



SCHOOL OF
**COMPUTING &
DATA SCIENCE**
The University of Hong Kong



S.T.A.R. lab.

ICDE 2026 Beyond SQL Workshop

BIRD-SQL: Towards Automatic Data-Centric Code Generation

Prof. Reynold Cheng
Head of AI and Data Science Division
School of Computing and Data Science
University of Hong Kong

<https://reynold.hku.hk>

ckcheng@cs.hku.hk

- Social Technology And Research Laboratory



SCHOOL OF
COMPUTING &
DATA SCIENCE
The University of Hong Kong



香港賽馬會慈善信託基金
The Hong Kong Jockey Club Charities Trust



HKU Med LKS Faculty of Medicine
School of Public Health
香港大學公共衛生學院

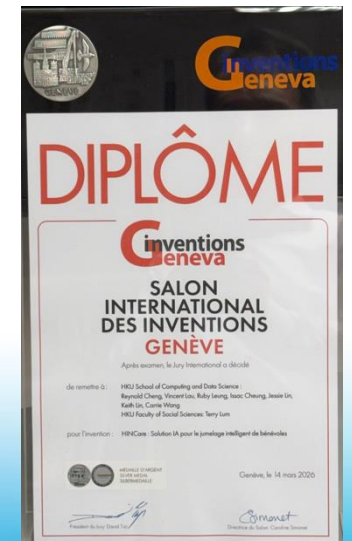


HKU SWSA Department of Social Work and Social Administration
The University of Hong Kong
香港大學社會工作及社會行政學系



HKU Musketeers Foundation
Institute of Data Science
香港大學同心基金數據科學研究院

- 7+ years in website, app, game, and video production
- 8 local and international awards





Intelligent Applications



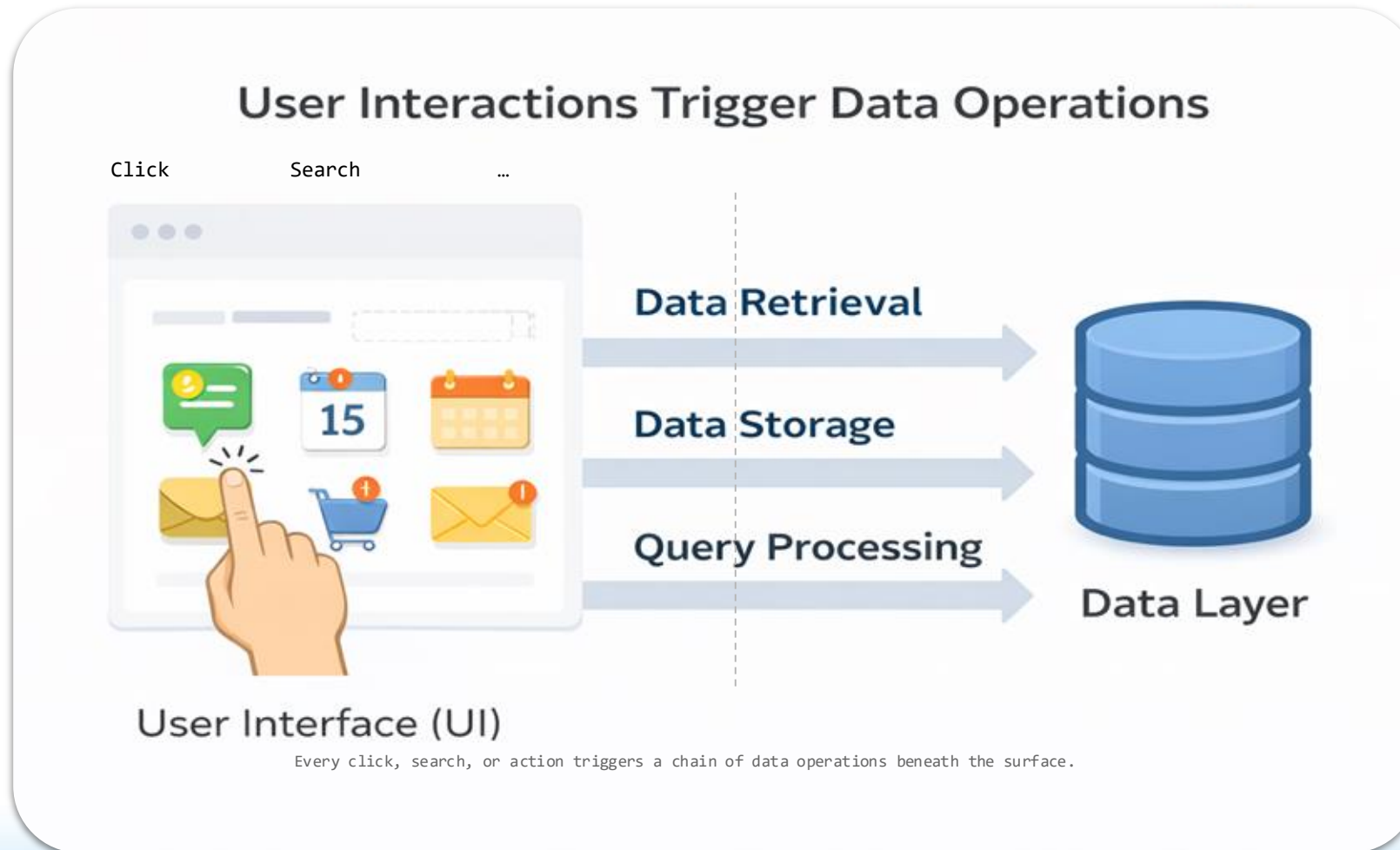
 **Social Media & Content Platform**
Instagram, TikTok, X (Twitter), ...

 **Productivity & Technology**
Google, Microsoft, OpenAI, ...

 **E-commerce & Finance**
Amazon, PayPal, ...

 **Entertainment & Lifestyle**
YouTube, Spotify, ...

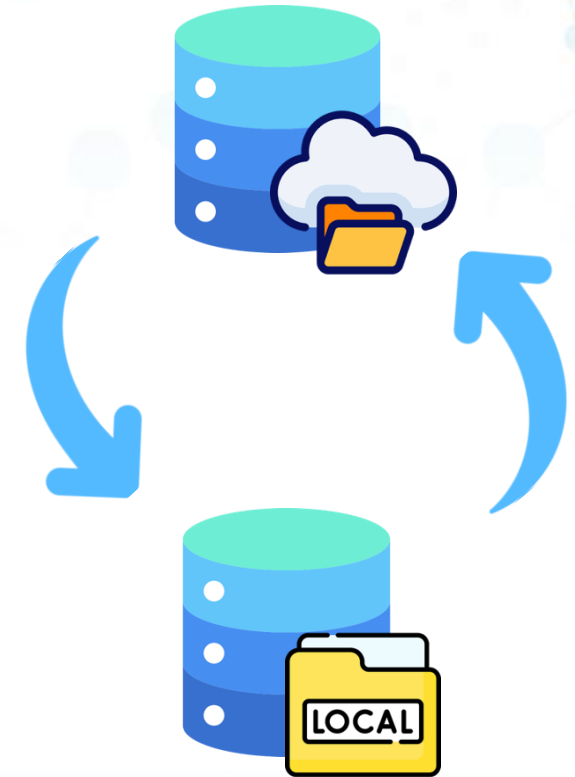
Behind the Surface: the Hidden Data Infrastructure



Database Access for Café Chatbot



Hey App! Show me that café I saved last week in Kyoto. Now that I'm back in New York, can you find me the closest café that's most similar to it?



What we see: the Application Layer



Application Layer:

- 🍊 Daily Schedule Management
- 🍊 Instant Social Networking
- 🍊 Interactive Computer Use
- 🍊 Digital World ...

Applications and Databases



Application Layer:

- 🍊 Daily Schedule Management
- 🍊 Instant Social Networking
- 🍊 Interactive Computer Use
- 🍊 Digital World ...

Database Layer:

- 🍊 **Efficiency & Organization** - Structured storage and optimized data retrieval
- 🍊 **Integrity & Reliability** - Consistent and accurate transaction handling
- 🍊 **Insight & Intelligence** - Transforming raw data into meaningful analysis
- 🍊 ...



Database Coding: Categories

Input Source	Code
Tables	Python Pandas
Databases	SQL
Documents	Private Lib Code
Spreadsheet / Notebook	Excel / Google Sheet Functions

Data Centric Code Generation (DCCG)



Database Layer:

- 🍊 **Efficiency & Organization** - Structured storage and optimized data retrieval
- 🍊 **Integrity & Reliability** - Consistent and accurate transaction handling
- 🍊 **Insight & Intelligence** - Transforming raw data into meaningful analysis
- 🍊 ...

Data Centric Code Generation

Automate coding for database interaction



Questions

- **Can AI streamline application development?**
 - Agentic solutions for automatic data coding
- **Are database AI models reliable and robust?**
 - Benchmarking for DB-AI models
- **Can database agents enable social applications?**
 - Family care and BIM applications

1. Data-Centric Code Generation (DCCG)




Automated generation of code driven by data semantics, modeling, and management

DCCG: Classification



Input Source	Target Code
Tables	Python Pandas
Databases	SQL
Documents	Private Lib Code
Spreadsheet / Notebook	Excel / Google Sheet Functions






-  Data Understanding
-  Human Computer Interaction (HCI)
-  Multi-modal Understanding & Design

DCCG: Classification




Input Source	Target Code
Tables	Python Pandas
Databases	SQL
Documents	Private Lib Code
Spreadsheet / Notebook	Excel / Google Sheet Functions




-  Data Understanding
-  Human Computer Interaction (HCI)
-  Multi-modal Understanding & Design

Text-to-SQL

- natural language questions → executable SQL queries

CARS DATA	
	Id
	MPG
	Horsepower

CARS NAMES	
	MakeId
	Model
	Make



What is id of the car with the max horsepower?

```
SELECT Id FROM CARS_DATA  
ORDER BY Horsepower DESC LIMIT 1
```



Text-to-SQL in Chatbot

Queries in a Conversation (Yelp)

My dad is visiting me in Mountain View. What do you think about dinner choices?

I found Cascal, a moderate Spanish restaurant with great tapas platters, appetizers, and drinks. It has a great ambience with live music on the weekends and both indoor and outdoor seating with heat lamps to keep guests warm.

Generated SQL:

```
SELECT * FROM restaurants  
WHERE location = "Mountain View"
```

How about Italian instead? I really like the food there when I visited Florence last year.

I found Vaso Azzurro Ristorante, a moderate Italian restaurant near Mountain View. It has a lively atmosphere and offers excellent customer service. Reviews say their pasta is authentic and perfectly seasoned.

Generated SQL:

```
SELECT * FROM restaurants  
WHERE location = "Mountain View"  
AND cuisines = ANY("Italian")
```

Oh, what is the ham called in Italian?

Prosciutto is the Italian word for ham.

GPT responds directly

Thanks. We'll be visiting Stanford. How about in Palo Alto?

I found Terún, a 4 star Italian and Pizza restaurant near Stanford in Palo Alto. Reviews say the pizza is great and they can seat customers outside without a reservation.

