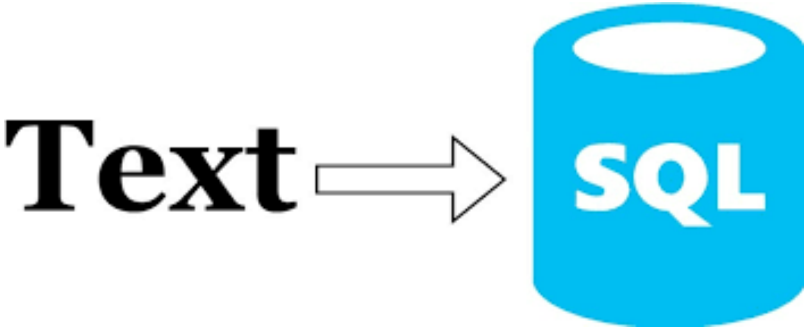
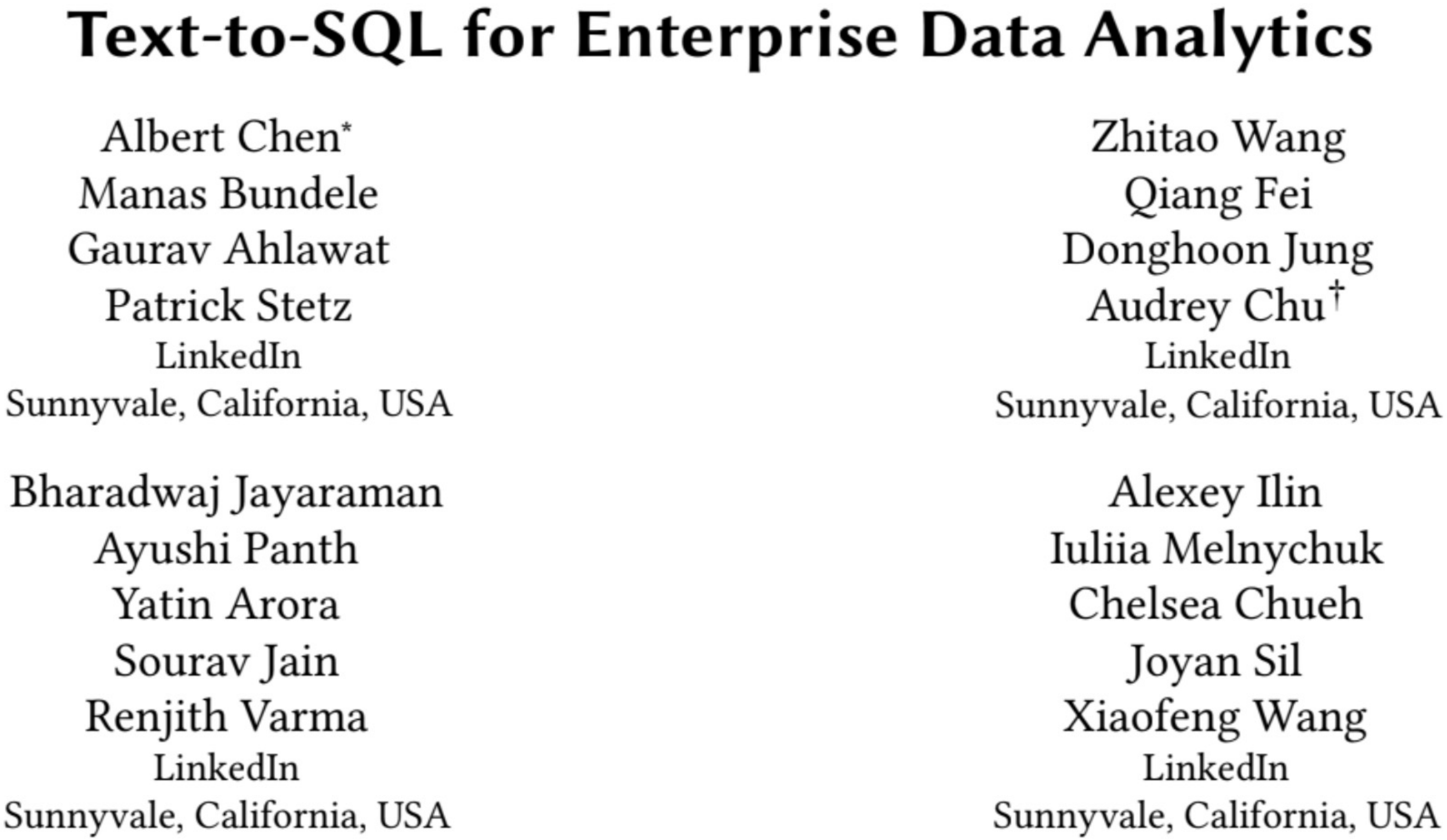


Agentic AI Use Case: Text-to-SQL for Enterprise Data Analytics

Yan Xu

Workshop on Agentic AI for Enterprise at KDD 2025



Text-to-SQL for Enterprise Data Analytics

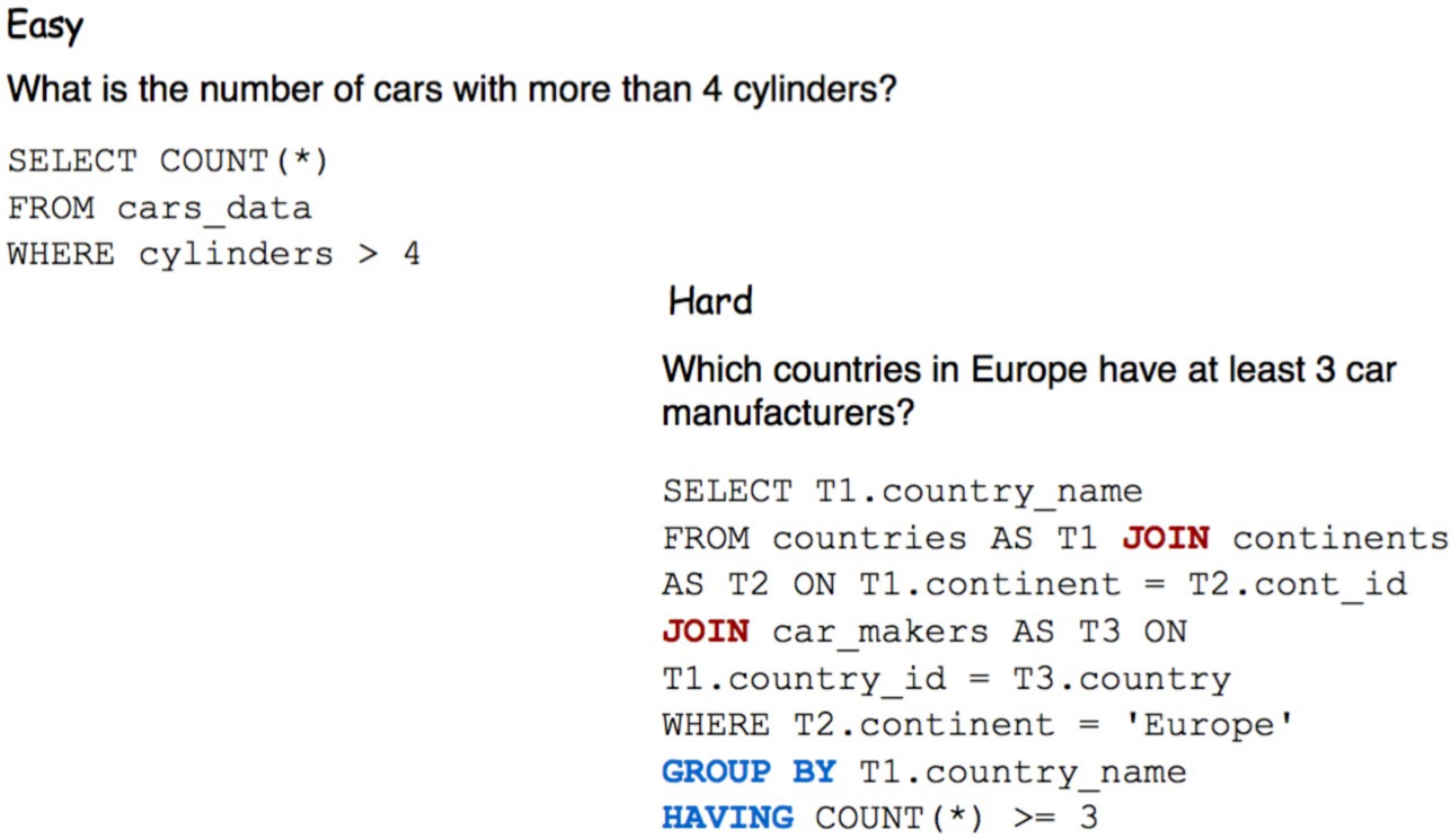
Albert Chen*
Manas Bundele
Gaurav Ahlawat
Patrick Stetz
LinkedIn
Sunnyvale, California, USA

Zhitao Wang
Qiang Fei
Donghoon Jung
Audrey Chu†
LinkedIn
Sunnyvale, California, USA

Bharadwaj Jayaraman
Ayushi Panth
Yatin Arora
Sourav Jain
Renjith Varma
LinkedIn
Sunnyvale, California, USA

Alexey Ilin
Iuliia Melnychuk
Chelsea Chueh
Joyan Sil
Xiaofeng Wang
LinkedIn
Sunnyvale, California, USA

Use Case: Text to SQL



Easy

What is the number of cars with more than 4 cylinders?

```
SELECT COUNT(*)
FROM cars_data
WHERE cylinders > 4
```

