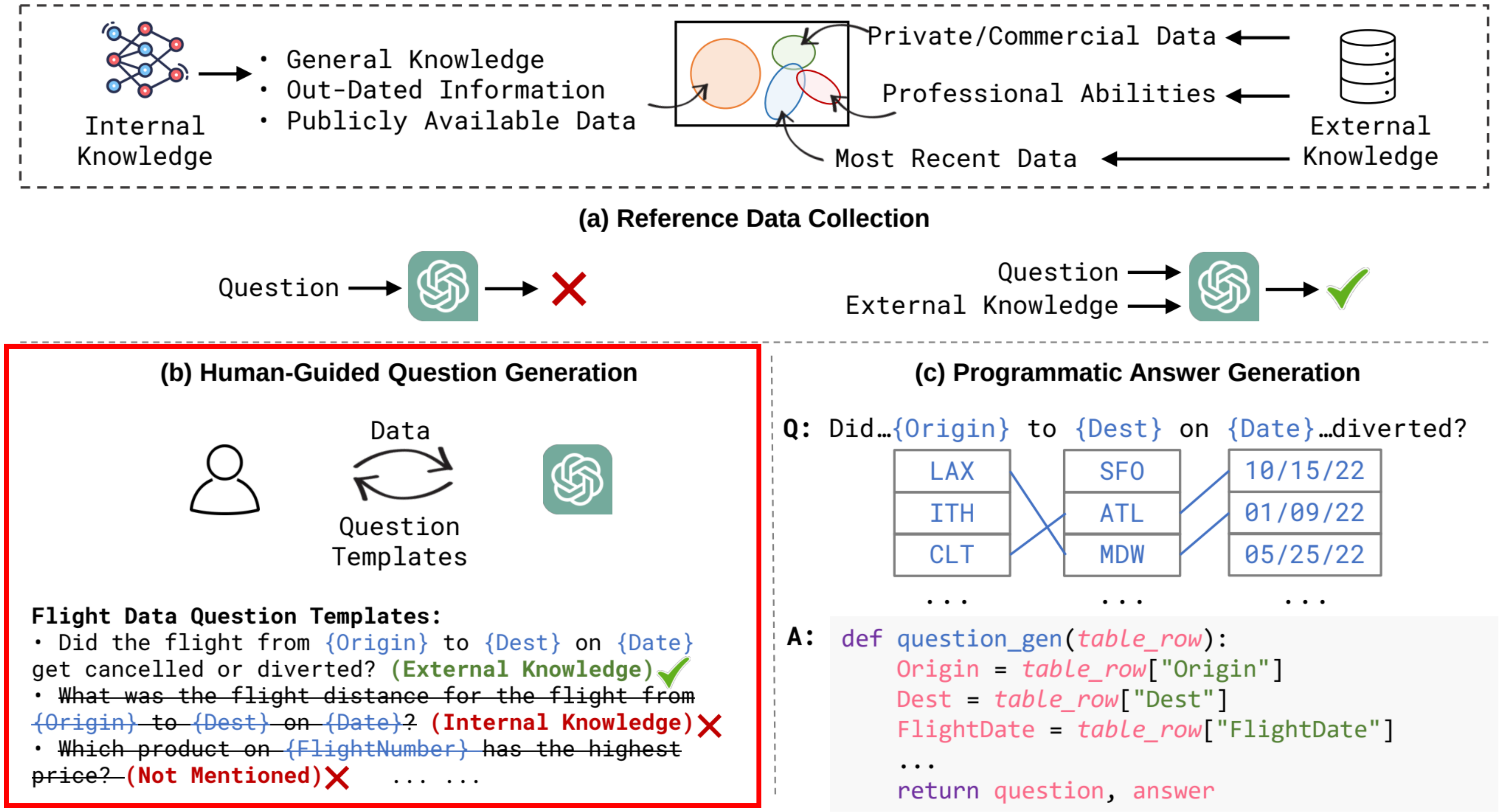
Q: Did…{Origin} to {Dest} on {Date}…diverted?

LAX	SFO	10/15/22
ITH	ATL	01/09/22
CLT	MDW	05/25/22
...	...	...

A:

```
def question_gen(table_row):
    Origin = table_row["Origin"]
    Dest = table_row["Dest"]
    FlightDate = table_row["FlightDate"]
    ...
    return question, answer
```

Curation of ToolQA Dataset



Internal Knowledge →

- General Knowledge
- Out-Dated Information
- Publicly Available Data

Private/Commercial Data ←

Professional Abilities ←

Most Recent Data ←

External Knowledge

(a) Reference Data Collection

Question → ✗

Question → External Knowledge → ✓

(b) Human-Guided Question Generation

Data

Question Templates

Flight Data Question Templates:

- Did the flight from {Origin} to {Dest} on {Date} get cancelled or diverted? **(External Knowledge)** ✓
- ~~What was the flight distance for the flight from {Origin} to {Dest} on {Date}?~~ **(Internal Knowledge)** ✗
- ~~Which product on {FlightNumber} has the highest price?~~ **(Not Mentioned)** ✗ ... ...