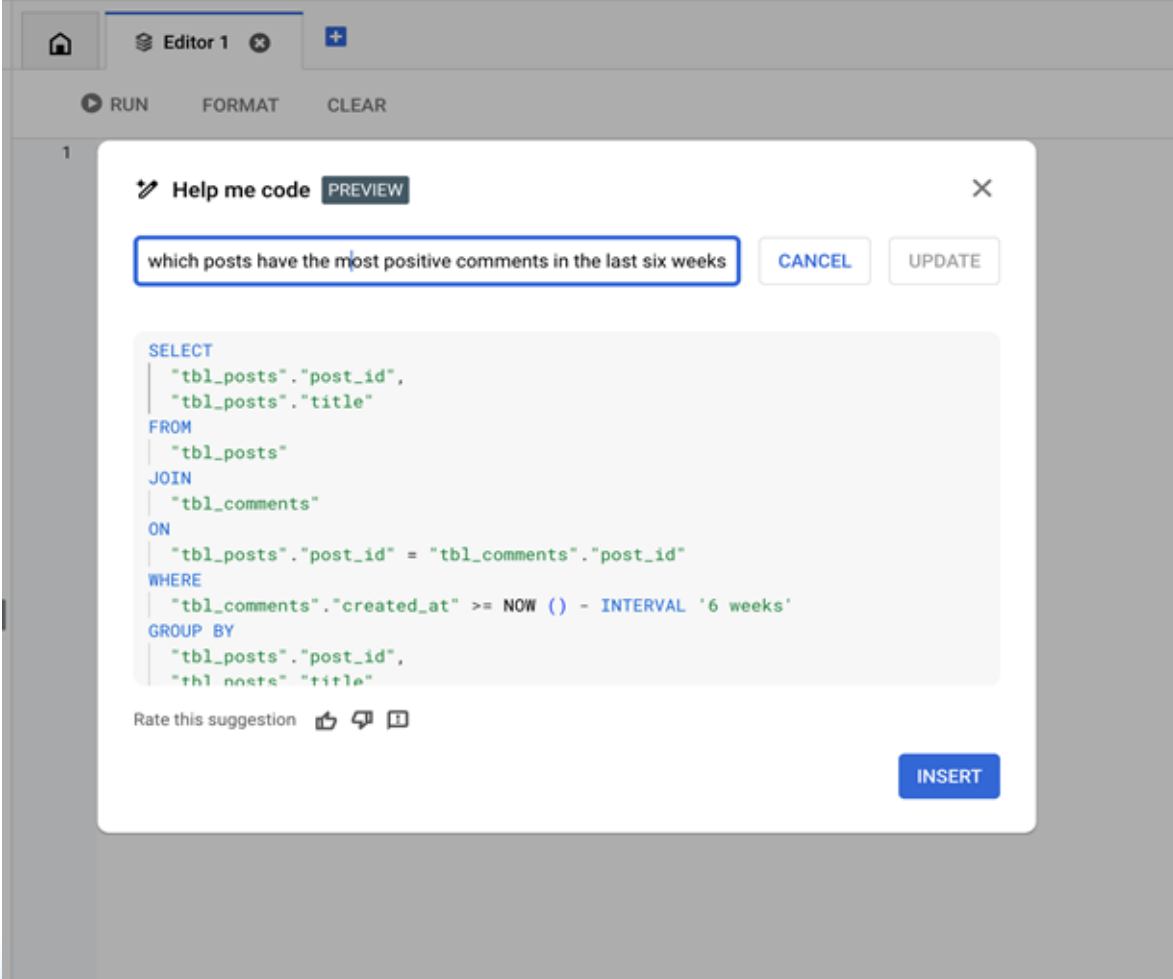
Generated SQL:

```
SELECT * FROM restaurants  
WHERE location = "Palo Alto"  
AND cuisines = ANY("Italian")
```

Text-to-SQL

Developers and AI agents can interact with databases using natural language.

Examples: CloudSQL, AlloyDB



Editor 1

RUN FORMAT CLEAR

1

Help me code PREVIEW

which posts have the most positive comments in the last six weeks CANCEL UPDATE

```
SELECT
| "tbl_posts"."post_id",
| "tbl_posts"."title"
FROM
| "tbl_posts"
JOIN
| "tbl_comments"
ON
| "tbl_posts"."post_id" = "tbl_comments"."post_id"
WHERE
| "tbl_comments"."created_at" >= NOW () - INTERVAL '6 weeks'
GROUP BY
| "tbl_posts"."post_id",
| "tbl_posts"."title"
```

Rate this suggestion

INSERT

Hallucinations in Text-to-SQL

List paper IDs, paper names, and paper descriptions for all papers.

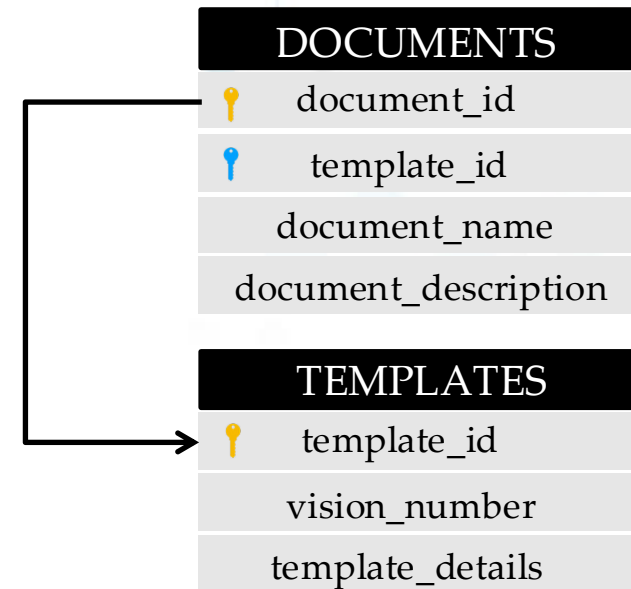
LLM :

```
SELECT paper_id, paper_name, paper_description  
FROM documents;
```



Gold :

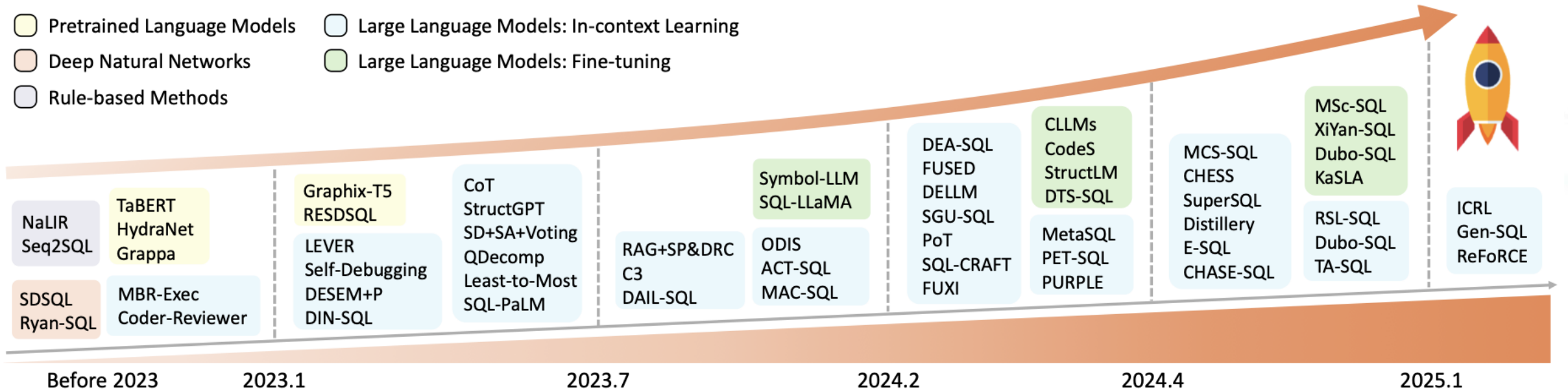
```
SELECT document_id, document_name,  
document_description FROM documents;
```



Graphix-T5: Mixing Pre-Trained Transformers with Graph-Aware Layers for Text-to-SQL Parsing
Li, J. and Hui, B. and ... Cheng, R and Li, Y. (AAAI 2023)

LLM-based Text-to-SQL Evolution

- Pretrained Language Models
- Large Language Models: In-context Learning
- Deep Natural Networks
- Large Language Models: Fine-tuning
- Rule-based Methods



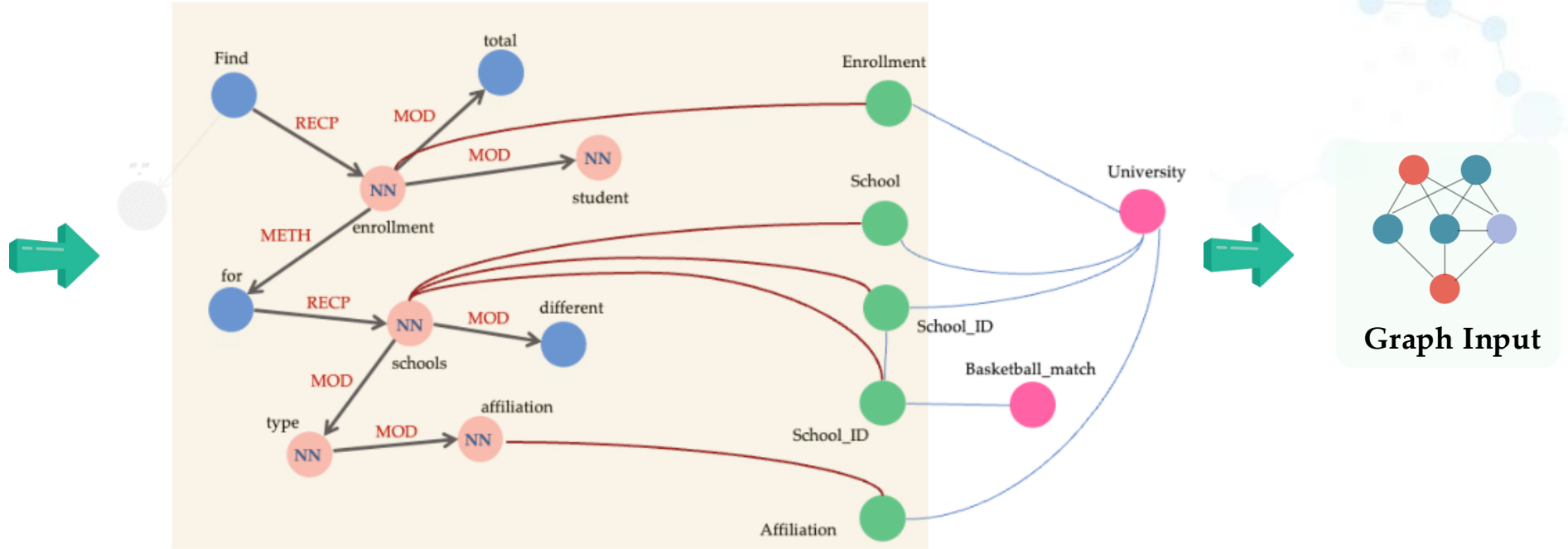
Graphix-T5

Question Input:

Find total enrollment student for schools that has different types of affiliation

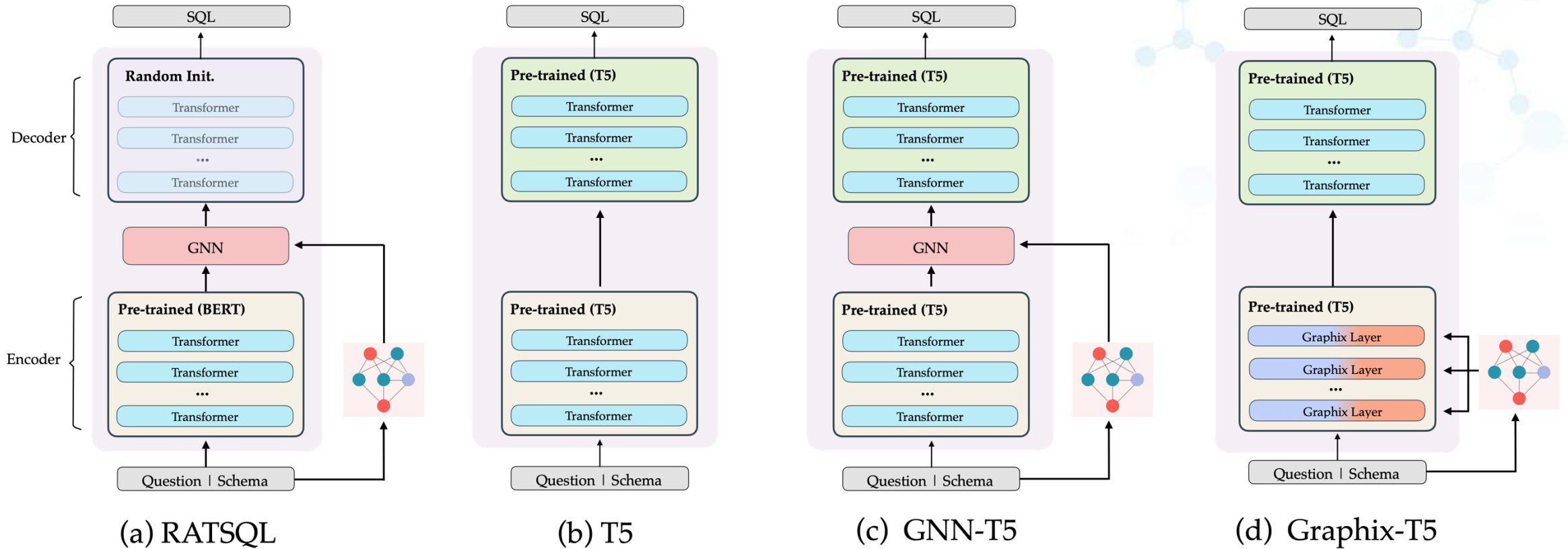
Database Input:

| University: School_ID, School, Enrollment ...
| Basketball_match: School_ID

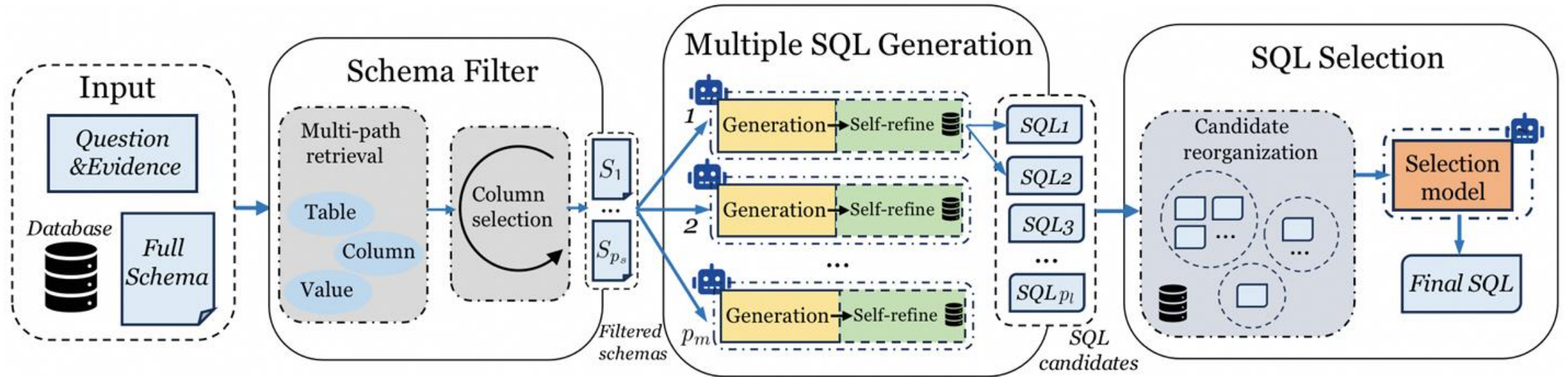


Graphix-T5: Mixing Pre-Trained Transformers with Graph-Aware Layers for Text-to-SQL Parsing
Li, J. and Hui, B. and ... Cheng, R and Li, Y. (AAAI 2023)