Hard

Which countries in Europe have at least 3 car manufacturers?

```
SELECT T1.country_name
FROM countries AS T1 JOIN continents
AS T2 ON T1.continent = T2.cont_id
JOIN car_makers AS T3 ON
T1.country_id = T3.country
WHERE T2.continent = 'Europe'
GROUP BY T1.country_name
HAVING COUNT(*) >= 3
```

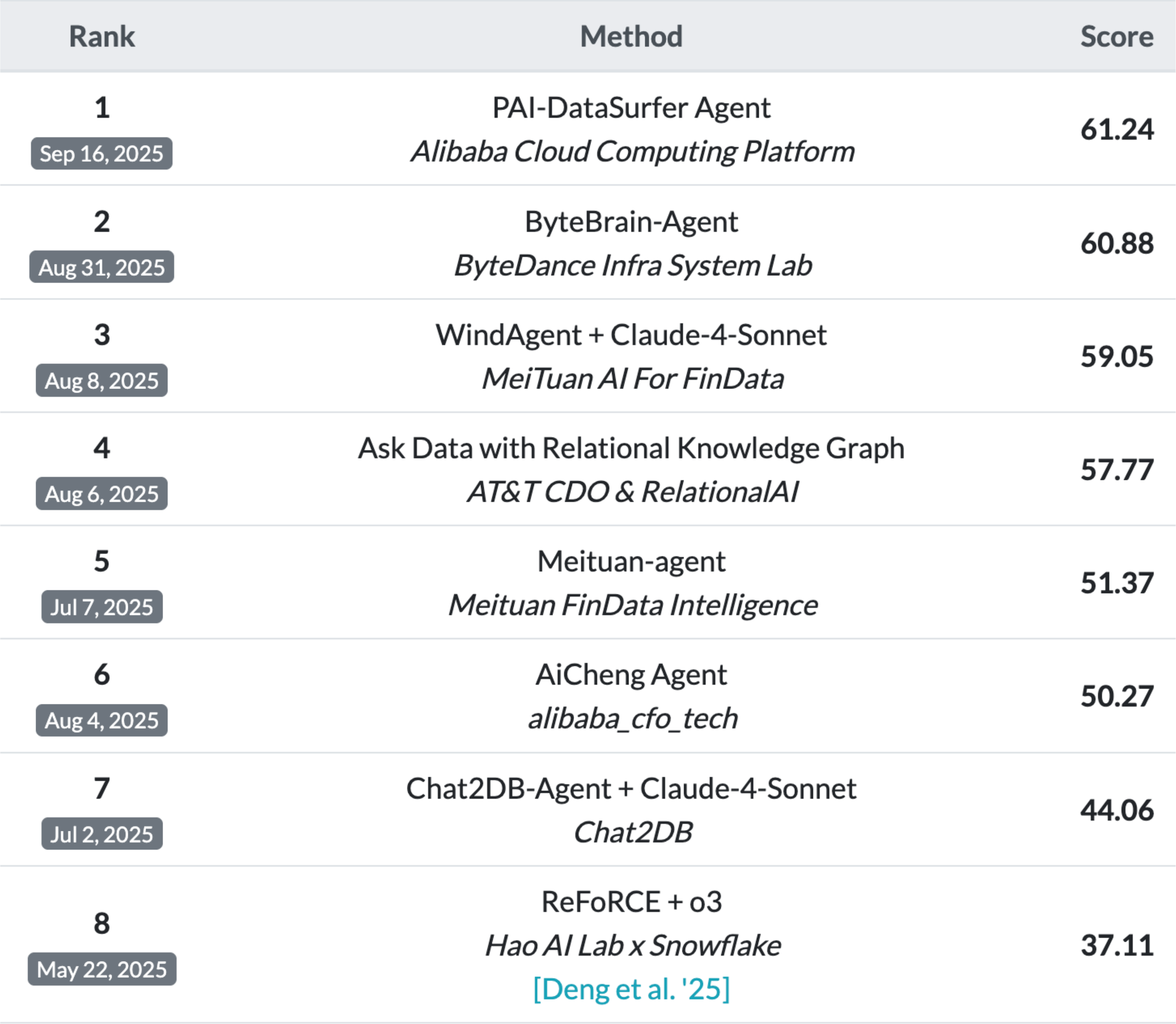
Key requirements

- Understanding domain-specific semantics
- Generalizing to a large and evolving data lake at million scale
- Delivering end-user utility

LLM performance

- **Spider 1.0**: text-to-SQL benchmarks, top models have achieved over 90% execution accuracy.
- **BIRD**: designed for more complex Text-to-SQL tasks, saw rapid improvement, reaching 77% execution accuracy in July 2025.
- **Spider 2.0**: better reflect real-world enterprise scenarios with large schemas and complex queries, gpt-4o performs at ~13%, o1-preview at ~23.77%,
- **Enterprise settings**: testing GPT-4o on an internal evaluation set by Snowflake resulted in an accuracy of 51%, emphasizing the challenges of real-world business use cases.

Spider 2.0 Leaderboard



Rank	Method	Score
1 Sep 16, 2025	PAI-DataSurfer Agent *Alibaba Cloud Computing Platform*	**61.24**
2 Aug 31, 2025	ByteBrain-Agent *ByteDance Infra System Lab*	**60.88**
3 Aug 8, 2025	WindAgent + Claude-4-Sonnet *MeiTuan AI For FinData*	**59.05**
4 Aug 6, 2025	Ask Data with Relational Knowledge Graph *AT&T CDO & RelationalAI*	**57.77**
5 Jul 7, 2025	Meituan-agent *Meituan FinData Intelligence*	**51.37**
6 Aug 4, 2025	AiCheng Agent *alibaba_cfo_tech*	**50.27**
7 Jul 2, 2025	Chat2DB-Agent + Claude-4-Sonnet *Chat2DB*	**44.06**
8 May 22, 2025	ReFoRCE + o3 *Hao AI Lab x Snowflake* [Deng et al. '25]	**37.11**

Text-to-SQL: Key contributions

- To understand data semantics, we construct a **knowledge graph** from table schemas, documentation, code repos, historical query logs, company jargon, and crowdsourced domain knowledge.
- To write queries, the **Query Writer Agent** uses multi-stage retrieval and ranking that identifies the most useful tables, columns, examples, and other context.
- To deliver end-user utility, we design an **interactive chat UI** that helps users understand the query and reply to the bot.