(c) Programmatic Answer Generation

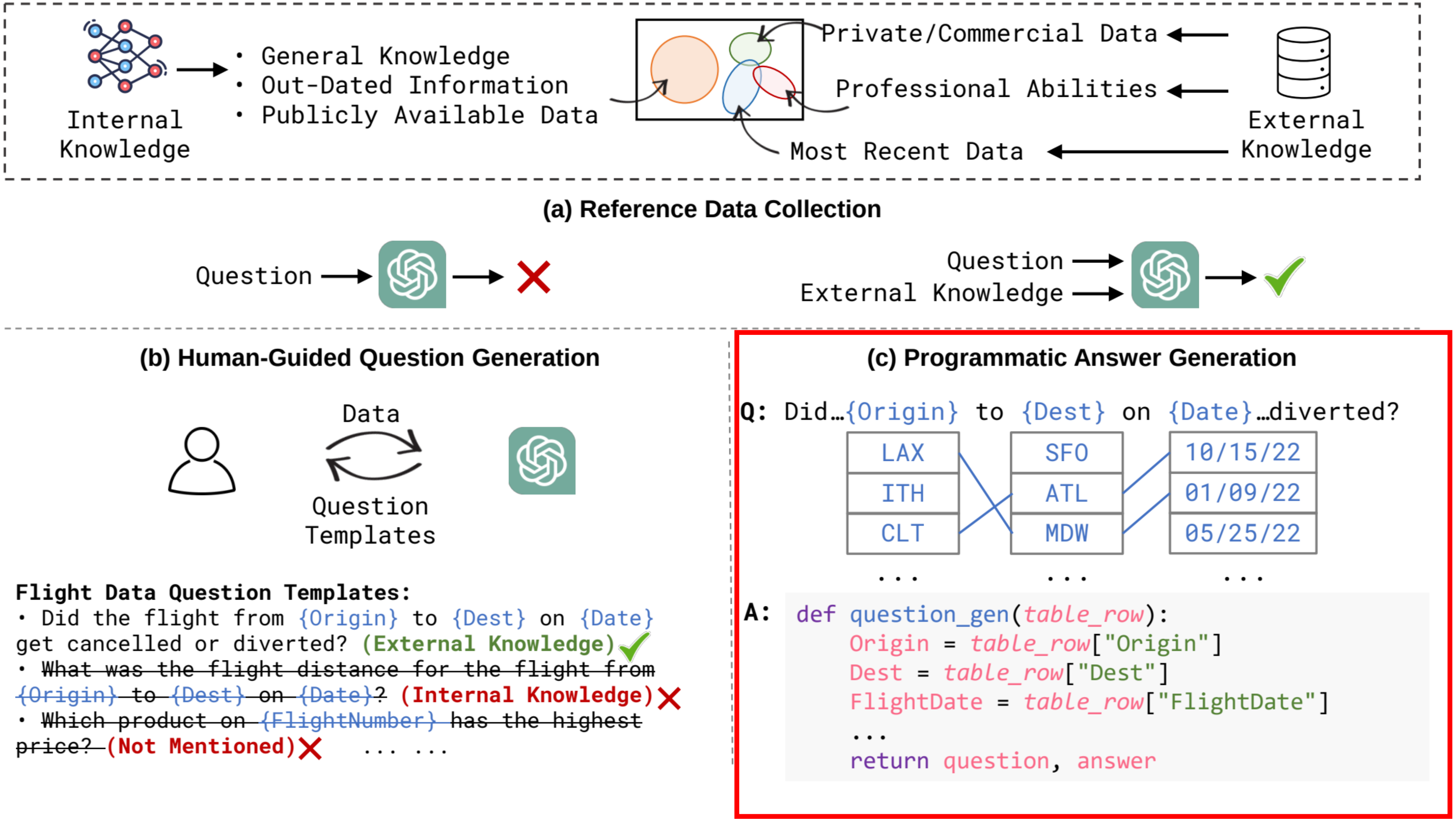
Q: Did…{Origin} to {Dest} on {Date}…diverted?

LAX	SFO	10/15/22
ITH	ATL	01/09/22
CLT	MDW	05/25/22
...	...	...