Graphix-T5

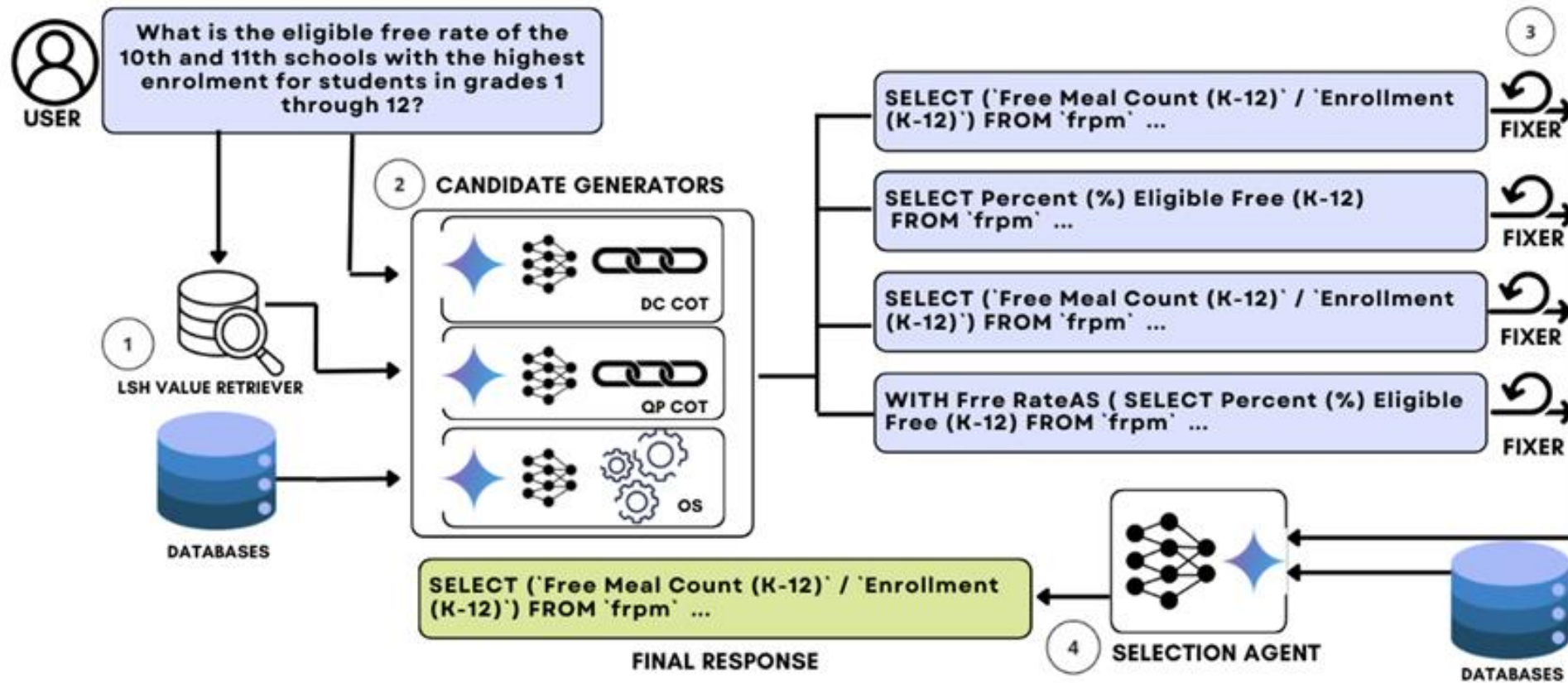


Xiyan-SQL



[3] Liu, Yifu, et al. "Xiyan-sql: A novel multi-generator framework for text-to-sql." *IEEE Transactions on Knowledge and Data Engineering* (2026).

CHASE-SQL



[5] Pourreza, Mohammadreza, et al.

"CHASE-SQL: MULTI-PATH REASONING AND PREFERENCE OPTIMIZED CANDIDATE SELECTION IN TEXT-TO-SQL.", *ICLR 2025*

SHARE

☀ Main Idea: Small

🗨 User Question

For all the set of cards that has Japanese translation, what is the percentage of them are only available in non-foil? In non-foil refers to isNonFoilOnly = 1

❗ Error SQL

```
WITH cte AS (  
  SELECT  
    isNonFoilOnly,  
    language  
  FROM sets  
)  
SELECT  
  CAST(SUM(CASE WHEN  
isNonFoilOnly = 1 AND language  
= 'Japanese' THEN 1 ELSE 0 END)  
AS REAL) * 100.0 / COUNT(*) AS  
percentage_japanese_nonfoil  
FROM cte;
```



SLMs

📍 Action Trajectories

```
# Step 1: Filter language as Japanese  
df1 = df.where ( element = sets.language, filter  
= 'Japanese')  
  
# Step 2: Further filter to non-foil rows  
df2 = df1.where ( element = sets.isNonFoilOnly,  
filter = 1)  
  
# Step 3: Calculate the percentage of non-foil  
Japanese rows among all rows  
res = df.select ( cast(df2.count(), real) * 100 /  
df1.count() )
```

Revise



SLMs

📍 Fixed Action Trajectories

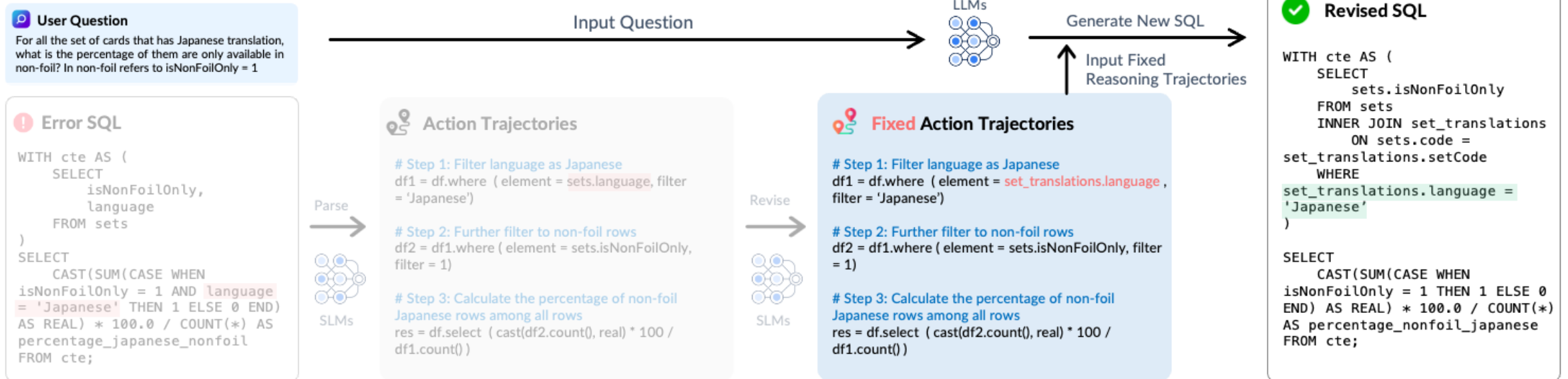
```
# Step 1: Filter language as Japanese  
df1 = df.where ( element = set_translations.language ,  
filter = 'Japanese')  
  
# Step 2: Further filter to non-foil rows  
df2 = df1.where ( element = sets.isNonFoilOnly, filter  
= 1)  
  
# Step 3: Calculate the percentage of non-foil  
Japanese rows among all rows  
res = df.select ( cast(df2.count(), real) * 100 /  
df1.count() )
```

- 1. BAM: Convert SQLs to *step-by-step* reasoning trajectories
- 2. SAM + LOM: Fix errors on *problematic* steps

[4] Qu, Ge, et al. "SHARE: An SLM-based hierarchical action CorREction assistant for text-to-SQL." *ACL* 2025.

SHARE

☀️ **Main Idea: Small Language Models, Infer Step-by-Step Reasoning behind Declarative SQLs**



- 1. BAM: Convert SQLs to step-by-step reasoning trajectories
- 2. SAM + LOM: Fix errors on problematic steps
- 3. LLM: Revise SQLs given fixed action trajectories

[4] Qu, Ge, et al. "SHARE: An SLM-based hierarchical action CorREction assistant for text-to-SQL." *ACL* 2025.

2. Benchmarking of Data Intelligence

How reliable is an AI model for data-centric tasks?

Evaluation framework for text-to-SQL



Spider 1.0

Spider 1.0



Yale Semantic Parsing and Text-to-SQL Challenge

What is Spider?

Nov 12, 2024: We have released [Spider 2.0 full paper, data and code](#). Follow the [guideline](#) to submit your scores to the [leaderboard](#)!

Aug 28, 2024: The early access version of [Spider 2.0](#) (a more realistic and challenging text-to-SQL task) is now available! We expect to release the whole dataset in 1-2 weeks. As this is a preliminary release, there may be errors. Your feedback would be invaluable in refining the dataset!

Spider is a large-scale [complex and cross-domain](#) semantic parsing and text-to-SQL dataset annotated by 11 Yale students. The goal of the Spider challenge is to develop natural language interfaces to cross-domain databases. It consists of 10,181 questions and 5,693 unique complex SQL queries on 200 databases with multiple tables covering 138 different domains. In Spider 1.0, different complex SQL queries and databases appear in train and test sets. To do well on it, systems must *generalize well to not only new SQL queries but also new database schemas*.

Leaderboard - Execution with Values

Our current models do not predict any value in SQL conditions so that we do not provide execution accuracies. However, we encourage you to provide it in the future submissions. For value prediction, your model should be able to 1) copy from the question inputs, 2) retrieve from the database content (database content is available), or 3) generate numbers (e.g. 3 in "LIMIT 3"). *Notice:* Test results after May 02, 2020 are reported on the new release (collected some annotation errors).

Rank	Model	Test
1 Nov 2, 2023	MiniSeek Anonymous Code and paper coming soon	91.2
1 Aug 20, 2023	DAIL-SQL + GPT-4 + Self-Consistency Alibaba Group (Gao and Wang et al., '2023) code	86.6
2 Aug 9, 2023	DAIL-SQL + GPT-4 Alibaba Group (Gao and Wang et al., '2023) code	86.2

[6] Yu, Tao, et al. "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task." EMNLP 2018.

Spider 2.0

Spider 2.0

Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows

ICLR 2025 **Oral**🔥

About Spider 2.0

Spider 2.0 is an evaluation framework comprising 632 real-world text-to-SQL workflow problems derived from enterprise-level database use cases. The databases in Spider 2.0 are sourced from real data applications, often containing over 1,000 columns and stored in local or cloud database systems such as BigQuery and Snowflake. This challenge calls for models to interact with complex SQL workflow environments, process extremely long contexts, perform intricate reasoning, and generate multiple SQL queries with diverse operations, often exceeding 100 lines, which goes far beyond traditional text-to-SQL challenges.

[Paper](#)

[Code](#)

[Twitter](#)

[Submit](#)

..