Knowledge Graph

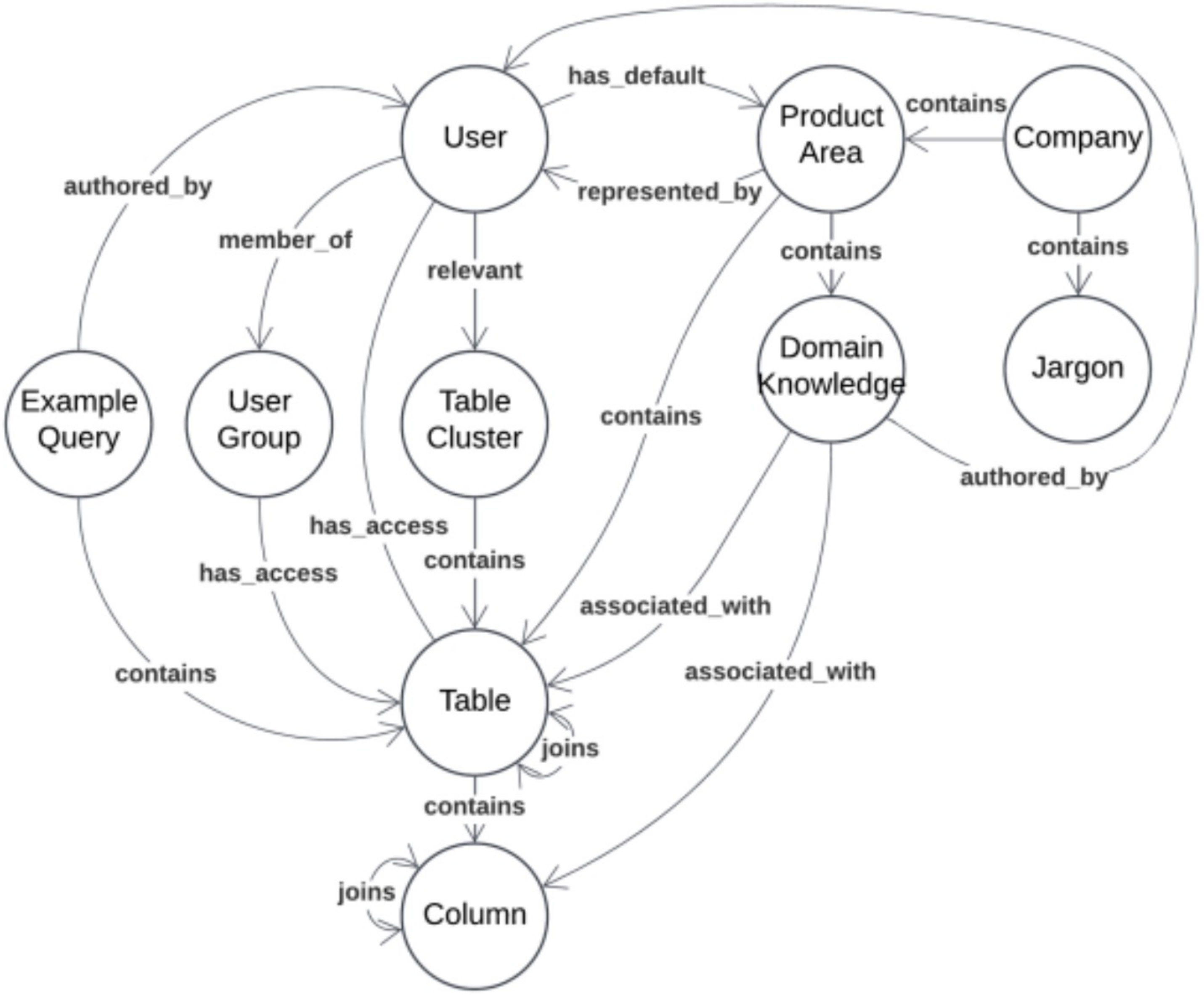
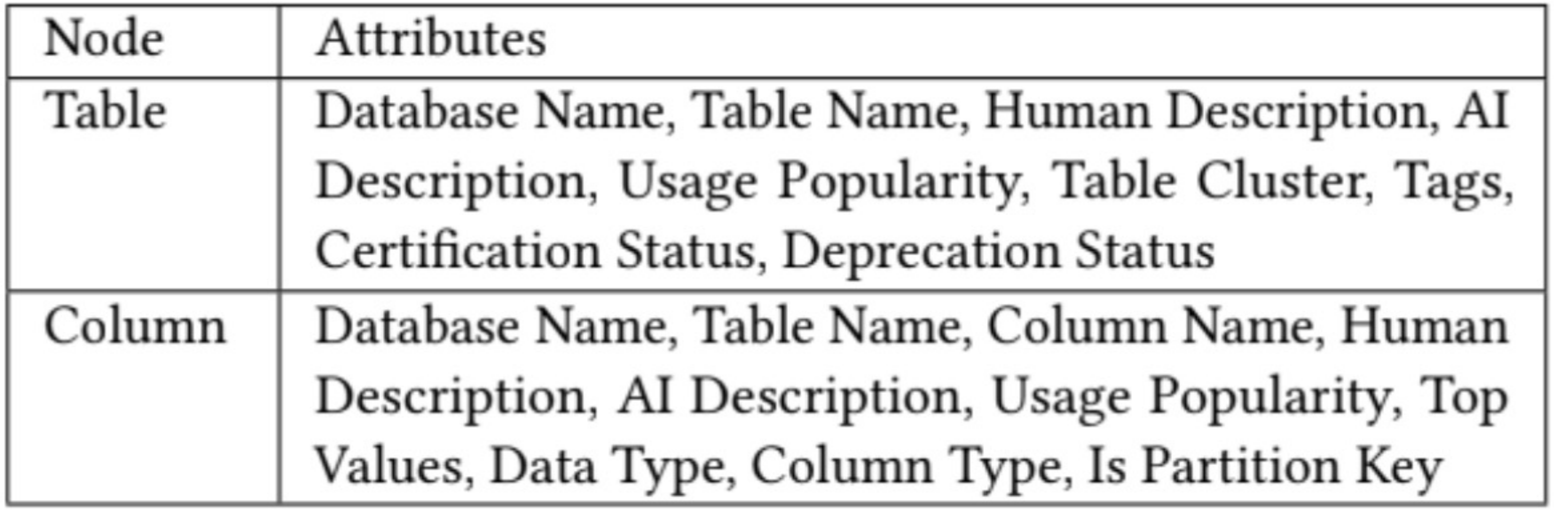


Figure 1: Knowledge graph for Text-to-SQL.

Attributes of table and column nodes



Node	Attributes
Table	Database Name, Table Name, Human Description, AI Description, Usage Popularity, Table Cluster, Tags, Certification Status, Deprecation Status
Column	Database Name, Table Name, Column Name, Human Description, AI Description, Usage Popularity, Top Values, Data Type, Column Type, Is Partition Key

Table 1: Attributes of the table and column nodes.

Table Cluster: User-Dataset Clustering

We have millions of tables, but not all are relevant to every product area or user.

- **Prepare User-Dataset Access Data:** Collect historical query logs for a specified period (e.g., three months) to create a matrix of user-dataset access counts, recording which users accessed which datasets.
- **Filter and Scale the Data:**
  - Reduce noise by filtering the matrix to include only tables (datasets) with a sufficient number of total and unique user accesses, removing rarely accessed or intermediate tables
  - Scale the user-dataset matrix so that the data for each user has a mean of 0 and a standard deviation of 1 across datasets.
- **Perform Dimensionality Reduction (ICA):** Apply Independent Component Analysis (ICA) to the scaled user-dataset matrix -> a score for each (table, component)
- **Create Dataset Clusters:**
  - For each component (representing a cluster), identify the top datasets with the strongest scores
  - Additionally, assign each dataset to the component (cluster) with its highest score. This ensures every dataset belongs to at least one cluster

Assign users to clusters

- Aggregate user-dataset access counts by cluster to determine user access counts for each cluster.
- For each user, identify and assign the top clusters based on their access counts, effectively creating personalized clusters for each user.