A:

```
def question_gen(table_row):
    Origin = table_row["Origin"]
    Dest = table_row["Dest"]
    FlightDate = table_row["FlightDate"]
    ...
    return question, answer
```

Curation of ToolQA Dataset



(a) Reference Data Collection

(b) Human-Guided Question Generation

Flight Data Question Templates:

- Did the flight from {Origin} to {Dest} on {Date} get cancelled or diverted? **(External Knowledge)** ✓
- ~~What was the flight distance for the flight from {Origin} to {Dest} on {Date}?~~ **(Internal Knowledge)** ✗
- ~~Which product on {FlightNumber} has the highest price?~~ **(Not Mentioned)** ✗ ... ...

(c) Programmatic Answer Generation

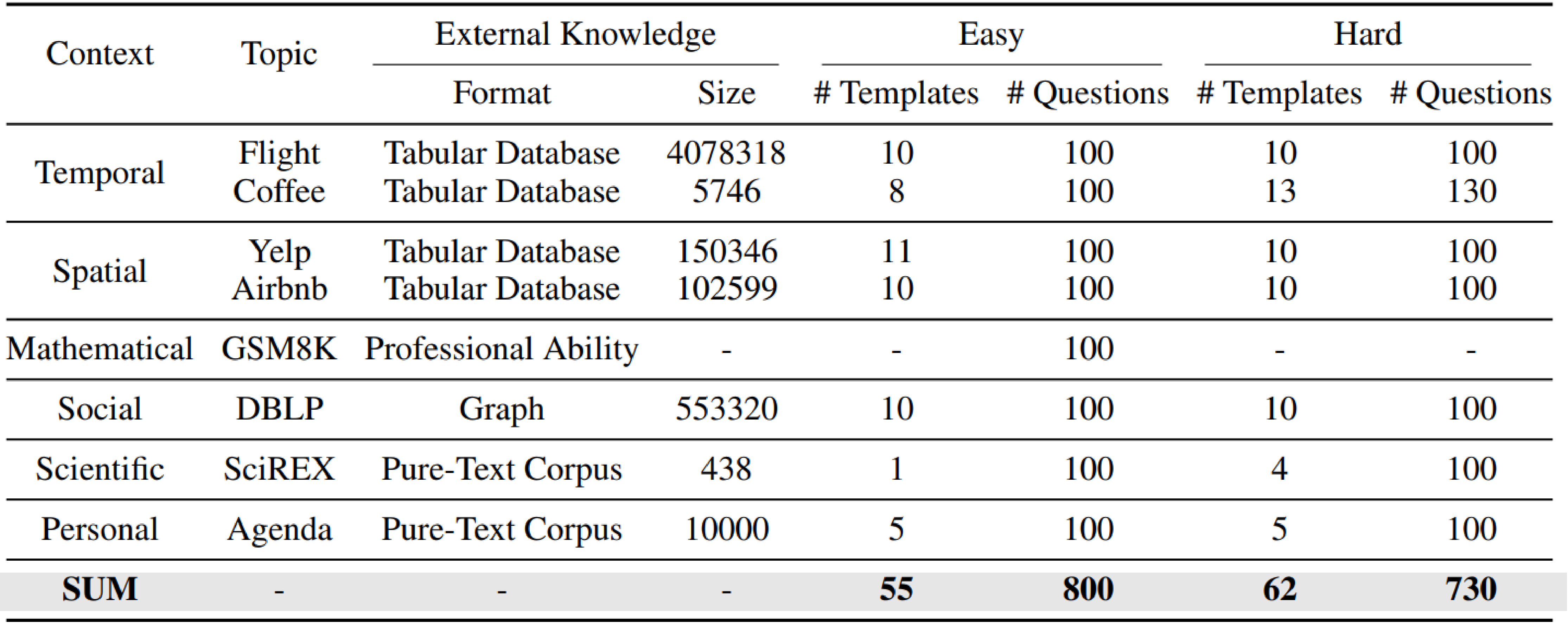
Q: Did…{Origin} to {Dest} on {Date}…diverted?

LAX	SFO	10/15/22
ITH	ATL	01/09/22
CLT	MDW	05/25/22
...	...	...

A:

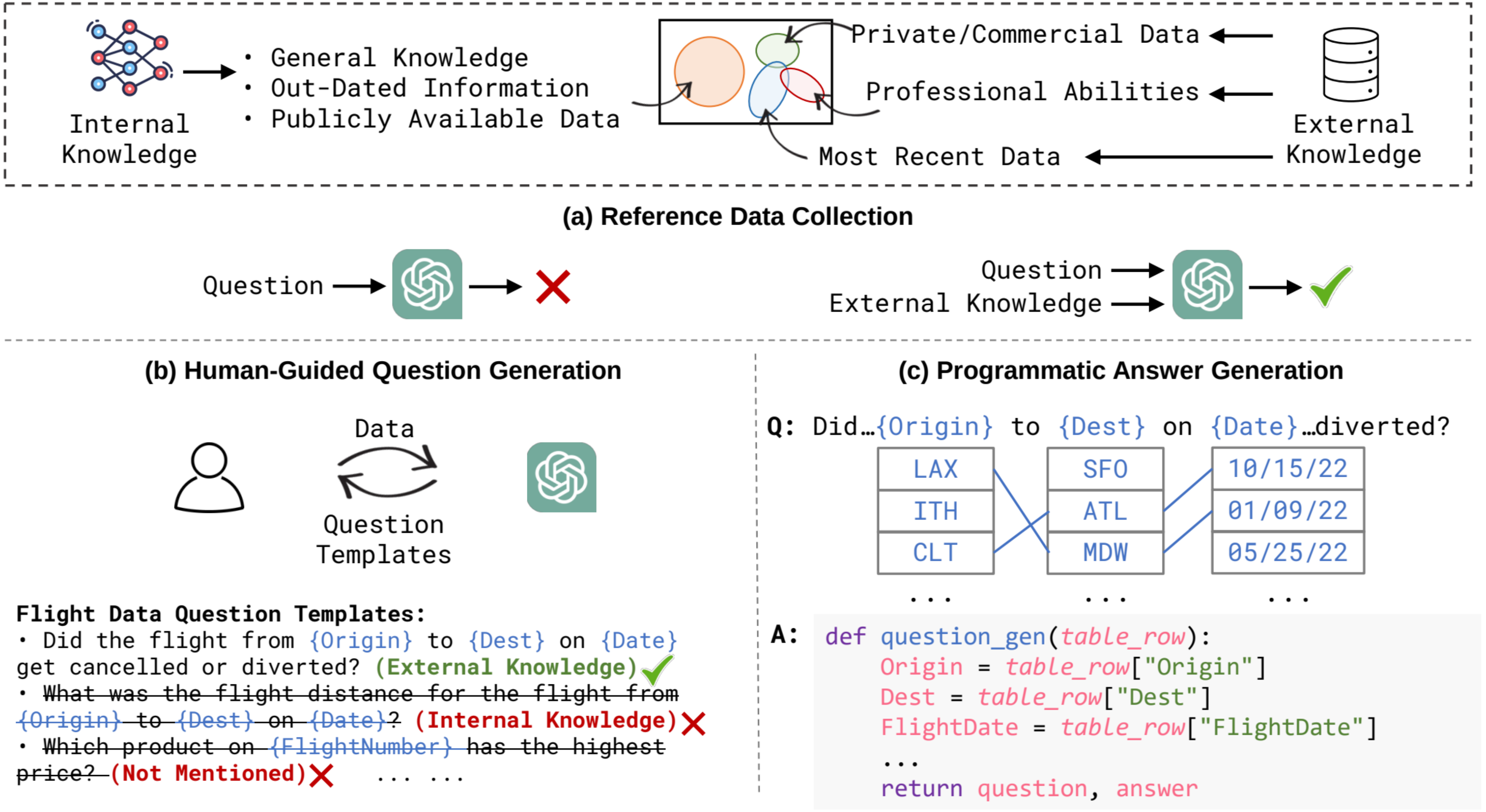
```
def question_gen(table_row):
    Origin = table_row["Origin"]
    Dest = table_row["Dest"]
    FlightDate = table_row["FlightDate"]
    ...
    return question, answer
```

Data Sources



Context	Topic	External Knowledge		Easy		Hard	
		Format	Size	# Templates	# Questions	# Templates	# Questions
Temporal	Flight	Tabular Database	4078318	10	100	10	100
	Coffee	Tabular Database	5746	8	100	13	130
Spatial	Yelp	Tabular Database	150346	11	100	10	100
	Airbnb	Tabular Database	102599	10	100	10	100
Mathematical	GSM8K	Professional Ability	-	-	100	-	-
Social	DBLP	Graph	553320	10	100	10	100
Scientific	SciREX	Pure-Text Corpus	438	1	100	4	100
Personal	Agenda	Pure-Text Corpus	10000	5	100	5	100
SUM	-	-	-	**55**	**800**	**62**	**730**

Curation of ToolQA Dataset



(a) Reference Data Collection

(b) Human-Guided Question Generation

Flight Data Question Templates:

- Did the flight from {Origin} to {Dest} on {Date} get cancelled or diverted? **(External Knowledge)** ✓
- ~~What was the flight distance for the flight from {Origin} to {Dest} on {Date}?~~ **(Internal Knowledge)** ✗
- ~~Which product on {FlightNumber} has the highest price?~~ **(Not Mentioned)** ✗ ... ...