Leaderboard

[Spider 2.0-Snow](#)

[Spider 2.0-DBT](#)

[Spider 2.0-lite](#)

[Spider 2.0-Snow](#) is a self-contained text-to-SQL task that includes well-prepared database metadata and documentation, includes 547 examples, all hosted on [Snowflake](#), which offers participants free quotas.

Methods with -* use special settings (ground-truth tables) and are not included in the ranking. Since we continually check the accuracy of the evaluation metrics (while the questions remain fixed), the scores may change slightly over time.

Rank	Method	Score
1 <small>Mar 1, 2026</small>	Genloop's Sentinel Agent v2 Pro Genloop	96.70
2 <small>Feb 18, 2026</small>	Native mini usenative.ai	96.53

[7] Lei, Fangyu, et al. "Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows." *ICLR 2025*

BIRD-SQL 2023






<https://bird-bench.github.io/>

Overall Leaderboard

Single-Model Leaderboard

Leaderboard - Execution Accuracy (EX)

	Model	Code	Size	Oracle Knowledge	Dev (%)	Test (%)
	Human Performance <i>Data Engineers + DB Students</i>			✓		92.96
 1 Sep 25, 2025	Agentar-Scale-SQL <i>Ant Group</i> [Pengfei Wang et al. '25]	[link]	UNK	✓	74.90	81.67
 2 Sep 22, 2025	AskData + GPT-4o <i>AT&T CDO - DSAIR</i> [Shkapenyuk et al. '25]		UNK	✓	76.14	80.88
 3 July 14, 2025	LongData-SQL <i>LongShine AI Research</i>		UNK	✓	74.32	77.53
4 Apr 16, 2025	CHASE-SQL + Gemini <i>Google Cloud</i> [Pourreza et al. '24]		UNK	✓	74.90	76.02

BIRD-SQL is Featured by Leading AI Companies

OpenAI Research Products Safety Company

Distil ranks 1st on BIRD-SQL benchmark

Distil, an AI solutions partner to Fortune 500 companies, recently placed 1st on the BIRD-SQL benchmark, the leading text-to-SQL benchmark. Distil's fine-tuned GPT-4o achieved an execution accuracy of 71.83% on the leaderboard and excelled across tasks like query reformulation, intent classification, chain-of-thought, and self-correction, with particularly high performance in SQL generation.

Leaderboard - Execution Accuracy (EX)

	Model	Size	Oracle Knowledge	Dev (%)	Test (%)
	Human Performance Data Engineers + DB Students	N/A	✓	N/A	92.96
1	Distillery + GPT-4o Distil AI Research	?	✓	67.21	71.83
2	ExSL + granite-34b-code IBM Research AI	34B	✓	67.47	68.27
3	RECAP + Gemini Google Cloud	?	✓	66.95	68.27
4	ByteBrain ByteDance Infra Lab	33B	✓	65.45	68.27

Gemini 2.0 Flash

Gemini 2.0 Flash builds on the success of 1.5 Flash, our most popular model yet for developers, with enhanced performance at similarly fast response times. Notably, 2.0 Flash even outperforms 1.5 Pro on key benchmarks, at twice the speed. 2.0 Flash also comes with new capabilities. In addition to supporting multimodal inputs like images, video and audio, 2.0 Flash now supports multimodal output like natively generated images mixed with text and steerable text-to-speech (TTS) multilingual audio. It can also natively call tools like Google Search, code execution as well as third-party user-defined functions.

CAPABILITY	BENCHMARK	DESCRIPTION	Gemini 1.5 Flash 002	Gemini 1.5 Pro 002	Gemini 2.0 Flash Experimental
General	MMLU-Pro	Enhanced version of popular MMLU dataset with questions across multiple subjects with higher difficulty tasks	67.3%	75.8%	76.4%
Code	Natural2Code	Code generation across Python, Java, C++, JS, Go. Held out dataset HumanEval-like, not leaked on the web	79.8%	85.4%	92.9%
	Bird-SQL (Dev)	Benchmark evaluating converting natural language questions into executable SQL	45.6%	54.4%	56.9%
	LiveCodeBench (Code Generation)	Code generation in Python. Code Generation subset covering more recent examples: 06/01/2024 - 10/05/2024	30.0%	34.3%	35.1%
	FACTS Grounding	Ability to provide factuality correct responses given documents and diverse user requests. Held out internal dataset	82.9%	80.0%	83.6%

1000+ Google citations since 2023

BIRD helps to unlock technical contributions

ImageNet

13.2  92.4

Deep Learning

BIRD-SQL: Question Types

Question Type	Sub Type	Question / SQL
Fundamental Type	Match-based	<p>How many gas stations in CZE has Premium gas?</p> <pre>SELECT COUNT(GasStationID) FROM gasstations WHERE Country = 'CZE' AND Segment = 'Premium'</pre>
	Ranking	<p>What are the titles of the top 5 posts with the highest popularity?</p> <pre>SELECT Title FROM posts ORDER BY ViewCount DESC LIMIT 5</pre>
	Comparison	<p>How many color cards with no borders have been ranked higher than 12000 on EDHRec?</p> <pre>SELECT COUNT(id) FROM cards WHERE edhrecRank > 12000 AND borderColor = 'borderless'</pre>
	Counting	<p>How many of the members' hometowns are from Maryland state?</p> <pre>SELECT COUNT(T2.member_id) FROM zip_code AS T1 INNER JOIN member AS T2 ON T1.zip_code = T2.zip WHERE T1.state = 'Maryland'</pre>
	Aggregation	<p>What is the average height of the superheroes from Marvel Comics?</p> <pre>SELECT AVG(T1.height_cm) FROM superhero AS T1 INNER JOIN publisher AS T2 ON T1.publisher_id = T2.id WHERE T2.publisher_name = 'Marvel Comics'</pre>



BIRD-SQL 2023

<https://bird-bench.github.io/>

Large and Realistic Database Values



What is the **average salary** of the worst performing managers?

```
SELECT AVG(CAST(REPLACE(SUBSTR(T1.salary, 4, 7) AS REAL)) FROM  
employee AS T1 JOIN position AS T2 ON T1.positionID = T2.positionID  
WHERE T1.performance = 'Poor' AND T2.positiontitle = 'Manager'
```

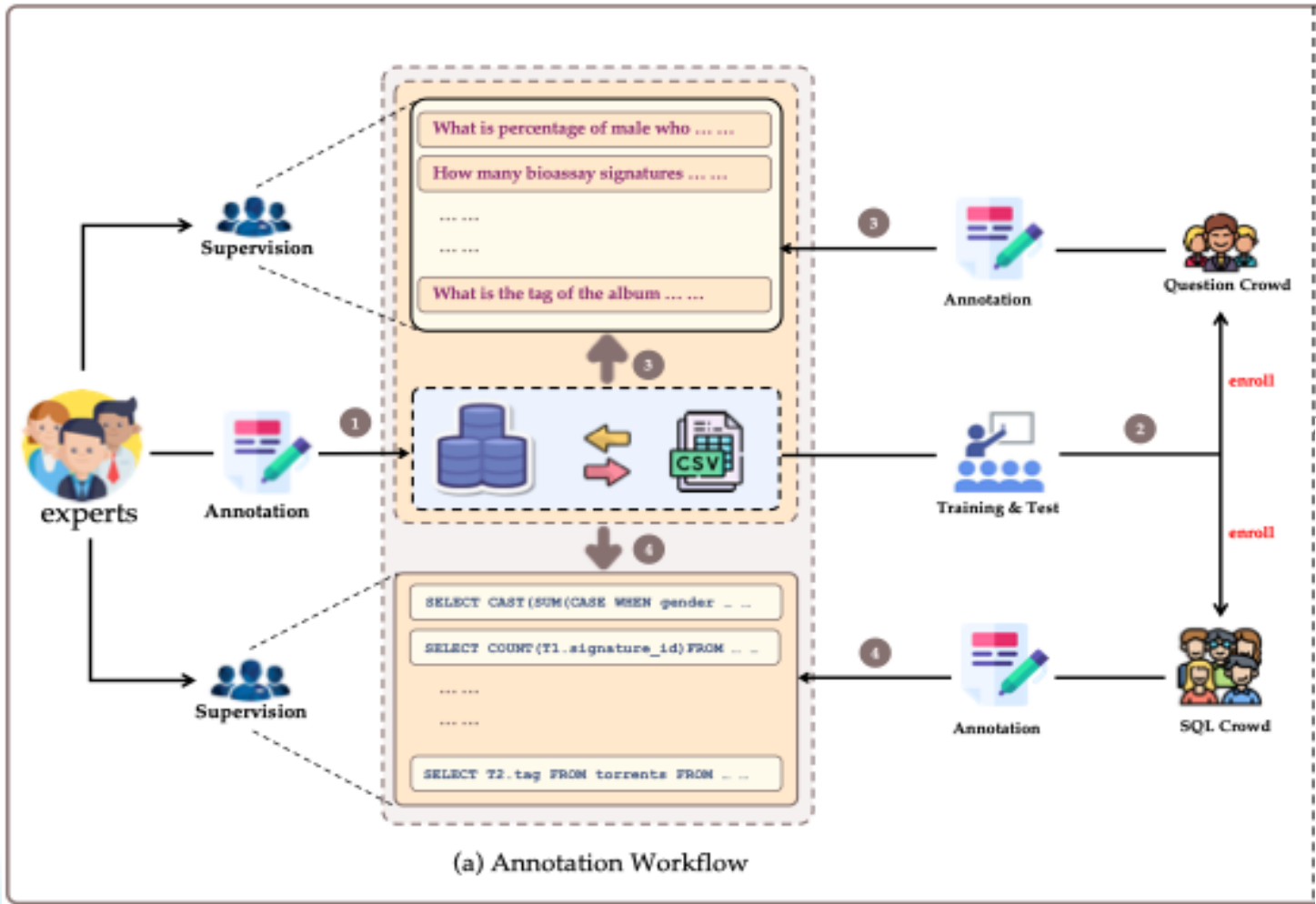


Reasoned Database:

Employees

emp_id	last_name	first_name	salary
0000	Milgrom	Santa	US\$87,800.00
0002	Adams	Sandy	US\$75,900.00
0040	Wood	Emily	US\$69,800.00
...

BIRD 2023 Construction



(a) Annotation Workflow

Query-to-SQL (BIRD 2025-26)



Text-to-SQL



*SELECT-Only
& Plain Text*

Text

Who is the best ...



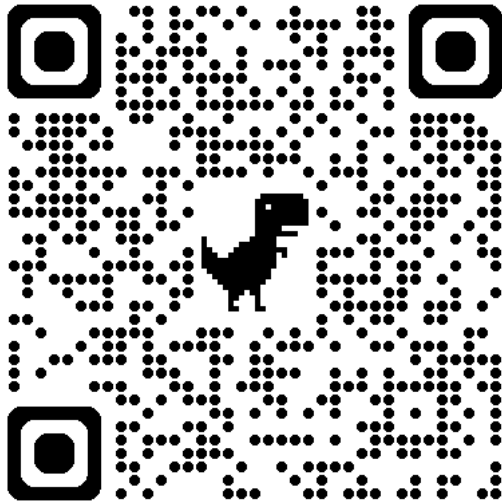
SQL

SELECT ...

... ..

ORDER BY

BIRD 2025 Home Page



BENCHMARK RESEARCH SITE
BIRD 2025

[Home](#) [Ecosystem](#) [Categories](#) [Projects](#) [Why It Matters](#) [Explore](#)

[Collaborate With BIRD 2025](#)

CATEGORIES

Organized around the key aspects of real-world data work

Each category tracks one capability area.



DB Querying

1 project

Evaluation for translating natural language into correct SQL across real-world DB environment.

[BIRD-SQL](#)



BI Analysis & CRUD

3 projects

Evaluation for end-to-end business intelligence workflows and database manipulation beyond simple SELECT queries.

[LiveSQLBench](#)

[BIRD-Interact](#)

[BIRD-Vision \(Data-Vista\)](#)



DB App Debugging

3 projects

Evaluation for diagnosing and fixing failures in database-powered applications and SQL-centric systems.

[BIRD-Critic 1.0](#)

[BIRD-Critic 1.5](#)

[BIRD-Critic 2.0](#)



Human-centric Interaction

2 projects

Evaluation for multi-turn, interactive, agentic collaboration with users on data tasks.

[BIRD-Interact](#)

[Tapilot-Crossing](#)



Data Science

2 projects

Evaluation related to broader data science workflows and interactive analytical problem solving.

[Tapilot-Crossing](#)

[DS-CodeTrans](#)



Digital

2 projects

A category for multimodal or broader digital data-intelligence tasks, including visual and hybrid data workflows.