Question routing: User intention classifier

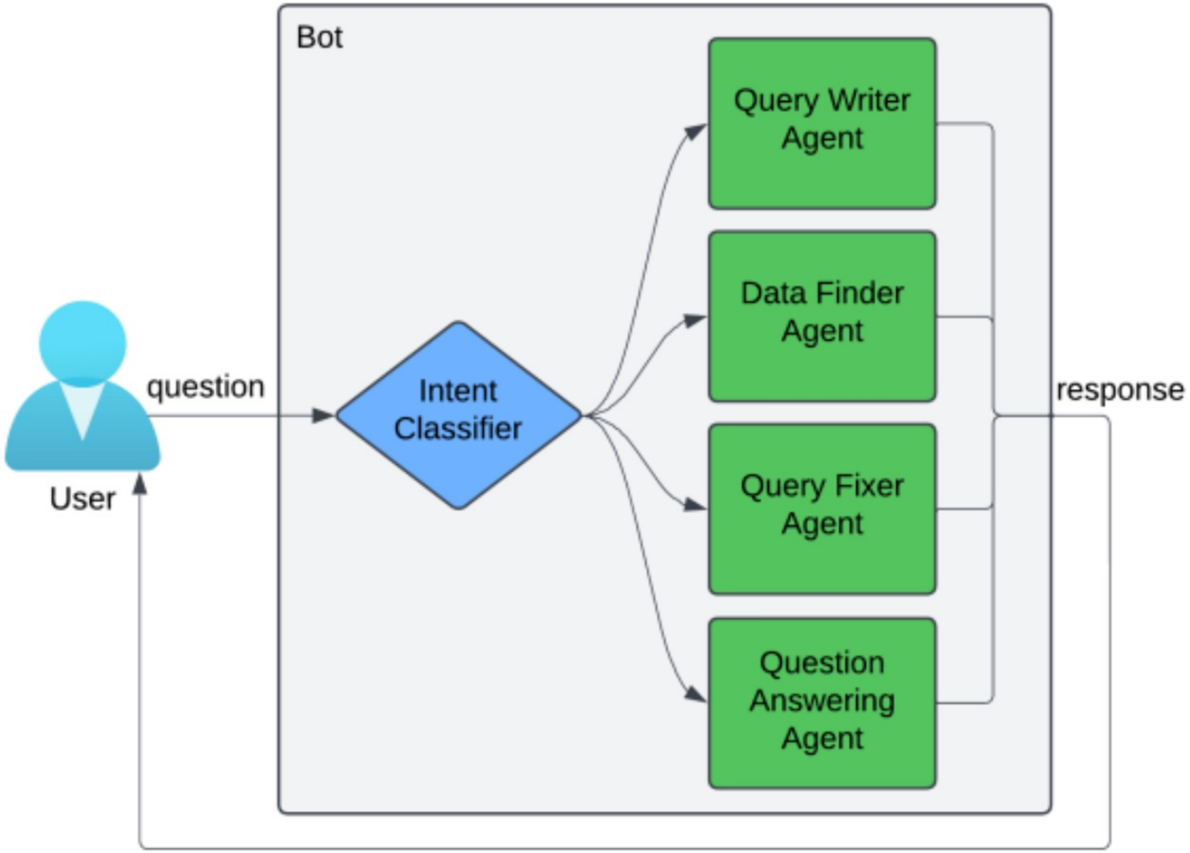
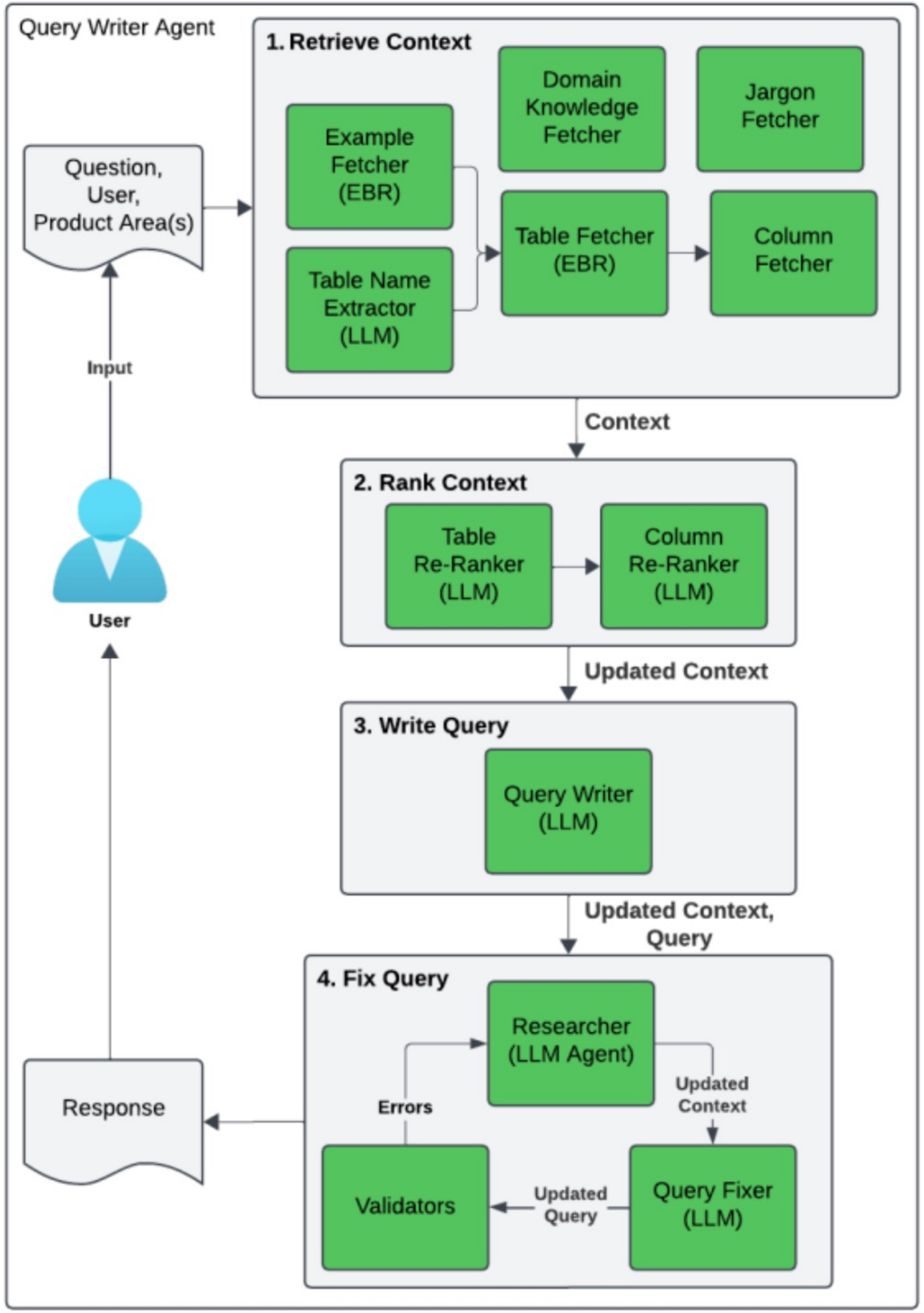
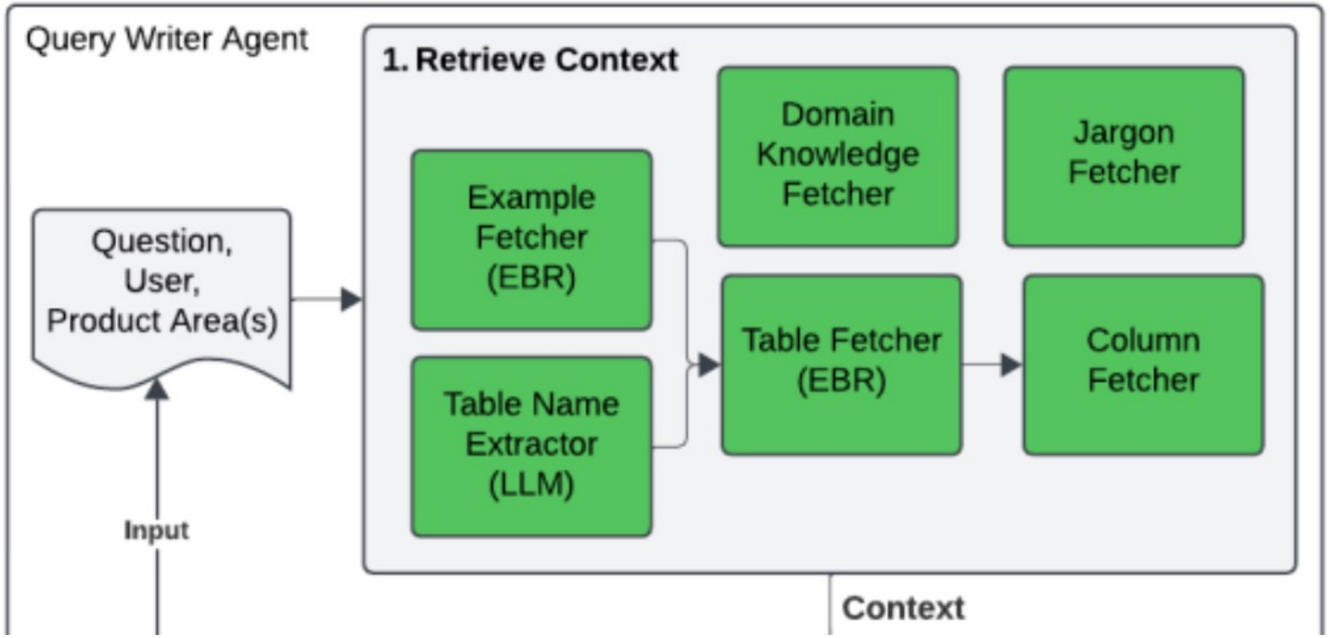


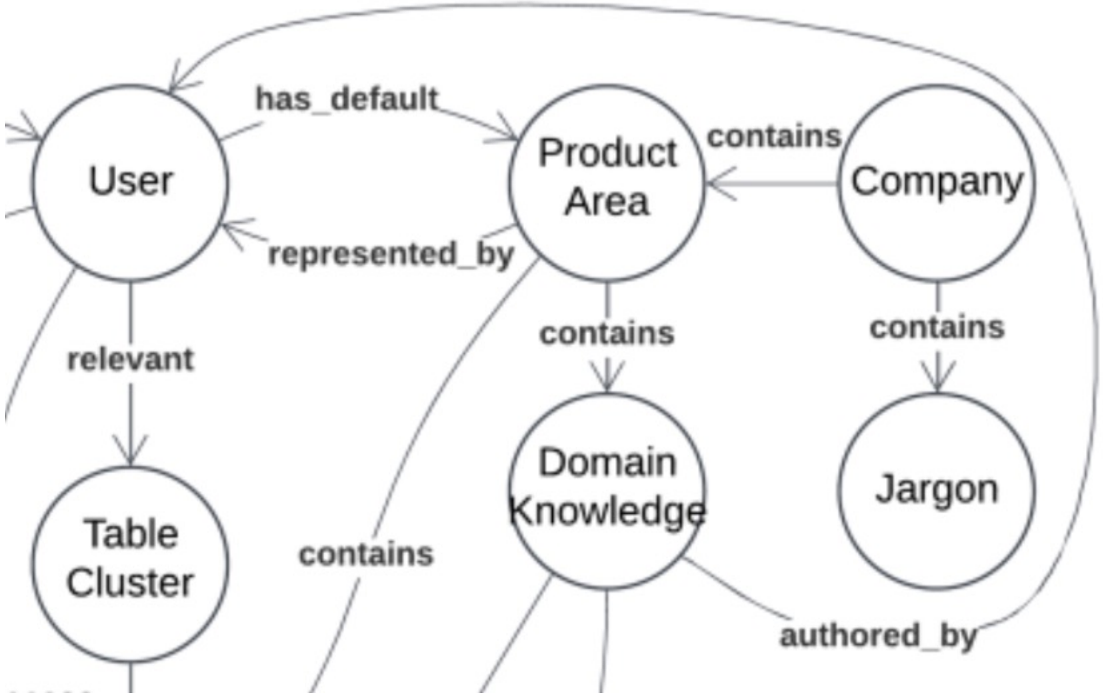
Figure 3: Multi-agent architecture supports various intents by routing questions to the appropriate agent.