(c) Programmatic Answer Generation

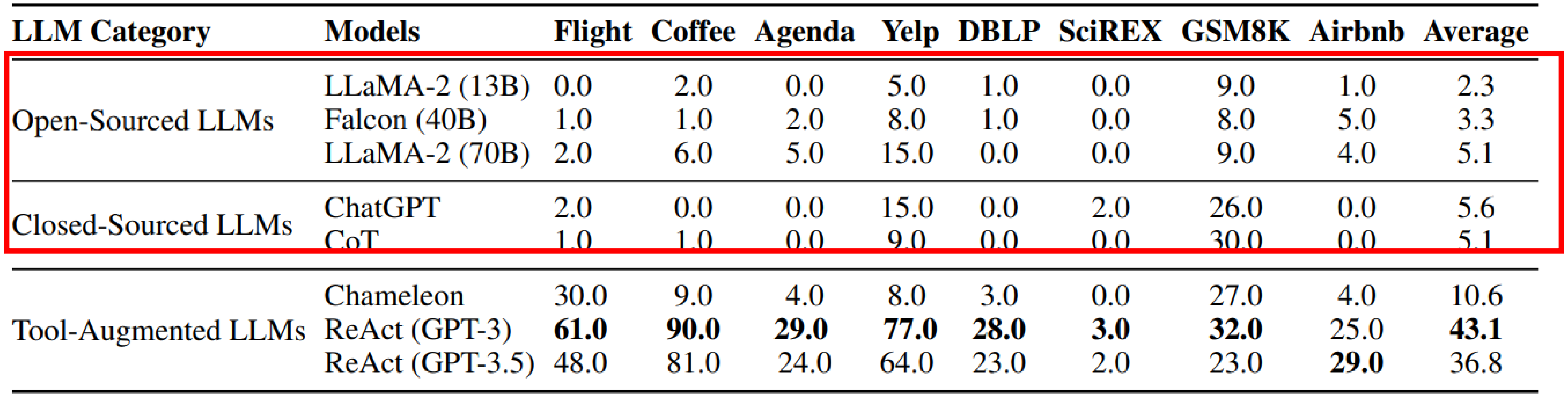
Q: Did…{Origin} to {Dest} on {Date}…diverted?

LAX	SFO	10/15/22
ITH	ATL	01/09/22
CLT	MDW	05/25/22
...	...	...

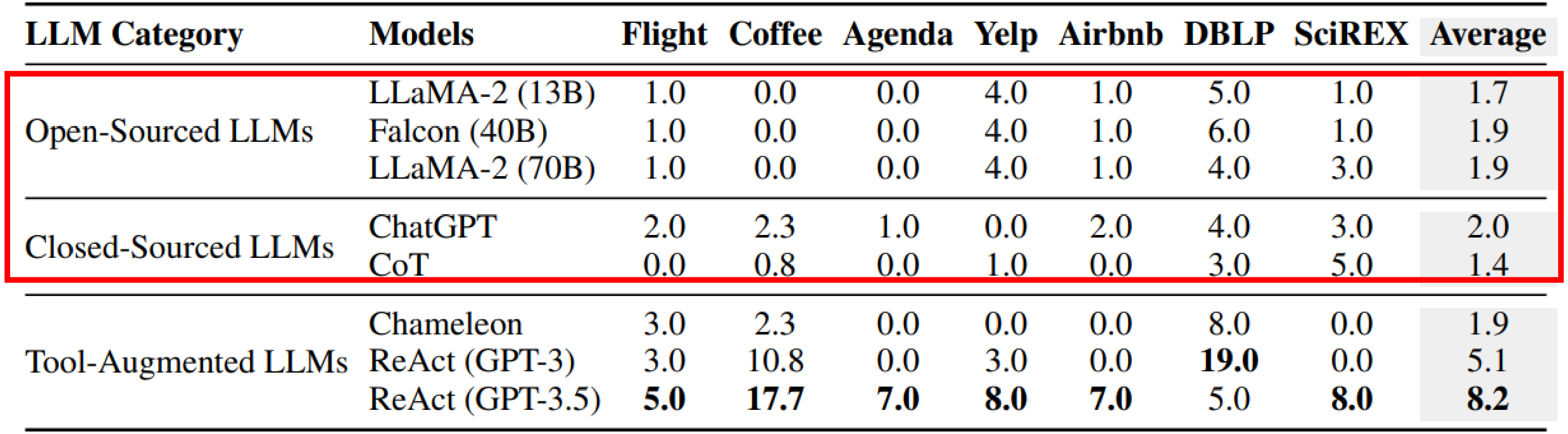
A:

```
def question_gen(table_row):
    Origin = table_row["Origin"]
    Dest = table_row["Dest"]
    FlightDate = table_row["FlightDate"]
    ...
    return question, answer
```