[BIRD-Critic 2.0](#)

[BIRD-Vision \(Data-Vista\)](#)



BIRD-Critic

BIRD-CRITIC 1.0

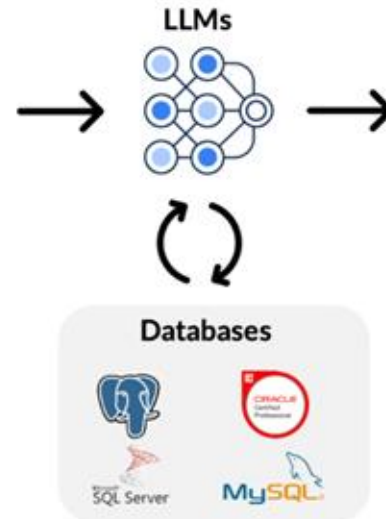
User Issue

I have a certain hierarchy of data in the `card_type` table, where each row represents a type of cards with a unique identifier, a type name, and a reference to its parent card through the `parent_uuid`. The data is structured in a way that cards can be grouped under parent cards, forming a tree-like hierarchy. I initially managed to create a recursive query that fetches the data, but the result isn't in the format I desire. The query correctly returns each card along with a list of parent uuids. However, instead of having the list of parent uuids, I would prefer to have a structured output where each parent card lists all its direct child cards grouped together, forming a tree-like structure. This would help me better visualize the hierarchy and relationships between the cards, with each parent card having an array of its children's uuid values. Can you guide me on how to achieve this transformation using SQL?

Error SQL

```
WITH RECURSIVE nodes AS (  
  SELECT  
    uuid,  
    card_name AS name,  
    ARRAY[]::uuid[] AS parents  
  FROM card_type  
  WHERE parent_uuid IS NULL  
  
  UNION ALL  
  
  SELECT  
    c.uuid,  
    c.card_name AS name,  
    nodes.parents || c.uuid  
  FROM card_type c  
  JOIN nodes ON nodes.uuid =  
  c.parent_uuid  
)  
SELECT * FROM nodes;
```

- Output Format
- Recursive Join
- Child Card Agg



Revised SQL

```
WITH RECURSIVE RCTE_NODES AS (  
  SELECT  
  
  . . .  
  
  , 1 AS lvl  
  , ARRAY[]::uuid[] AS children  
  , true AS has_next  
  AS card_type  
  WHERE parent_uuid IS NULL  
  
  . . .  
  
  , (EXISTS(SELECT 1 FROM card_type  
  AS cat2  
  
  . . .  
  
  SELECT root_uuid AS uuid, root_name,  
  array_agg(children) AS children  
  FROM RCTE_NODES  
  WHERE has_next = false  
  GROUP BY root_uuid, root_name  
  ORDER BY root_uuid
```

Types of User Issues



Query-like

(SELECT-ONLY)

“ I have a query that calculates the number of referrals each user has made. However, I want to **count** a referral only if the referred user has activated their premium account. How can I achieve this?”

Leaderboard:

Leaderboard - BIRD-CRITIC-1.0-Open (570) Verified

OVERALL POSTGRESQL MYSQL SQL SERVER ORACLE

Rank	Model	SR (%)	Institute	Link	Date	Tier
	Human Performance	76.67			2025-07-03	
1	AlphaData <i>CMCC-Department of Digital Intelligence</i>	48.25			2026-02-27	Leading
2	Gemini-3.1-Pro-Preview	45.96			2026-04-08	Elite
3	XiYan-SQL-CRITIC	44.37			2025-11-20	Elite
4	ByteBrain-Agent + Deepseek-r1 <i>ByteDance Infra System Lab x Renmin University of China</i>	43.33			2025-06-10	Elite
5	AlphaData+Deepseek-r1 <i>CMCC-Department of Digital Intelligence</i>	43.15			2026-01-14	Elite
6	Claude-Opus-4-6	40.88			2026-03-02	Elite
7	GPT-5-High	34.96			2025-09-04	Elite
8	Grok-4	33.68			2025-07-18	Superior
9	DeepSeek-R1	33.51			2025-04-20	Superior
10	Claude-Sonnet-4.5	33.33			2026-03-02	Superior
11	o3-mini-2025-01-31	33.33			2025-04-20	Superior
12	o1-preview-2024-09-12	33.33			2025-04-20	Superior
13	GPT-5	33.02			2025-08-15	Superior



BIRD-Interact

BIRD-Interact

Single-turn Text-to-SQL



What are schools whose mailing city address is in Fresno?
Please return their CDSCode.

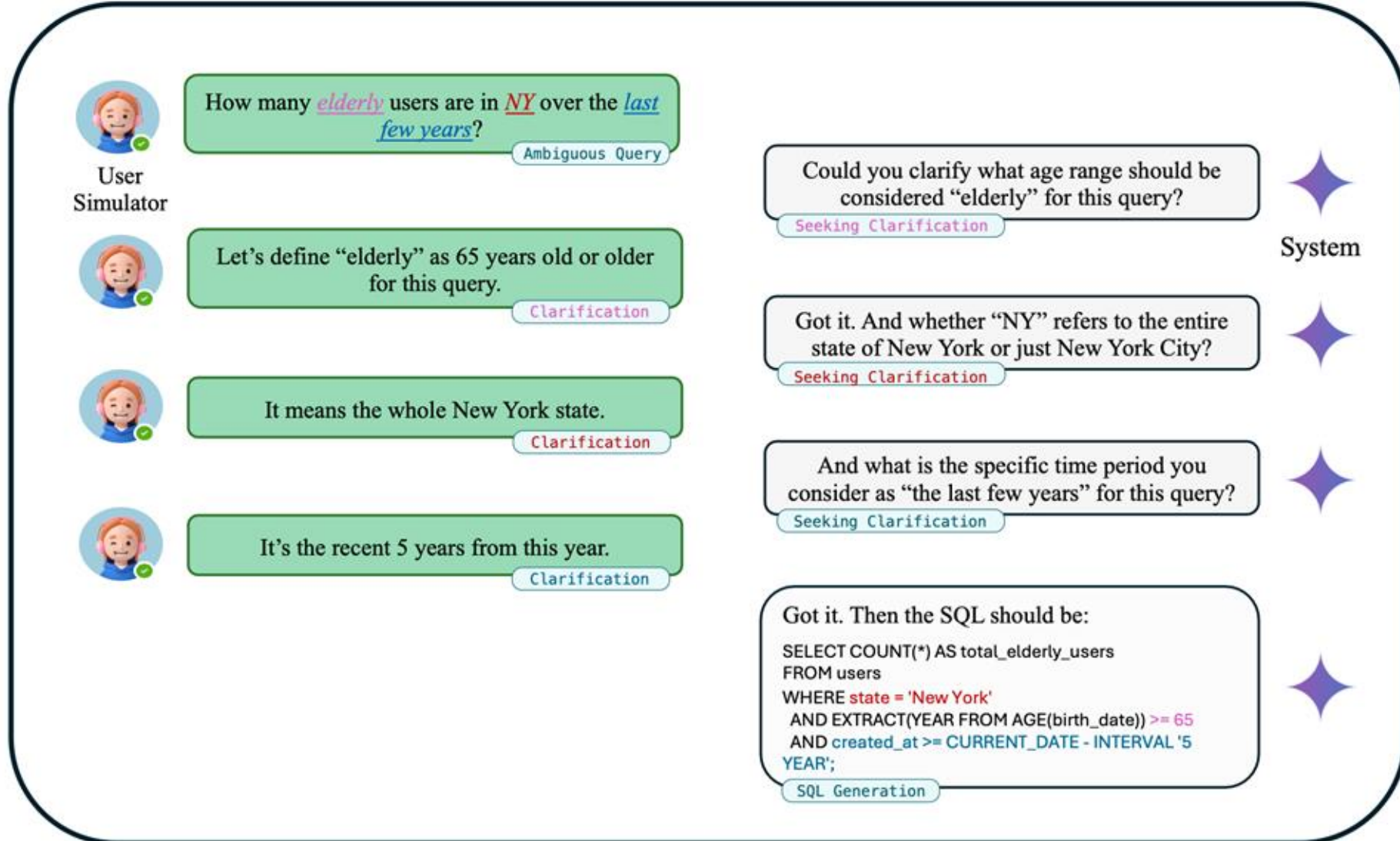
A. Ask question

```
SELECT T2.CDSCode FROM schools AS T2 WHERE  
T2.MailCity = 'Fresno'
```



B. Semantic Parsing

Resolving Ambiguity in Multiple Turns



Interaction Time Scaling Law:

ITS Law: A model satisfies this law if, given enough interactive turns, its performance can match or even surpass that of the idealized single-turn task.

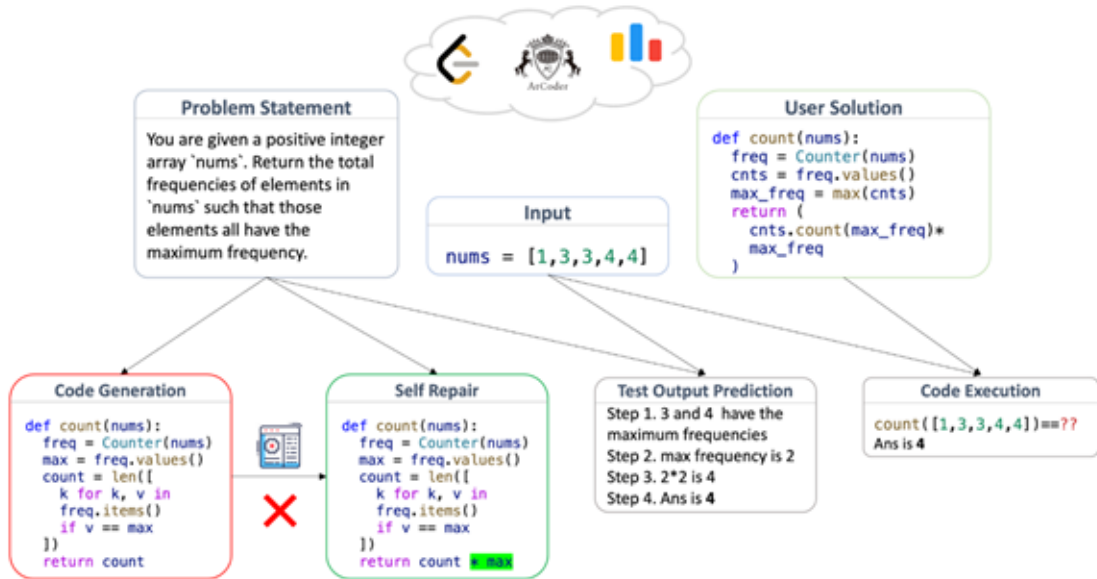


Figure 4: The performance of different LLMs with different user patience on BIRD-INTERACT-LITE. The red line denotes *a*-Interact mode (-a); the blue line denotes *c*-Interact mode (-c). And the dotted line (*Idealized* Performance) denotes the performance under ambiguity-free single-turn text-to-SQL.