Query Writer Agent architecture



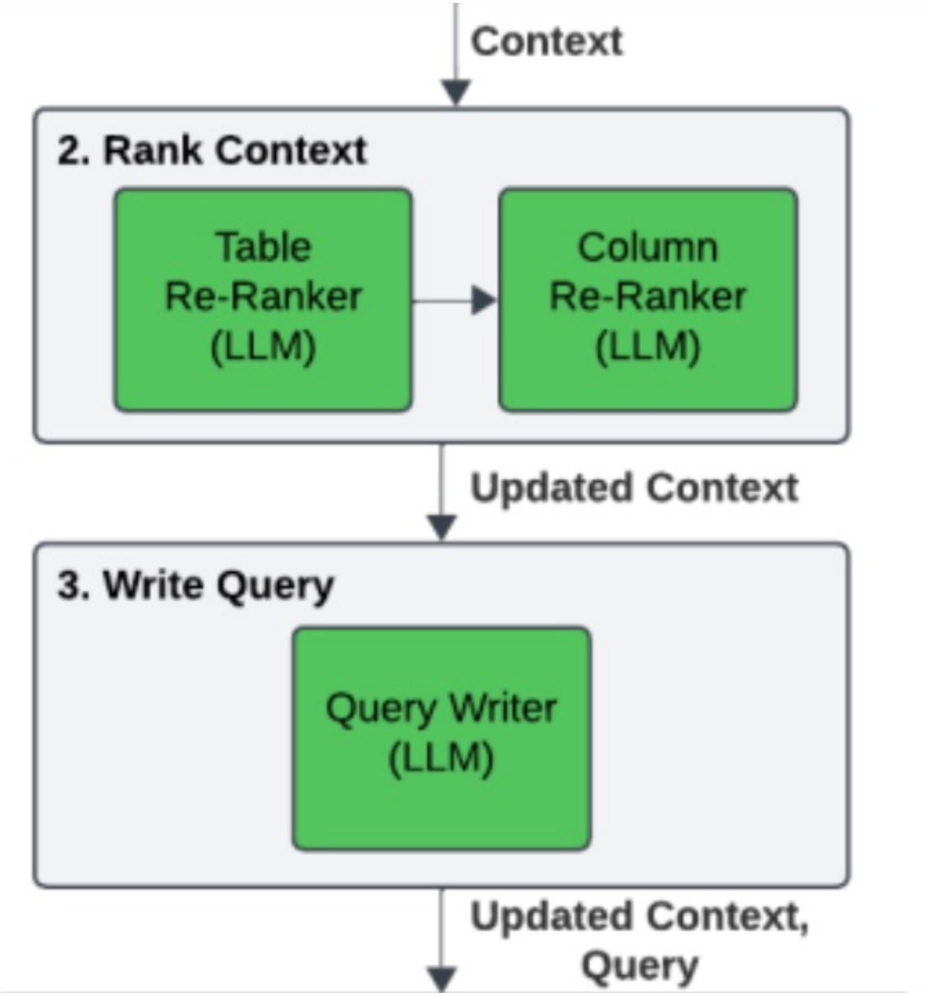
Retrieve Context





EBR: embedding-based retrieval

Rank Context and Write Query



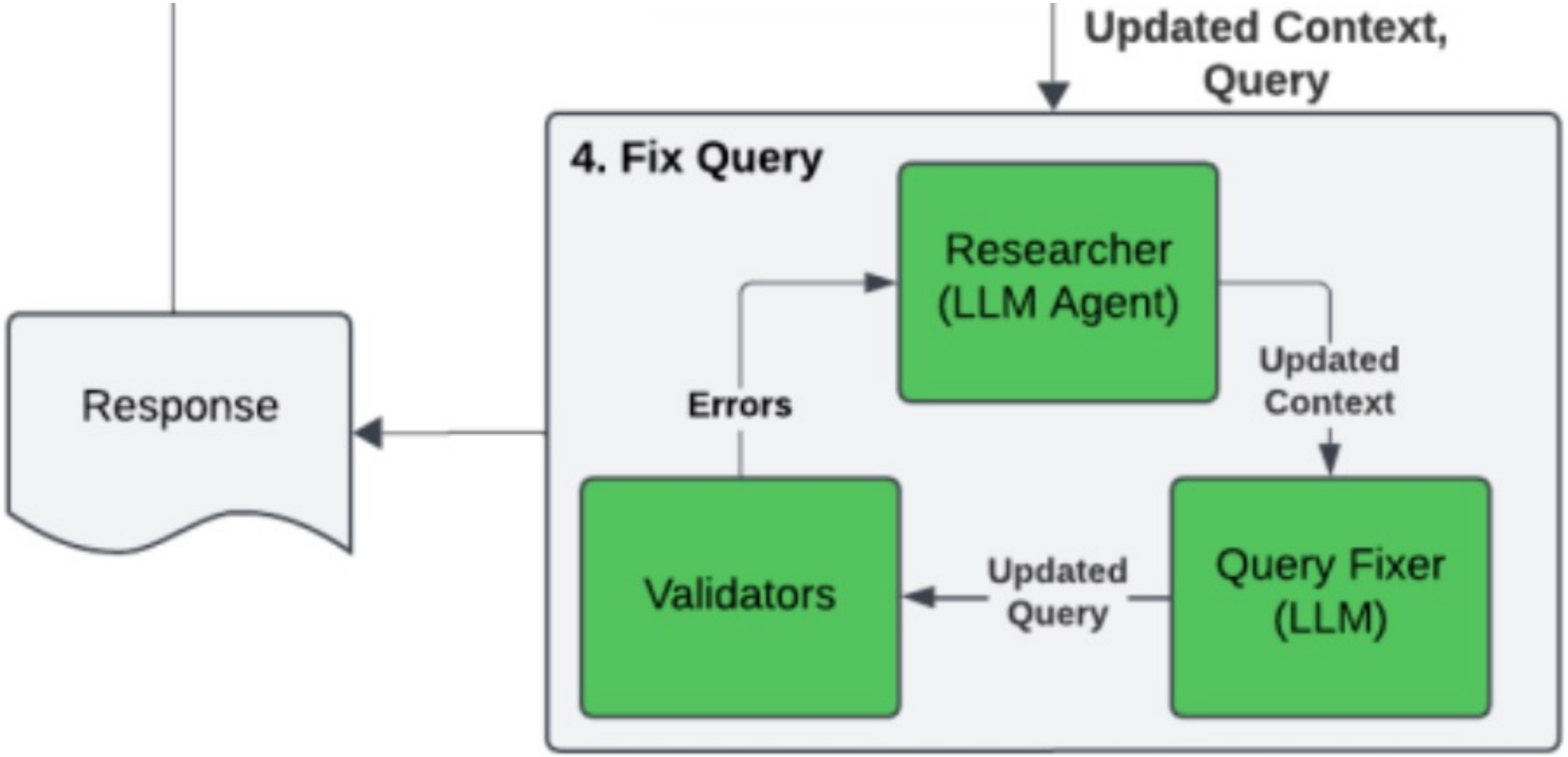
To narrow down the initial set of candidate tables (identified during context retrieval) to a smaller, more focused set that is most pertinent to the user's natural language question.

Re-Ranker LLM Prompt context:

- Table names and descriptions
- Example queries associated with tables
- Information about commonly joined or co-queried tables
- Relevant domain knowledge records
- Explanations of internal jargon detected in the user's query

Output: Identify the top K tables to be passed to the subsequent query writing

Fix Query


Updated Context,
Query
4. Fix Query
Researcher
(LLM Agent)
Updated
Context
Errors
Response
Validators
Updated
Query
Query Fixer
(LLM)