Main Results

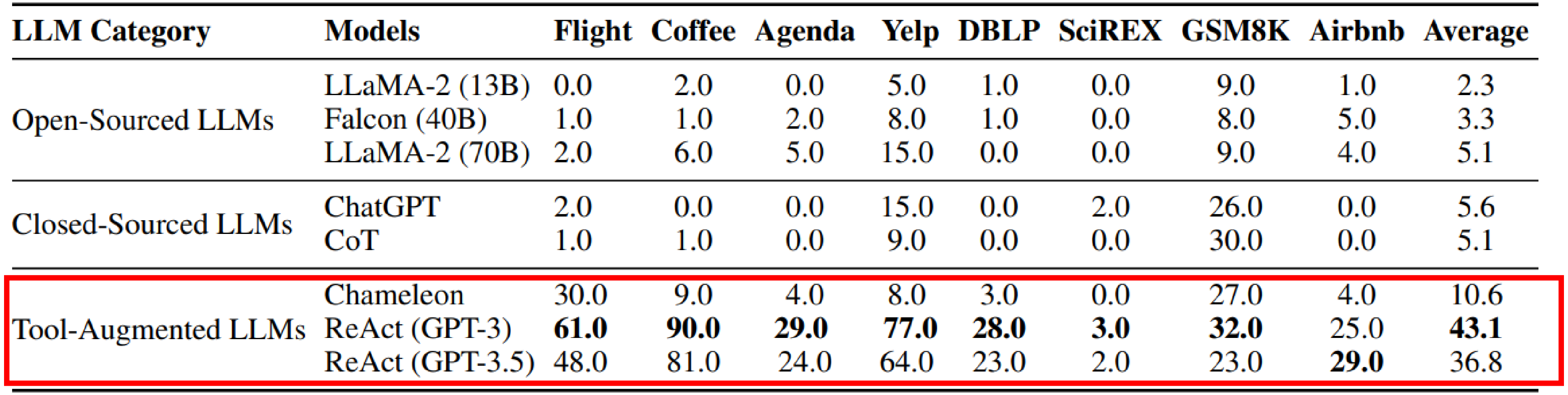


LLM Category	Models	Flight	Coffee	Agenda	Yelp	DBLP	SciREX	GSM8K	Airbnb	Average
Open-Sourced LLMs	LLaMA-2 (13B)	0.0	2.0	0.0	5.0	1.0	0.0	9.0	1.0	2.3
	Falcon (40B)	1.0	1.0	2.0	8.0	1.0	0.0	8.0	5.0	3.3
	LLaMA-2 (70B)	2.0	6.0	5.0	15.0	0.0	0.0	9.0	4.0	5.1
Closed-Sourced LLMs	ChatGPT	2.0	0.0	0.0	15.0	0.0	2.0	26.0	0.0	5.6
	CoT	1.0	1.0	0.0	9.0	0.0	0.0	30.0	0.0	5.1
Tool-Augmented LLMs	Chameleon	30.0	9.0	4.0	8.0	3.0	0.0	27.0	4.0	10.6
	ReAct (GPT-3)	**61.0**	**90.0**	**29.0**	**77.0**	**28.0**	**3.0**	**32.0**	25.0	**43.1**
	ReAct (GPT-3.5)	48.0	81.0	24.0	64.0	23.0	2.0	23.0	**29.0**	36.8