BIRD-LiveSQLBench

SQL Periodic Update



🏆 G-Pass@k Leaderboard 🏆

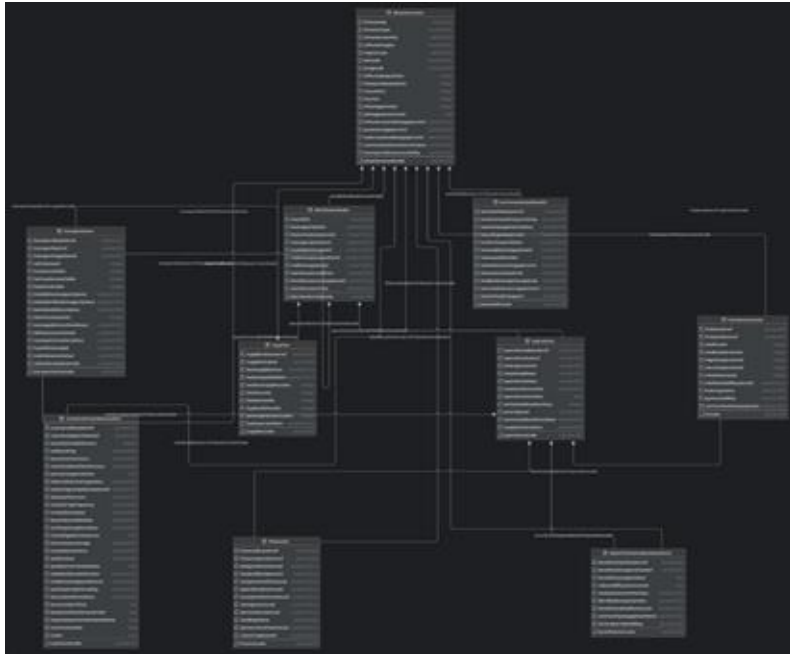
LiveMathBench-202412

LLMs	Greedy	G-Pass@16_0.5	G-Pass@16_0.75	G-Pass@16_1.0	mG-Pass@16
Llama-3.1-8B-Instruct 🤖	24.0	18.2	11.3	4.5	10.4
Qwen2.5-7B-Instruct 🤖	37.0	36.5	27.2	16.0	25.8
Llama-3.3-70B-Instruct 🤖	40.3	36.2	28.9	19.1	27.5
InternLM3-8B-Instruct 🤖	44.5	43.0	35.4	23.0	33.6
Claude-3.5-Sonnet 🤖	46.7	44.1	36.2	26.6	35.3



Multi-Scale RDBs

Base-Scale (36-100 cols)



Large-Scale (>1000 cols)





DDL and DML support

“ Create a report analyzing artists' album release patterns and sonic evolution over time ... ”

SQL

```
WITH artist_releases AS (  
  SELECT  
    art.zid AS artist_id,  
    art.nam AS artist_name,  
    alb.zid AS album_id,  
    alb.anam AS album_name,  
    alb.atyp AS album_type,  
    ...  
    WHEN ar.days_since_previous IS NULL THEN 'First Album'  
    WHEN ar.days_since_previous < 365 THEN 'Less than 1 year'  
    WHEN ar.days_since_previous < 730 THEN '1-2 years'  
    ELSE 'Over 2 years'  
  END AS release_cycle  
FROM artist_releases ar  
JOIN album_features af ON ar.album_id = af.album_id  
WHERE ar.artist_name IN (  
  SELECT artist_name  
  FROM artist_releases  
  GROUP BY artist_id, artist_name  
  HAVING COUNT(*) >= 3  
)  
ORDER BY ar.artist_name, ar.release_date DESC  
LIMIT 100;
```

“ I want to automatically **update** a summary for each artist whenever the popularity of one of their tracks changes ... ”

SQL

```
-- Create or replace function to update the artist_summary table  
CREATE OR REPLACE FUNCTION update_popular_track_count()  
RETURNS TRIGGER AS $$  
BEGIN  
  -- Update the total_popular_tracks count for the artist  
  -- First, we find the artist's ID based on the track ID  
  WITH artist_id AS (  
    SELECT art  
    FROM sp_track_artists  
    JOIN sp_tracks ON sp_tracks.zid = sp_track_artists.trk  
    WHERE sp_tracks.spid = NEW.sid  
  )  
  -- Now, we update the count of popular tracks for this artist  
  UPDATE artist_summary  
  SET total_popular_tracks = (  
    SELECT COUNT(*)  
    ...  
  )  
  -- Create trigger to fire after update on pval column of sp_track_info  
  CREATE TRIGGER trigger_update_popular_tracks  
  AFTER UPDATE OF pval ON sp_track_info  
  FOR EACH ROW  
  EXECUTE PROCEDURE update_popular_track_count();
```



Multi-Form Knowledge Bank

```
{
  "type": "domain_knowledge",
  "entries": [
    {
      "name": "Disease Stage",
      "description": "Defines disease stage.",
      "definition": "If their stage event pathologic stage is 'Stage III' or 'Stage IV', then disease stage is 'Advanced Stage', else 'Early Stage'."
    },
    {
      "name": "Tumor Grade",
      "description": "Defines tumor grade.",
      "definition": "If their neoplasm histologic grade is 'G3' or 'G4', then tumor grade is 'High Grade', else 'Low Grade'."
    },
    {
      "name": "Concerning Patient",
      "description": "Defines a concerning patient based on staging and clinical status.",
      "definition": "A high-risk patient will be considered as concerning."
    },
    {
      "name": "High Risk Patient",
      "description": "Defines a high-risk patient based on staging and clinical status.",
      "definition": "A patient is considered high-risk if their stage event pathologic stage is 'Stage III' or 'Stage IV', OR if their vital status is"
    },
    {
      "name": "High Risk Smoker",
      "description": "Defines a high-risk smoker based on number pack years smoked.",
      "definition": "A smoker is considered high risk if their number pack years smoked is more than 30. And moderate means greater than 10 but lower t"
    },
    {
      "name": "Advanced Disease",
      "description": "Defines an advanced disease based on staging.",
      "definition": "A patient is considered advanced disease if their stage event pathologic stage is 'Stage III' or 'Stage IV'."
    },
    {
      "name": "Recommended Follow-up Interval",
      "description": "Defines patient recommended follow-up interval.",
      "definition": "For 'Imaging', the recommended interval is 3 months for high risk patient and 6 month for those not high risk but have barretts, s"
    },
    {
      "name": "Follow-up Types",
      "description": "Defines follow-up types.",
      "definition": "There are two types of follow-up including 'Imaging' and 'Lab Work'."
    },
    {
      "name": "Complete Follow-up",
      "description": "Defines what constitutes complete follow-up information.",
      "definition": "Complete follow-up includes information on drugs, radiations, and general follow-up status (treatment_and_followup.has_drugs_infor"
    }
  ]
}
```

Esophageal Cancer Clinical Knowledge Bank

Risk Factors and Patient Classification

Patients with cumulative exposure exceeding 20 pack years are classified as heavy smokers, representing a substantially elevated risk profile. Those with positive but lower exposure rates are designated as light smokers, with a more moderate risk assessment. For more granular risk assessment, smokers can be classified as high risk (more than 30 pack years), moderate risk (between 10-30 pack years), or low risk (less than 10 pack years).

Body Mass Index provides a standardized assessment of nutritional status and has demonstrated associations with both risk and prognosis:

$$BMI = \frac{\text{Weight}}{(\text{Height}/100)^2}$$

Nutritional status is categorized based on BMI values: below 18.5 indicates underweight, between 18.5-24.9 represents normal weight, 25-29.9 signifies overweight, and 30 or greater constitutes obesity. A nutrition alert is triggered when BMI falls below 17 or exceeds 35, indicating potential nutritional intervention requirements.

Alcohol intake represents another established risk factor. Frequent consumption, defined as intake exceeding three times weekly, carries particular significance. The combination of frequency and volume provides the most comprehensive risk assessment, with daily heavy consumption representing the highest risk category.

Patient risk score offers a comprehensive assessment using weighted risk factors based on clinical significance. The score is calculated by adding assigned points to specific risk factors: 3 points for heavy smoking history (20+ pack years), 2 points for documented Barrett's esophagus, and 4 points for advanced disease (Stage III or IV).

Follow-up and Treatment Approaches

Follow-up protocols vary based on patient risk classification. Two primary follow-up types exist: "Imaging" and "Lab Work." For imaging, the recommended interval is 3 months for high-risk patients, 6 months for non-high-risk patients with Barrett's esophagus, and 12 months for all others. For laboratory work, the recommended interval is 2 months for high-risk patients and 6 months for all others. Immediate follow-up attention is warranted when patients with Stage III or IV disease either lack proper follow-up documentation or haven't been seen in over 6 months.

Treatment approaches are categorized based on timing relative to surgery. Neoadjuvant therapy refers to treatment administered before primary surgical intervention, aimed at tumor downsizing or improving operative outcomes. This is indicated by a documented history of pre-surgical treatment. Adjuvant therapy encompasses treatments delivered after primary surgery to reduce recurrence risk, indicated by the presence of postoperative treatment information. Treatment categories include "Neoadjuvant Only," "Adjuvant Only," or "Neither."

Performance status provides critical insights into treatment tolerance and prognosis using the Eastern Cooperative Oncology Group scale. Good performance encompasses scores of 0-1, moderate performance includes scores of 2-3, and poor performance is designated by scores exceeding 3. These classifications strongly influence treatment selection and intensity.

Leaderboard:

Model Base

Agent

Only models with results for *all* selected Base datasets are shown. Switch benchmark sets to compare Base and Large independently.

Rank / Window	Model	Organization	Success Rate (%) ↓	Avg. Cost (USD) / Task	Link
1 2025-05 – 2025-09	Gemini 3.1 Pro	Google	43.10	-	Link
2 2025-05 – 2025-09	Claude Opus 4.6	Anthropic	39.43	\$ 0.0990	Link
3 2025-05 – 2025-09	GPT-5.5 (xhigh)	OpenAI	37.36	-	Link
4 2025-05 – 2025-09	GPT-5.5 (low)	OpenAI	37.24	-	Link
5 2025-05 – 2025-09	Kimi K2.6	Moonshot AI	36.43	-	Link
6 2025-05 – 2025-09	GLM-5.1	Zhipu AI	35.29	-	Link
7 2025-05 – 2025-09	Qwen3.6 Max	Alibaba	33.79	-	Link
8 2025-05 – 2025-09	GPT-5.4	OpenAI	33.56	-	Link
9 2025-05 – 2025-09	GPT-5.3 Codex	OpenAI	33.33	-	Link

Model Base

Agent

Only models with results for *all* selected Base datasets are shown. Switch benchmark sets to compare Base and Large independently.

Rank / Window	Model	Organization	Success Rate (%) ↓	Avg. Cost (USD) / Task	Link
1 2025-05 – 2025-09	Claude Opus 4.6 (OpenHands CLI)	Anthropic	41.60	-	Link
2 2025-05 – 2025-09	Claude Sonnet 4.5 (OpenHands CLI)	Anthropic	38.99	-	Link
3 2025-05 – 2025-09	Kimi 2.5 (OpenHands CLI)	Moonshot AI	35.99	-	Link
4 2025-05 – 2025-09	MiniMax M2.1 (OpenHands CLI)	MiniMax	31.62	-	Link
5 2025-05 – 2025-09	Qwen3 Coder 480B (OpenHands CLI)	Alibaba	26.77	-	Link
6 2025-05 – 2025-09	GLM 4.7 (OpenHands CLI)	Zhipu AI	23.99	-	Link

Why does AI need an index?

- As AI becomes more capable and general-purpose, its capability can no longer be captured by **a single benchmark score**.



- AI evaluation is shifting:

single benchmark
accuracy



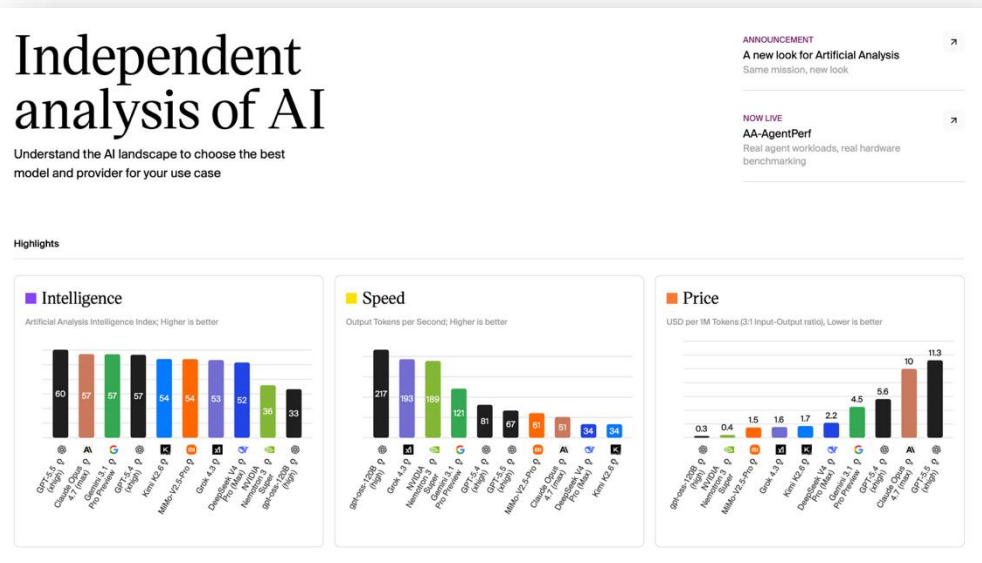
multidimensional
capability indices

Why does AI need an index?