Researcher agent

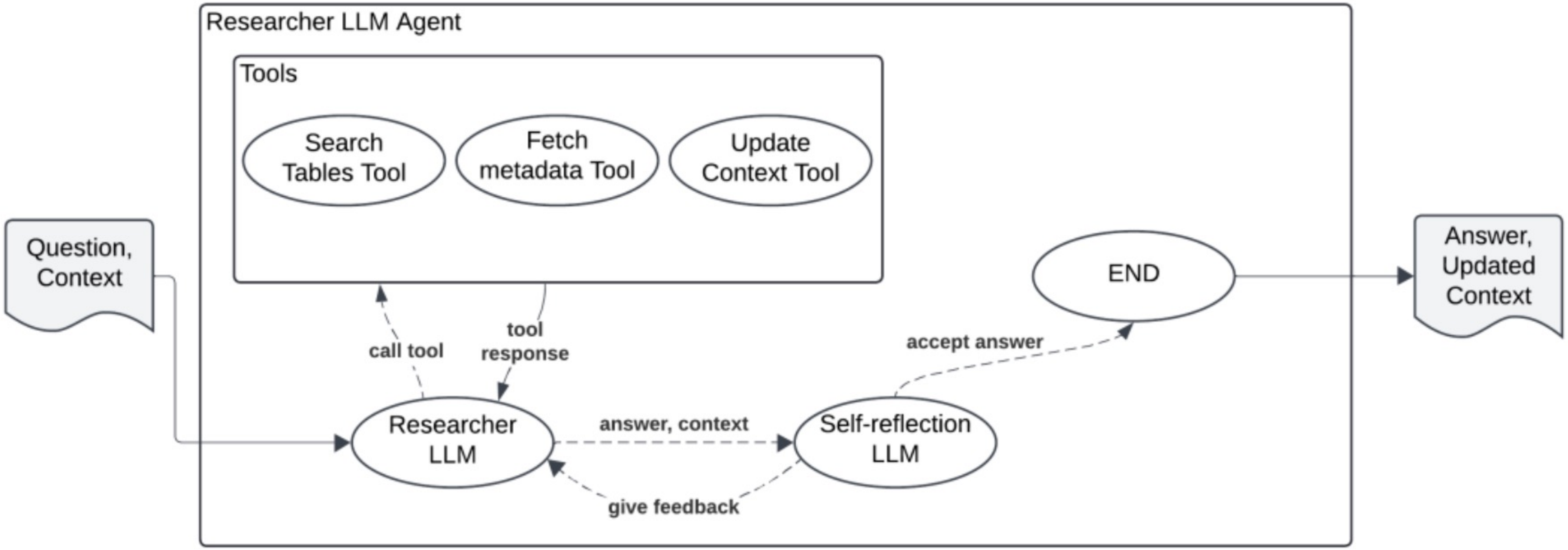
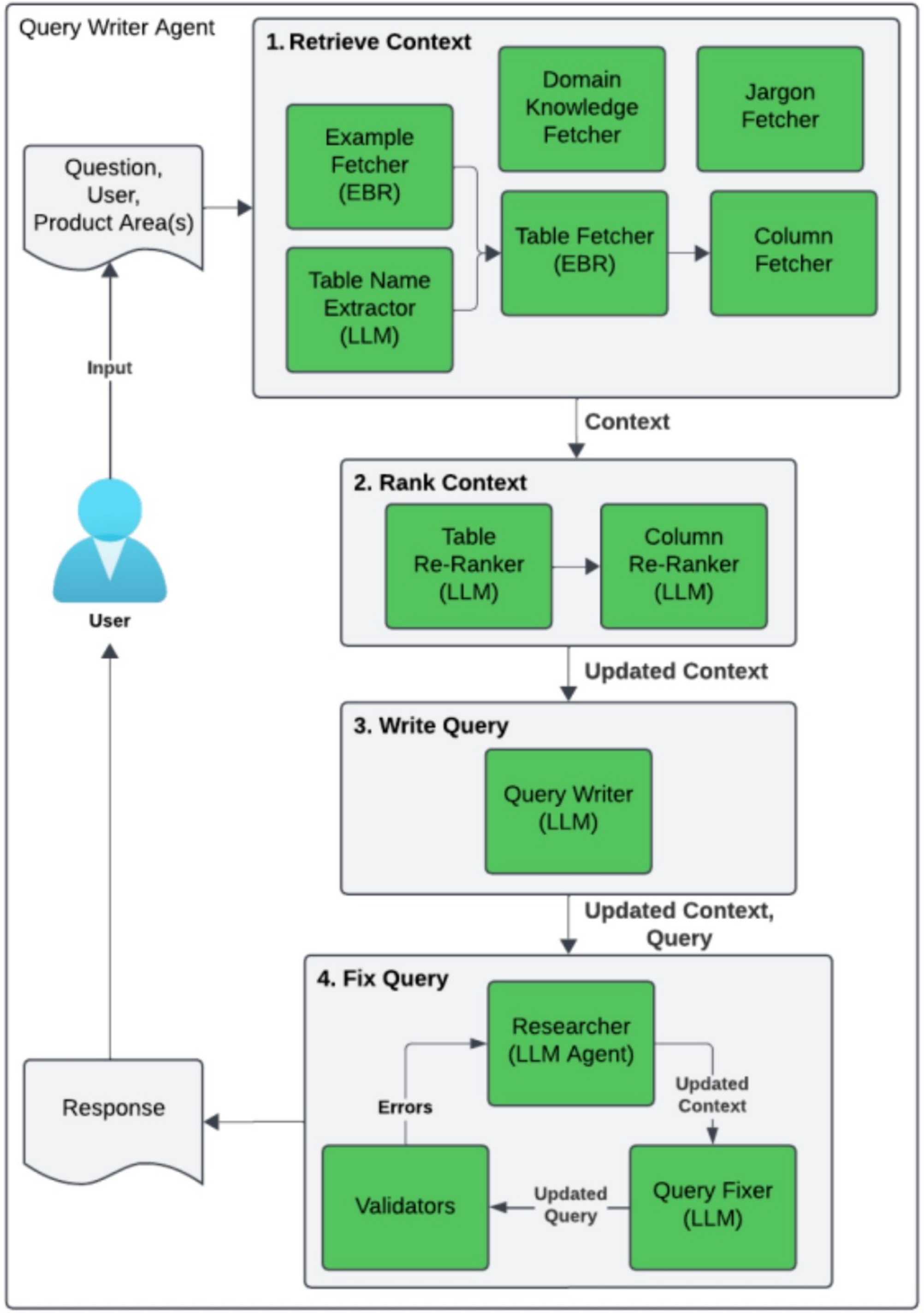
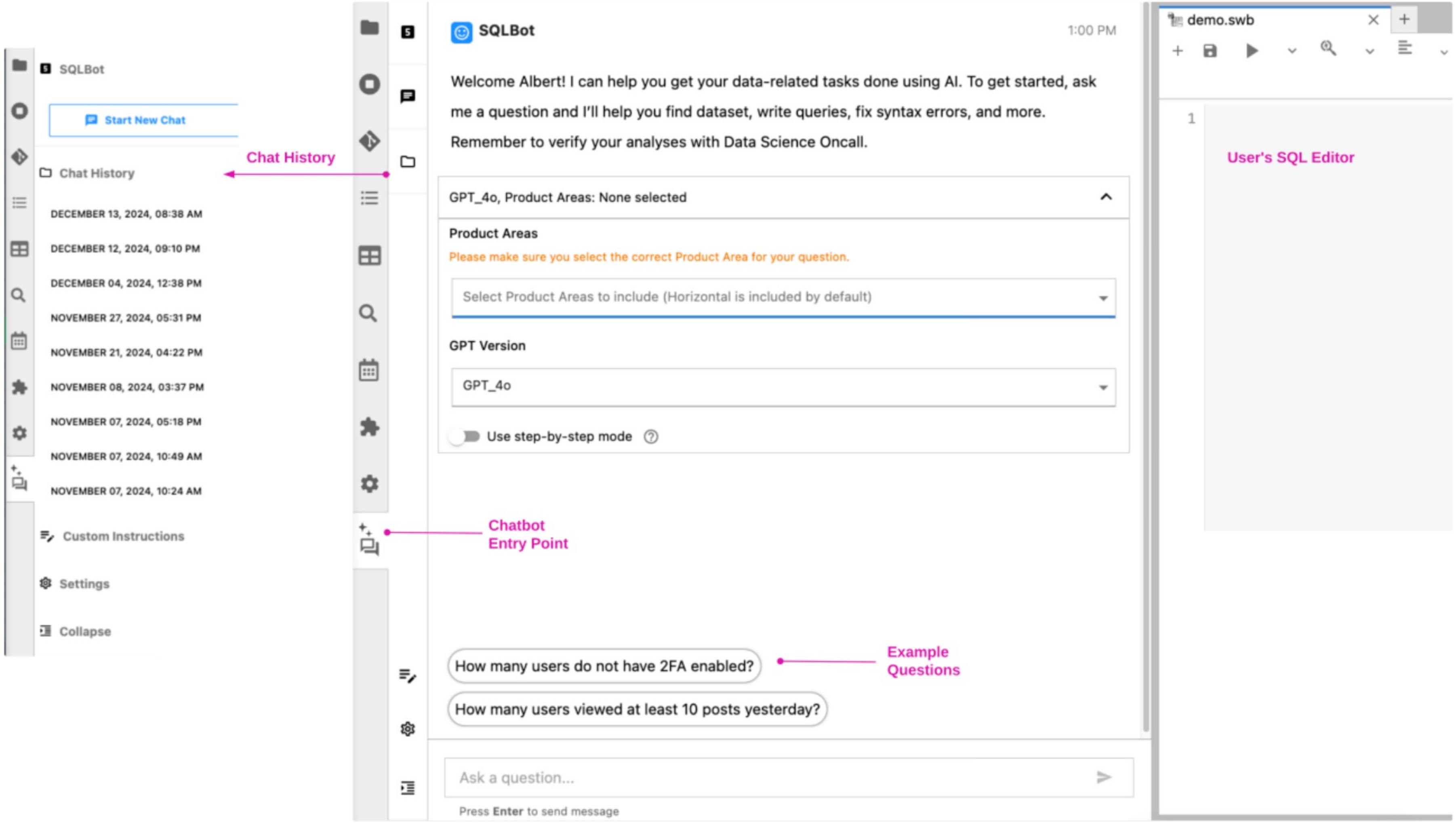


Figure 8: Researcher LLM Agent architecture. The Researcher is used within query fixing to search for tables to resolve hallucination.

Recap: Query Writer Agent architecture



Use Case: Text to SQL



DECEMBER 13, 2024, 08:38 AM
DECEMBER 12, 2024, 09:10 PM
DECEMBER 04, 2024, 12:38 PM
NOVEMBER 27, 2024, 05:31 PM
NOVEMBER 21, 2024, 04:22 PM
NOVEMBER 08, 2024, 03:37 PM
NOVEMBER 07, 2024, 05:18 PM
NOVEMBER 07, 2024, 10:49 AM
NOVEMBER 07, 2024, 10:24 AM

SQLBot 1:00 PM

Welcome Albert! I can help you get your data-related tasks done using AI. To get started, ask me a question and I'll help you find dataset, write queries, fix syntax errors, and more. Remember to verify your analyses with Data Science Oncall.

GPT_4o, Product Areas: None selected

Product Areas

Please make sure you select the correct Product Area for your question.

Select Product Areas to include (Horizontal is included by default)

GPT Version

GPT_4o

Use step-by-step mode

How many users do not have 2FA enabled?

How many users viewed at least 10 posts yesterday?

Chat History

Chatbot Entry Point

Example Questions

User's SQL Editor

Figure 5: Chatbot is available in sidebar next to the SQL editor.