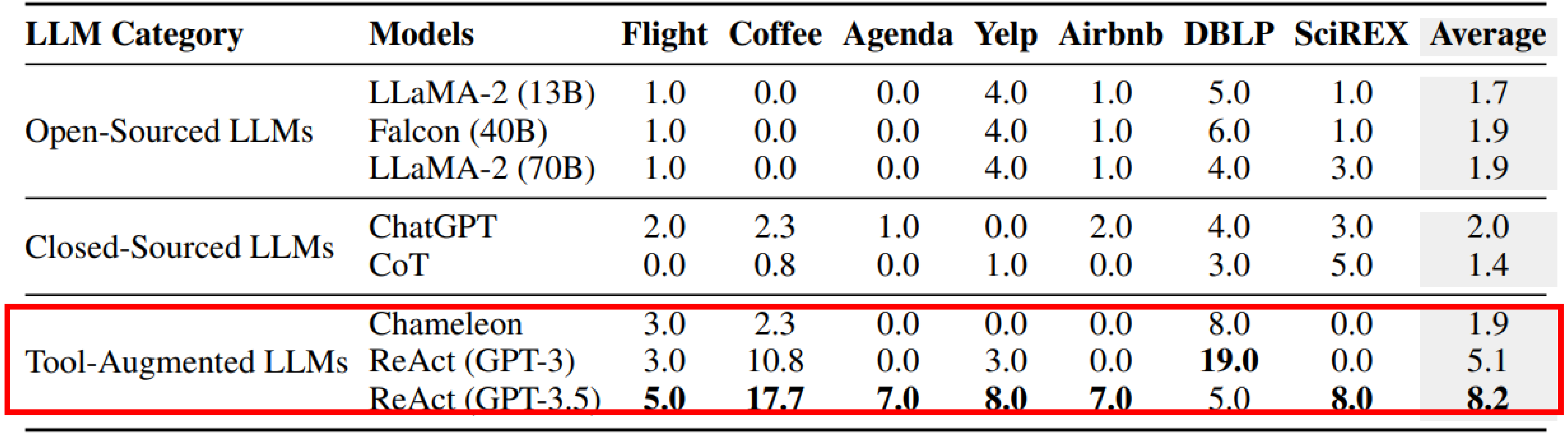


LLM Category	Models	Flight	Coffee	Agenda	Yelp	Airbnb	DBLP	SciREX	Average
Open-Sourced LLMs	LLaMA-2 (13B)	1.0	0.0	0.0	4.0	1.0	5.0	1.0	1.7
	Falcon (40B)	1.0	0.0	0.0	4.0	1.0	6.0	1.0	1.9
	LLaMA-2 (70B)	1.0	0.0	0.0	4.0	1.0	4.0	3.0	1.9
Closed-Sourced LLMs	ChatGPT	2.0	2.3	1.0	0.0	2.0	4.0	3.0	2.0
	CoT	0.0	0.8	0.0	1.0	0.0	3.0	5.0	1.4
Tool-Augmented LLMs	Chameleon	3.0	2.3	0.0	0.0	0.0	8.0	0.0	1.9
	ReAct (GPT-3)	3.0	10.8	0.0	3.0	0.0	**19.0**	0.0	5.1
	ReAct (GPT-3.5)	**5.0**	**17.7**	**7.0**	**8.0**	**7.0**	5.0	**8.0**	**8.2**