Some domains and their Indices:

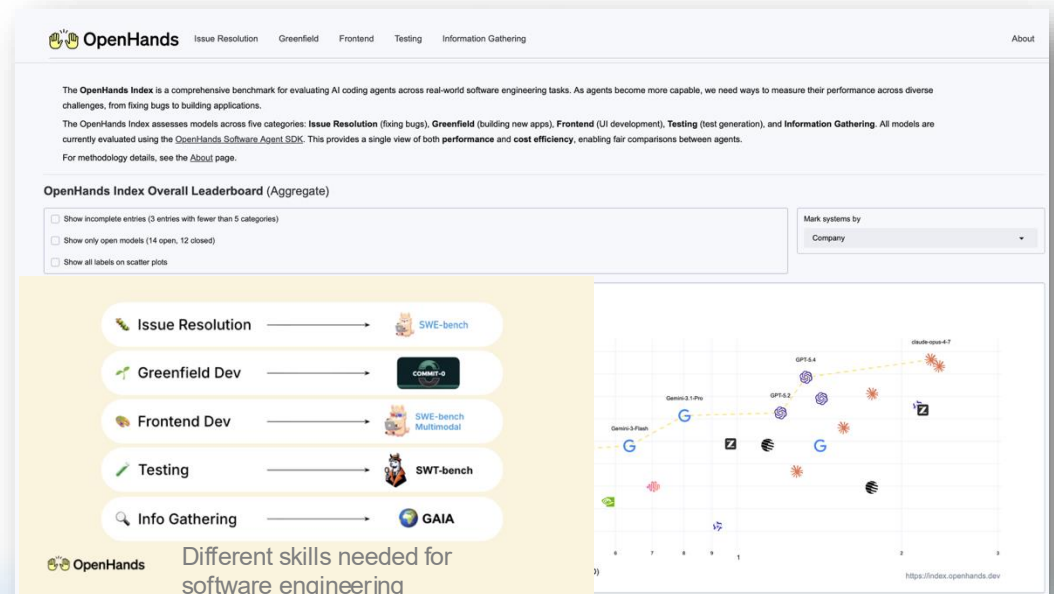
General Domain: Artificial Analysis Index

a horizontal index for general AI model capability, covering reasoning, knowledge, mathematics, coding, science, and agentic tasks.



Software Engineering: OpenHands Index

a vertical index for software engineering agents, covering issue resolution, frontend development, testing, greenfield coding, information gathering, cost, and runtime.



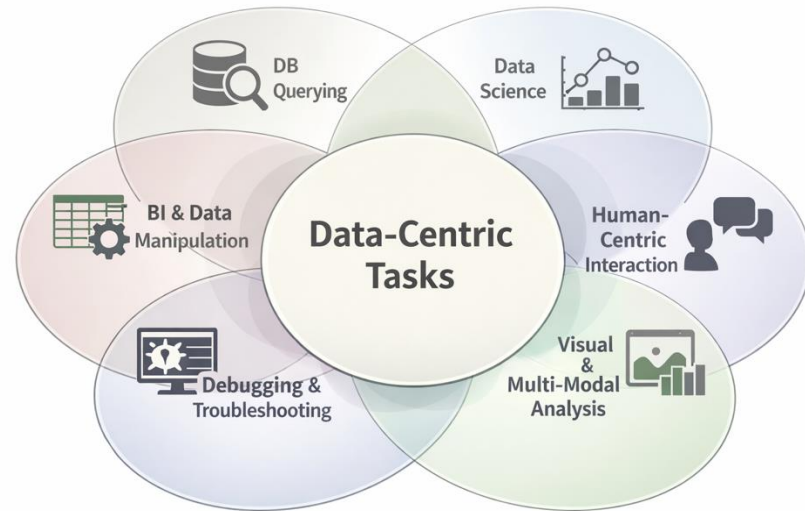
Limitations of Data-Centric Benchmarks

Classic data-centric benchmark tests **narrow skill**:

- **WikiSQL**: single-table natural language to SQL
- **Spider**: complex, cross-domain text-to-SQL
- **BIRD**: large-scale database-grounded text-to-SQL
- **BIRD-Critic**: SQL debugging tasks
- ...

Multi-Aspects of Data Intelligence

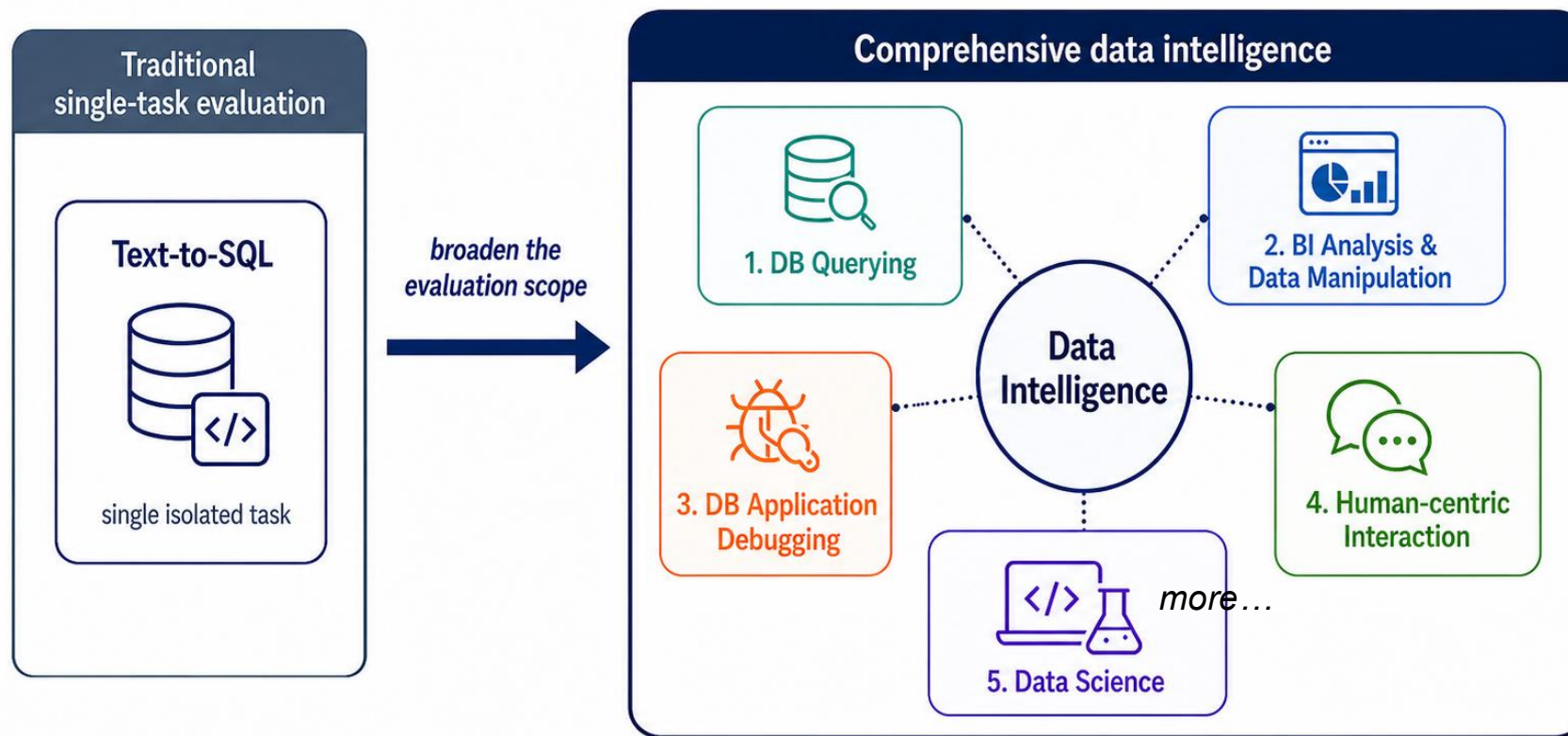
BUT, real data work also requires many aspects, e.g. **BI analysis, data transformation, debugging, human-AI interaction, multimodal data understanding, and data science reasoning.**



Data Intelligence Index

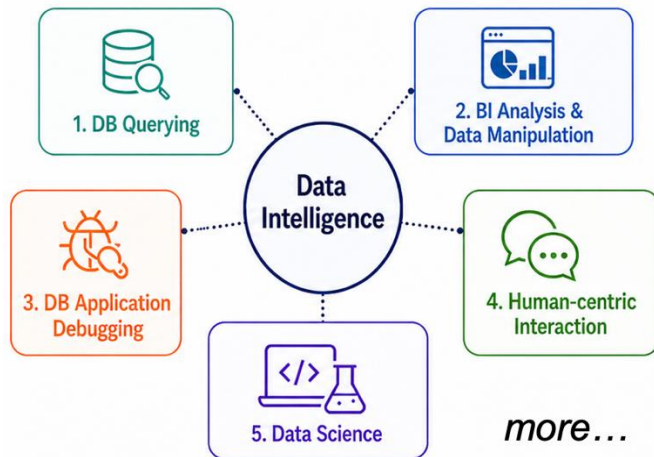
From Single-Task Evaluation to Comprehensive Data Intelligence

Data work is more than Text-to-SQL



Data Intelligence Index

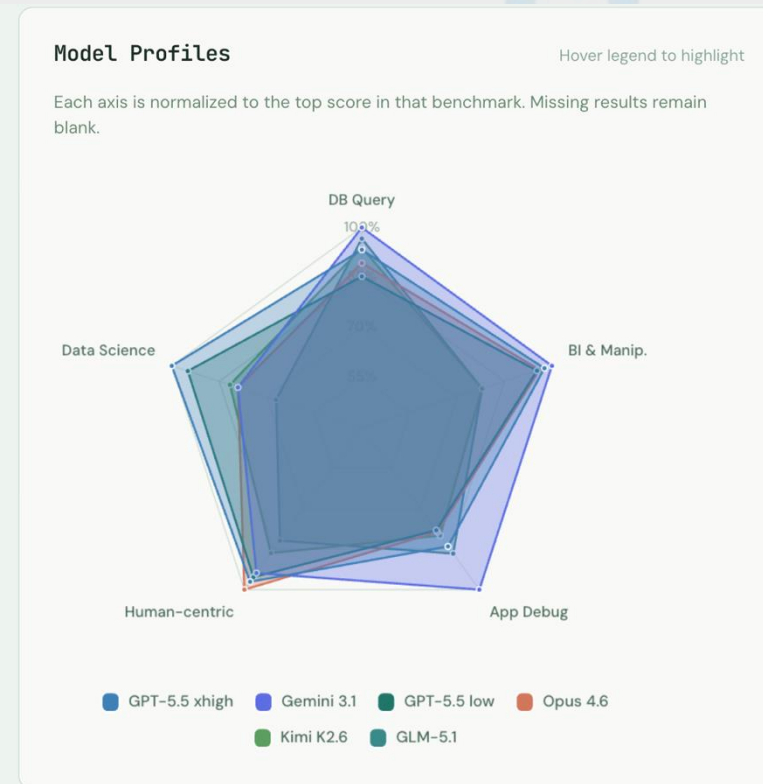
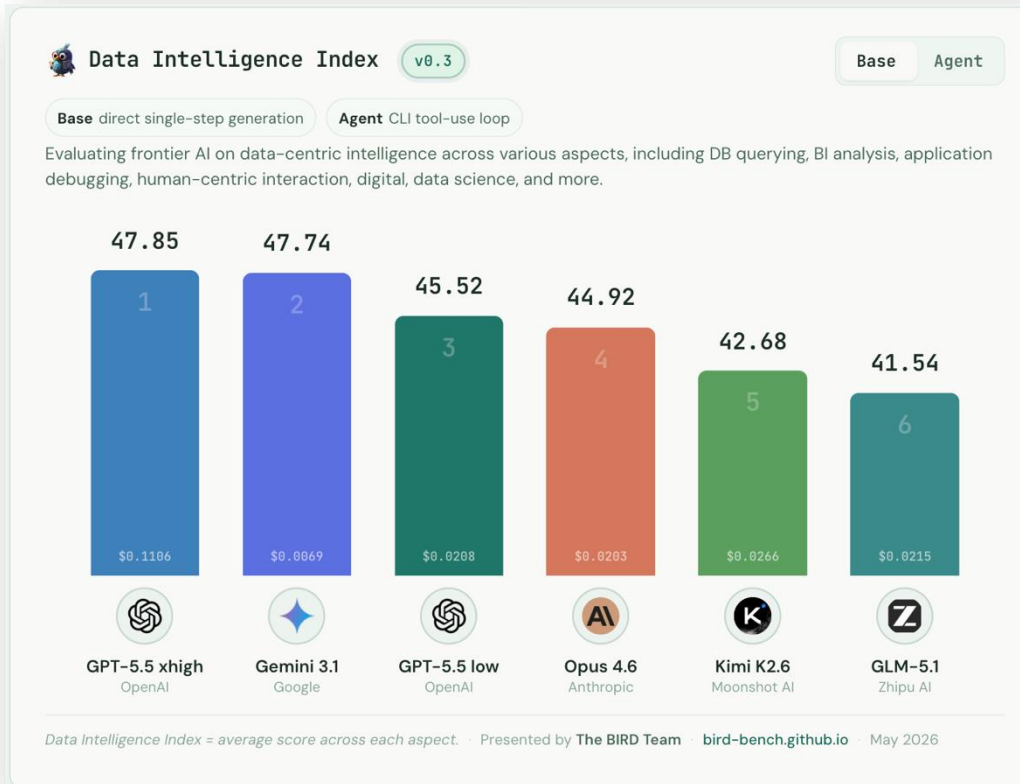
Data Intelligence Index: from single-task to comprehensive data-intelligence.