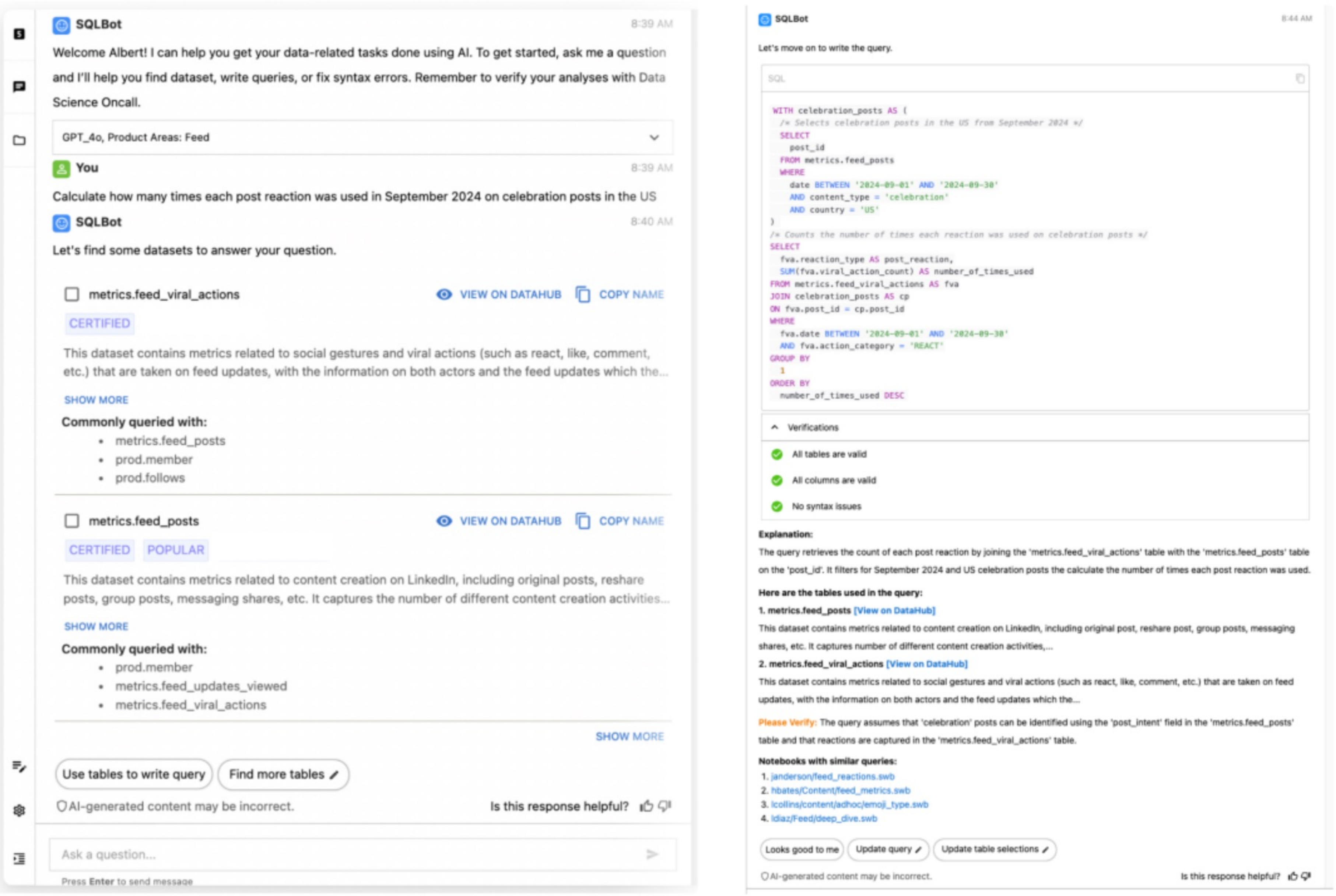


Table Output

SQLBot 8:39 AM

Welcome Albert! I can help you get your data-related tasks done using AI. To get started, ask me a question and I'll help you find dataset, write queries, or fix syntax errors. Remember to verify your analyses with Data Science Oncall.

GPT_4o, Product Areas: Feed

You 8:39 AM

Calculate how many times each post reaction was used in September 2024 on celebration posts in the US

SQLBot 8:40 AM

Let's find some datasets to answer your question.

☐ metrics.feed_viral_actions

CERTIFIED

This dataset contains metrics related to social gestures and viral actions (such as react, like, comment, etc.) that are taken on feed updates, with the information on both actors and the feed updates which the...

Commonly queried with:
- metrics.feed_posts
- prod.member
- prod.follows

☐ metrics.feed_posts

CERTIFIED POPULAR

This dataset contains metrics related to content creation on LinkedIn, including original posts, reshare posts, group posts, messaging shares, etc. It captures the number of different content creation activities...

Commonly queried with:
- prod.member
- metrics.feed_updates_viewed
- metrics.feed_viral_actions

AI-generated content may be incorrect.

Query Output

SQLBot 8:44 AM

Let's move on to write the query.

SQL

```sql
WITH celebration_posts AS (
  /* Selects celebration posts in the US from September 2024 */
  SELECT
    post_id
  FROM metrics.feed_posts
  WHERE
    date BETWEEN '2024-09-01' AND '2024-09-30'
    AND content_type = 'celebration'
    AND country = 'US'
)
/* Counts the number of times each reaction was used on celebration posts */
SELECT
  fva.reaction_type AS post_reaction,
  SUM(fva.viral_action_count) AS number_of_times_used
FROM metrics.feed_viral_actions AS fva
JOIN celebration_posts AS cp
ON fva.post_id = cp.post_id
WHERE
  fva.date BETWEEN '2024-09-01' AND '2024-09-30'
  AND fva.action_category = 'REACT'
GROUP BY
  1
ORDER BY
  number_of_times_used DESC
```

Verifications
- ✓ All tables are valid
- ✓ All columns are valid
- ✓ No syntax issues

Explanation:

The query retrieves the count of each post reaction by joining the 'metrics.feed_viral_actions' table with the 'metrics.feed_posts' table on the 'post_id'. It filters for September 2024 and US celebration posts the calculate the number of times each post reaction was used.

Here are the tables used in the query:

1. metrics.feed_posts [View on DataHub]

This dataset contains metrics related to content creation on LinkedIn, including original post, reshare post, group posts, messaging shares, etc. It captures number of different content creation activities,...

2. metrics.feed_viral_actions [View on DataHub]

This dataset contains metrics related to social gestures and viral actions (such as react, like, comment, etc.) that are taken on feed updates, with the information on both actors and the feed updates which the...

Please Verify: The query assumes that 'celebration' posts can be identified using the 'post_intent' field in the 'metrics.feed_posts' table and that reactions are captured in the 'metrics.feed_viral_actions' table.

Notebooks with similar queries:
1. janderson/feed_reactions.swb
2. hbates/Content/feed_metrics.swb
3. lcollins/content/adhoc/emoji_type.swb
4. ldiaz/Feed/deep_dive.swb

AI-generated content may be incorrect.

Figure 6: Rich display elements in query output help users understand responses and ask follow-up questions.

Evaluation

- We defined an **internal benchmark set** to measure performance for our enterprise. We asked product area domain experts in our company to provide questions and ground truth queries.
- Since questions may have multiple valid solutions (e.g., using equivalent tables or columns), we accommodate multiple ground truth responses per question.
- In total, we collected **133 questions across 10 product areas and 167 ground truth tables**. 60% of the questions have more than one valid response.

Evaluation

Since our end user experience is **an interactive chatbot**, we focus on metrics that **assess utility**:

- Did it find the right tables columns
- Is the query correct or close to correct
- Did it respond in a timely manner

Evaluation

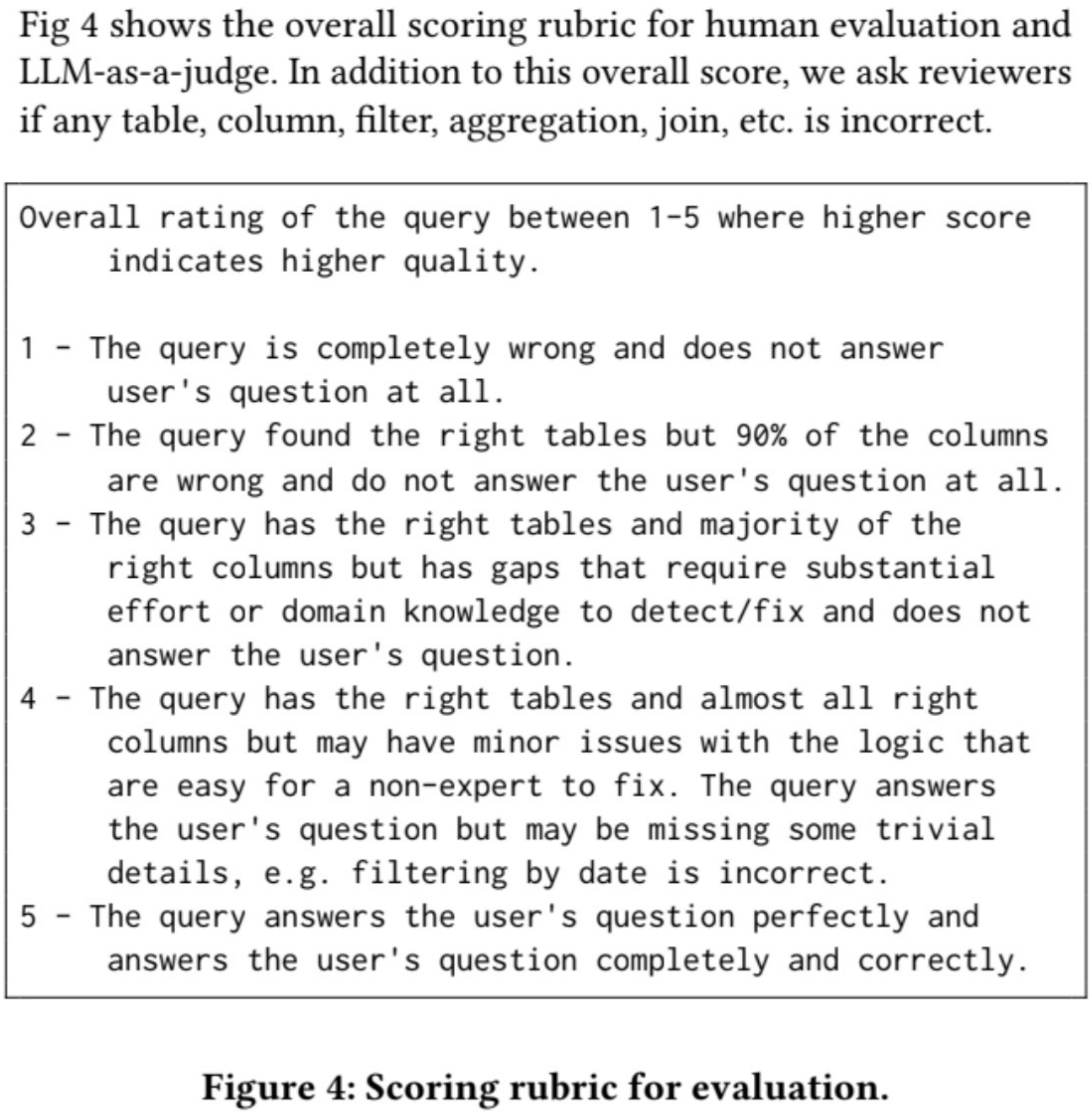


Fig 4 shows the overall scoring rubric for human evaluation and LLM-as-a-judge. In addition to this overall score, we ask reviewers if any table, column, filter, aggregation, join, etc. is incorrect.

```
Overall rating of the query between 1-5 where higher score
     indicates higher quality.

1 - The query is completely wrong and does not answer
     user's question at all.
2 - The query found the right tables but 90% of the columns
     are wrong and do not answer the user's question at all.
3 - The query has the right tables and majority of the
     right columns but has gaps that require substantial
     effort or domain knowledge to detect/fix and does not
     answer the user's question.
4 - The query has the right tables and almost all right
     columns but may have minor issues with the logic that
     are easy for a non-expert to fix. The query answers
     the user's question but may be missing some trivial
     details, e.g. filtering by date is incorrect.
5 - The query answers the user's question perfectly and
     answers the user's question completely and correctly.
```

Figure 4: Scoring rubric for evaluation.