Main Results

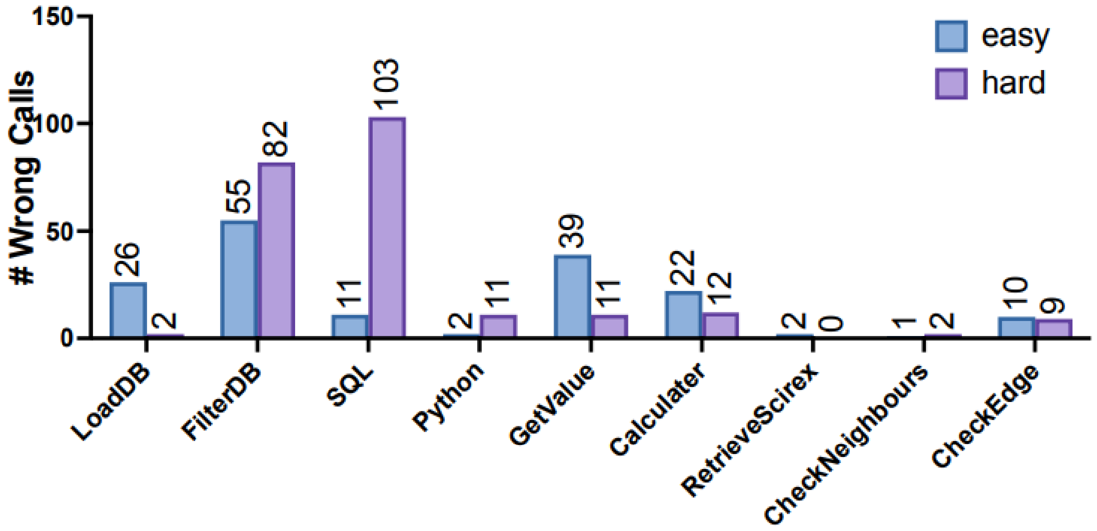


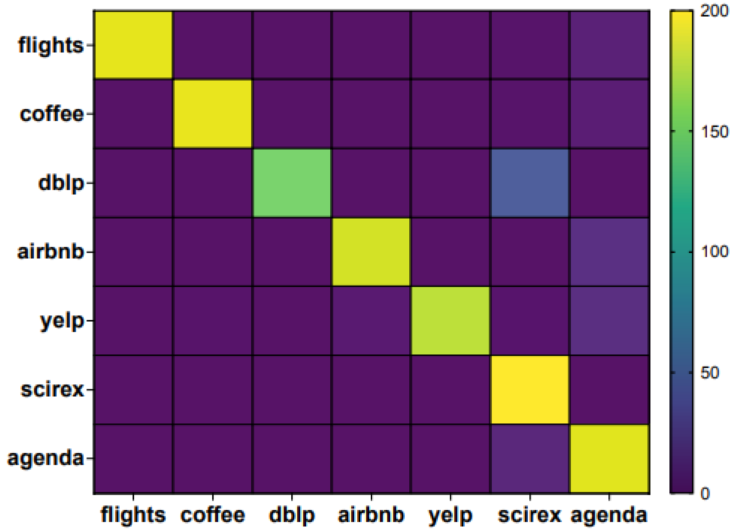
LLM Category	Models	Flight	Coffee	Agenda	Yelp	DBLP	SciREX	GSM8K	Airbnb	Average
Open-Sourced LLMs	LLaMA-2 (13B)	0.0	2.0	0.0	5.0	1.0	0.0	9.0	1.0	2.3
	Falcon (40B)	1.0	1.0	2.0	8.0	1.0	0.0	8.0	5.0	3.3
	LLaMA-2 (70B)	2.0	6.0	5.0	15.0	0.0	0.0	9.0	4.0	5.1
Closed-Sourced LLMs	ChatGPT	2.0	0.0	0.0	15.0	0.0	2.0	26.0	0.0	5.6
	CoT	1.0	1.0	0.0	9.0	0.0	0.0	30.0	0.0	5.1
Tool-Augmented LLMs	Chameleon	30.0	9.0	4.0	8.0	3.0	0.0	27.0	4.0	10.6
	ReAct (GPT-3)	**61.0**	**90.0**	**29.0**	**77.0**	**28.0**	**3.0**	**32.0**	25.0	**43.1**
	ReAct (GPT-3.5)	48.0	81.0	24.0	64.0	23.0	2.0	23.0	**29.0**	36.8

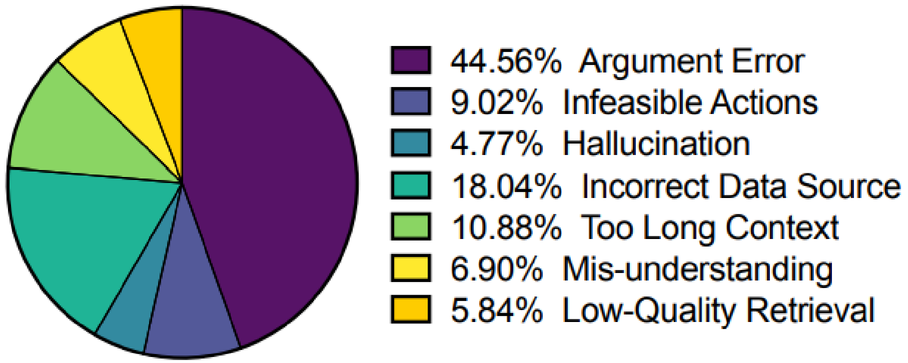


LLM Category	Models	Flight	Coffee	Agenda	Yelp	Airbnb	DBLP	SciREX	Average
Open-Sourced LLMs	LLaMA-2 (13B)	1.0	0.0	0.0	4.0	1.0	5.0	1.0	1.7
	Falcon (40B)	1.0	0.0	0.0	4.0	1.0	6.0	1.0	1.9
	LLaMA-2 (70B)	1.0	0.0	0.0	4.0	1.0	4.0	3.0	1.9
Closed-Sourced LLMs	ChatGPT	2.0	2.3	1.0	0.0	2.0	4.0	3.0	2.0
	CoT	0.0	0.8	0.0	1.0	0.0	3.0	5.0	1.4
Tool-Augmented LLMs	Chameleon	3.0	2.3	0.0	0.0	0.0	8.0	0.0	1.9
	ReAct (GPT-3)	3.0	10.8	0.0	3.0	0.0	**19.0**	0.0	5.1
	ReAct (GPT-3.5)	**5.0**	**17.7**	**7.0**	**8.0**	**7.0**	5.0	**8.0**	**8.2**

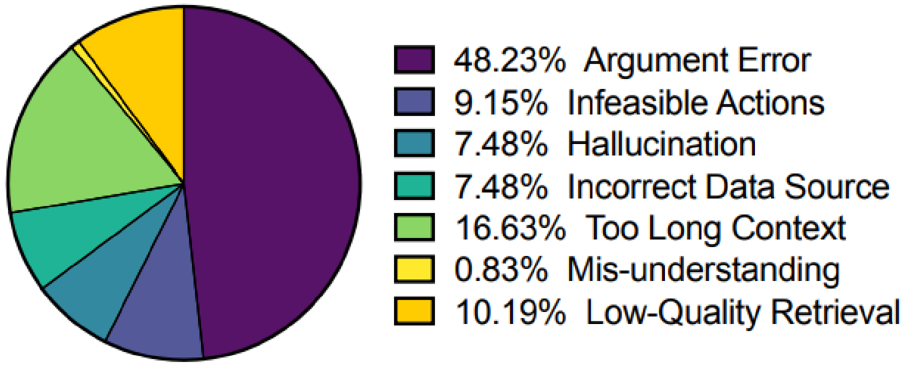
Error Analysis







(a) Easy questions.



(b) Hard questions.

Thanks!
Any Questions?