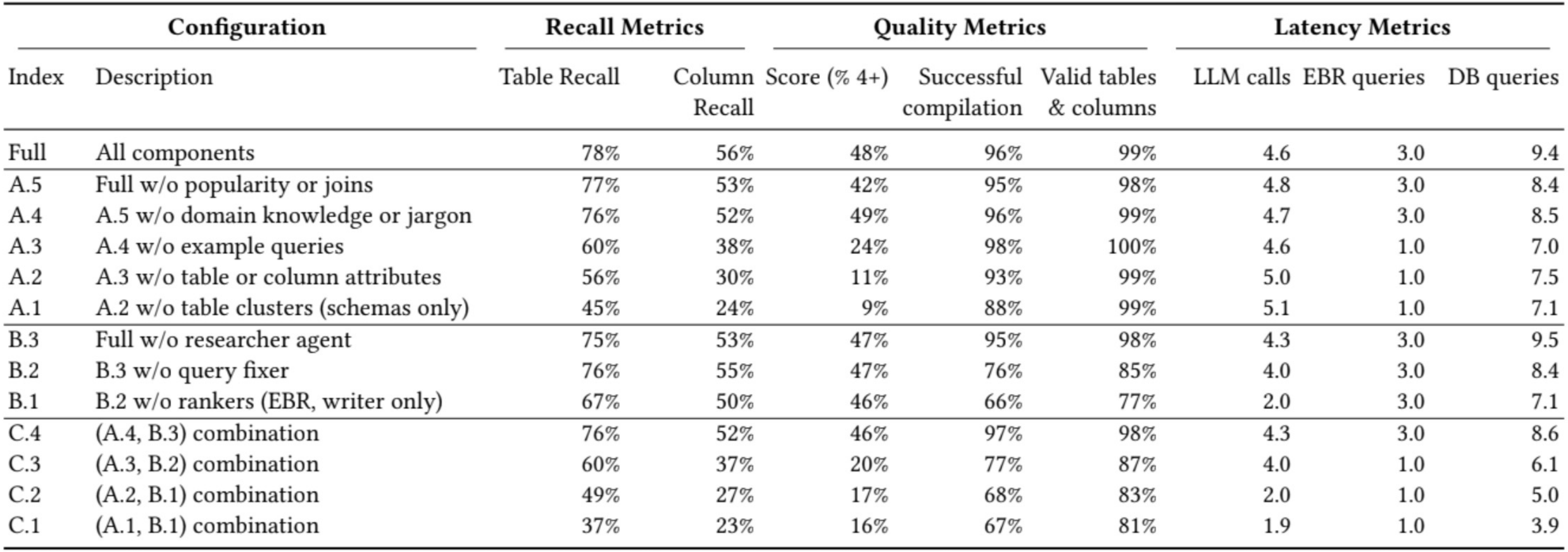
Experiment setup

- For benchmarking, we used **E5-large-v2** for example embeddings and **text-embedding-ada-002** for table/column embeddings.
- We used **gpt-4o-mini-2024-07-18** for the Researcher LLM and **gpt-4o-2024-05-13** for the other LLMs, with temperature 0.
- We did a single run for each configuration. LLM-as-a-judge scoring is done with **gpt-4o-2024-05-13**. All models were accessed through the Azure OpenAI Service.

Results

Scores: use both human evaluators and LLM-as-a-judge to rate queries.

our full model configuration runs in under 60 seconds per question on average.



	Configuration		Recall Metrics		Quality Metrics			Latency Metrics		
	Index	Description	Table Recall	Column Recall	Score (% 4+)	Successful compilation	Valid tables & columns	LLM calls	EBR queries	DB queries
	Full	All components	78%	56%	48%	96%	99%	4.6	3.0	9.4
Knowledge graph components	A.5	Full w/o popularity or joins	77%	53%	42%	95%	98%	4.8	3.0	8.4
	A.4	A.5 w/o domain knowledge or jargon	76%	52%	49%	96%	99%	4.7	3.0	8.5
	A.3	A.4 w/o example queries	60%	38%	24%	98%	100%	4.6	1.0	7.0
	A.2	A.3 w/o table or column attributes	56%	30%	11%	93%	99%	5.0	1.0	7.5
	A.1	A.2 w/o table clusters (schemas only)	45%	24%	9%	88%	99%	5.1	1.0	7.1
Models	B.3	Full w/o researcher agent	75%	53%	47%	95%	98%	4.3	3.0	9.5
	B.2	B.3 w/o query fixer	76%	55%	47%	76%	85%	4.0	3.0	8.4
	B.1	B.2 w/o rankers (EBR, writer only)	67%	50%	46%	66%	77%	2.0	3.0	7.1
Both	C.4	(A.4, B.3) combination	76%	52%	46%	97%	98%	4.3	3.0	8.6
	C.3	(A.3, B.2) combination	60%	37%	20%	77%	87%	4.0	1.0	6.1
	C.2	(A.2, B.1) combination	49%	27%	17%	68%	83%	2.0	1.0	5.0
	C.1	(A.1, B.1) combination	37%	23%	16%	67%	81%	1.9	1.0	3.9

Table 2: Ablation study. A.1-A.5 show the effect of ablating knowledge graph components. B.1-B.3 show the effect of ablating modeling components. C.1-C.4 show the effect of ablating both.

Deployment

- Our chatbot has been available in the querying platform since July 2024 with steady usage of over 300 weekly active users.
- Our most popular intents are query fixing, query writing, and data discovery.
- Power users have over 100 chat sessions / month.
- About 20% of weekly active users return the following week.
- 33% of chat sessions lead to code pasted from the chatbot into the SQL editor.
- We conducted a survey to understand users' satisfaction with the chatbot. 39% of users rated its queries as "Very good (require minor modifications)" or "Excellent (queries are correct)." 95% of users rated them at least "Passes (queries require some modifications)."

Conclusion: Paving the Way for Enterprise Text-to-SQL

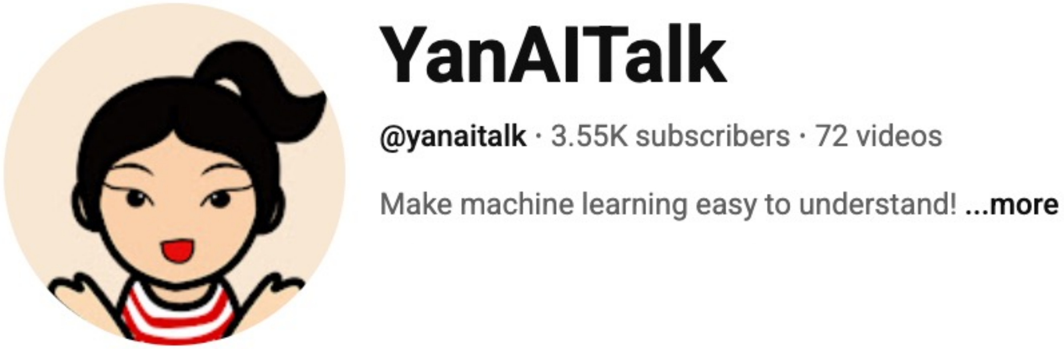
- **Semantic Understanding is Paramount**: The knowledge graph, incorporating diverse enterprise data like documentation, query logs, and user-generated domain knowledge, was crucial for capturing nuanced data semantics and improving query quality from 9% to 49% correct/near-correct responses (vs. schemas only).
- **Robust Agentic Architecture is Essential**: A multi-stage Query Writer Agent (retrieval, ranking, writing, and fixing with a Researcher LLM Agent) significantly boosted compilation success and eliminated schema hallucination, ensuring reliable query generation.
- **User-Centric Design Drives Adoption:** An interactive, multi-agent chatbot UI tailored for diverse user intents (query writing, fixing, data discovery) and integrated into workflows, has shown strong user engagement and high satisfaction ratings.

Upcoming meetups – Open to proposals and guest speakers!

- Agentic AI real-world use cases
- Latest trends in AI research
- Build AI agents: Hands-on sessions
- Panel discussions

Slides posted at:
https://github.com/YanXuHappygela/LLM-reading-group

Recordings posted at:



YanAITalk

@yanaitalk · 3.55K subscribers · 72 videos

Make machine learning easy to understand! **...more**