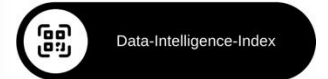
- **DB Querying** — *natural language to SQL over real-world database schemas*
- **BI Analysis & Data Manipulation** — *BI reasoning plus database operations such as UPDATE, CREATE, and DELETE*
- **DB Application Debugging** — *diagnosing and fixing broken SQL/database applications*
- **Human-centric Interaction** — *multi-turn, ambiguous, collaborative data interaction*
- **Data Science** — *data-science pipelines and code translation across data tasks*

Data Intelligence Index

Current Rankings



More results ↓



Data Intelligence Index

🏆 Claude Opus 4.6 Leads

Dominates overall with **45.7** index score. However, even the top model scores under 50, signaling that "Data Intelligence" remains an unsolved research frontier.

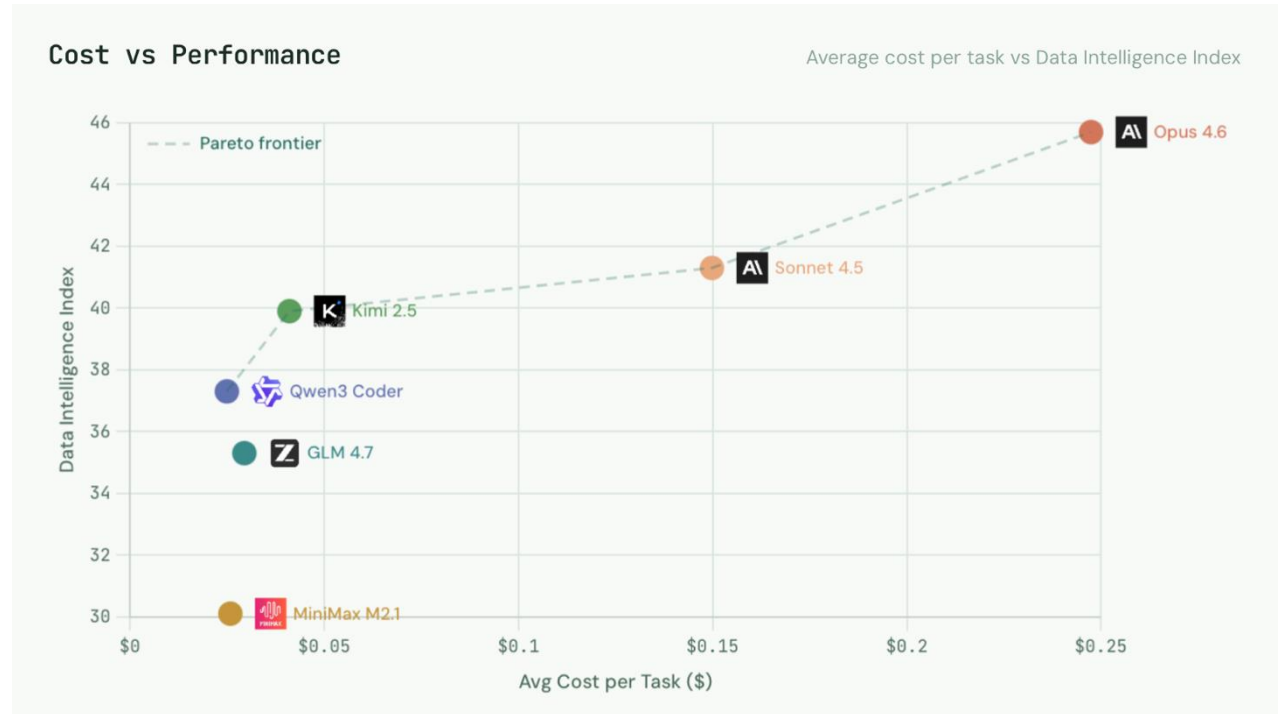
👤 Human-Centric Gap

Interaction and debugging are major bottlenecks. BIRD-Interact scores peak at only **28.8**, highlighting the difficulty of multi-turn agentic reasoning.

💰 Cost-Performance Efficiency

Kimi 2.5 and **Qwen3 Coder** provide exceptional value. Kimi 2.5 achieves ~40 score at 1/4th the cost of leading models.

COST VS PERFORMANCE PARETO FRONTIER



Data Intelligence Index

1



We evaluated 6 frontier models in the current Base setting.

The top results are tightly clustered: GPT-5.5 xhigh leads with a DII score of 47.85, followed by Gemini 3.1 at 47.74, GPT-5.5 low at 45.52, and Opus 4.6 at 44.92.

2



The overall race is close.

The gap between GPT-5.5 xhigh and Gemini 3.1 is only 0.11 DII points, showing that no single model dominates the full data-centric stack.

3



Gemini 3.1 is strongest on SQL-centric accuracy.

Gemini 3.1 leads Mini-Dev (multi-dialect) with 50.00, LiveSQLBench DII with 42.33, and BIRD-Critic with 50.00, making it the most consistent model across DB querying, BI/data manipulation, and app debugging.

4



Human-centric interaction favors Opus 4.6.

Opus 4.6 has the best Human-centric aspect score at 58.28, ahead of GPT-5.5 xhigh at 56.58 and GPT-5.5 low at 55.58.

5



GPT-5.5 xhigh stands out in data science.

On DS-CodeTrans, GPT-5.5 xhigh scores 52.67, ahead of GPT-5.5 low at 50.00 and Kimi K2.6 at 43.00.

6



Coverage still matters.

GLM-5.1 is competitive on Mini-Dev (multi-dialect) (48.33) and BIRD-Critic (43.33) but trails on DS-CodeTrans (35.33), while Kimi K2.6 is stronger on BIRD-Interact (60.33) than on LiveSQLBench DII (33.00).

3. AIXDB for Social Good

Family care and BIM applications

The Jockey Club SMART Family-Link Project



Computer Science

Prof. Reynold Cheng
 Professor, Department of Computer Science,
 Associated Director of Institute of Data Science, HKU



School of Nursing

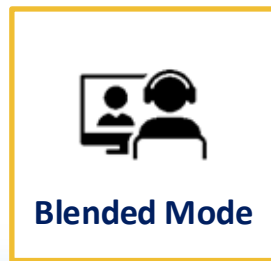
Prof. Kelvin Wang
 Professor, School of Nursing,
 Li Ka Shing Faculty of Medicine, HKU



Social Work

Prof. Daniel Wong
 Chair Professor, Department of Social Work and
 Social Administration, HKBU

Prof. Ng Siu Man
 Professor, Department of Social Work and Social
 Administration, HKU



Sector Capacity Building for Innovation through Multi-disciplinary Collaboration

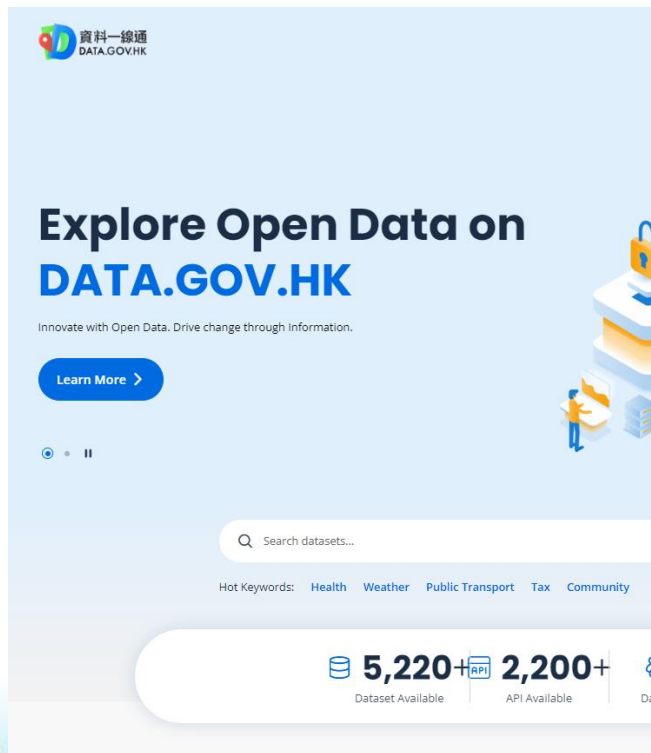


11 NGO partners

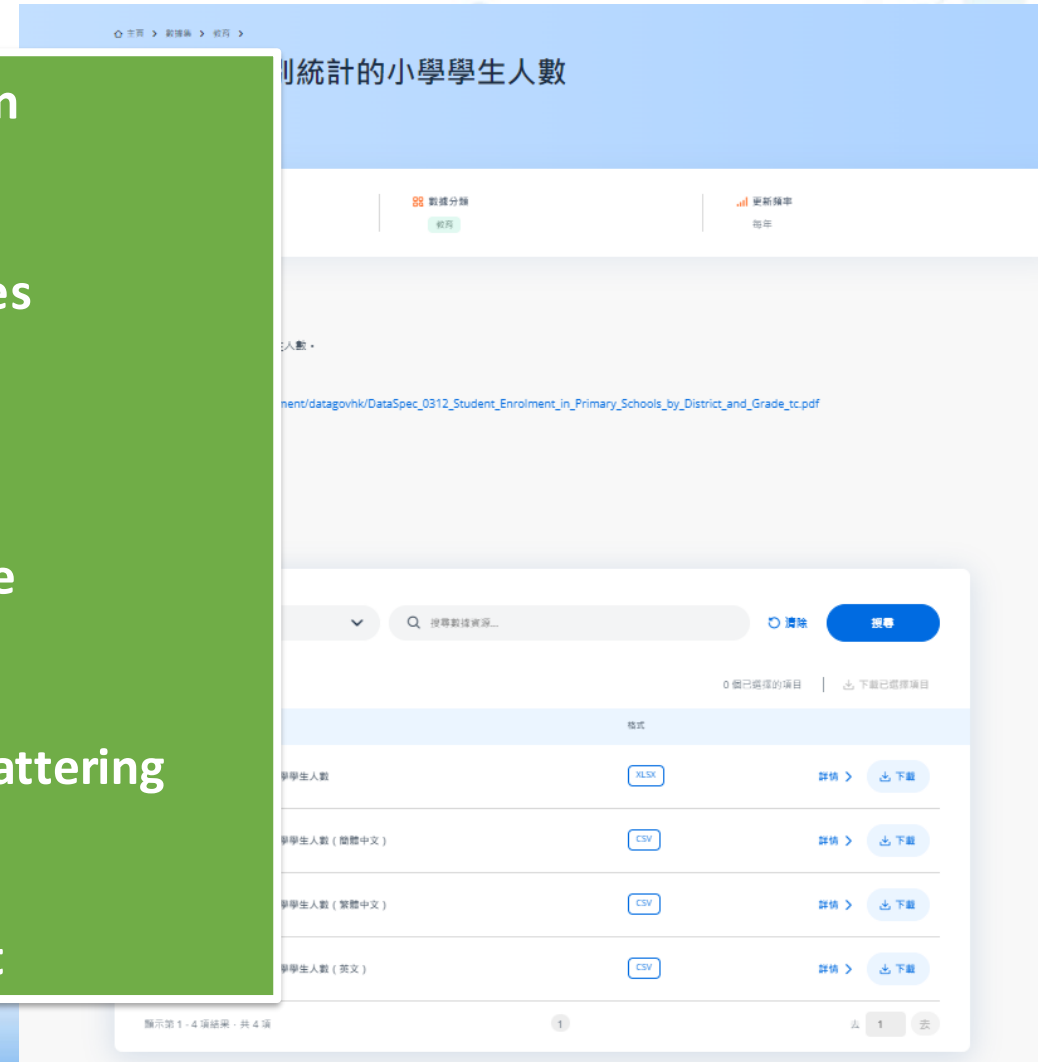


Open Family Data

- 10 open datasets from three go



Child Protection
Elder Abuse
Ethnic Minorities
Population
Poverty
Sexual Violence
Single Parents
Spouse/Cohabitant Battering
Suicide
Unemployment



DataBot

AI chatbot for social data analysis



 **DataBot**

Hello! 有咩可以幫到你？

我想知道2021年男性失業率係幾多？

 **DataBot** Unemployment

2021年男性失業率係6.1%。

Source: Census and Statistics Department

女性又係幾多？



Case study

	How many physical elder abuse cases were recorded in Hong Kong from 2018 to 2022?	
--	--	--



Application of Graphix-T5 in BIM

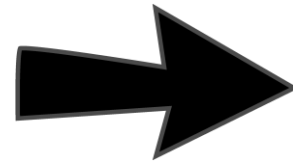
- Application in **Building Information Management (BIM)** system

With Mengtian Yin & Prof. Llewellyn Tang in Faculty of Architecture.>



a) The picture that user requests

Semantic Parsing



Execute BIMQL

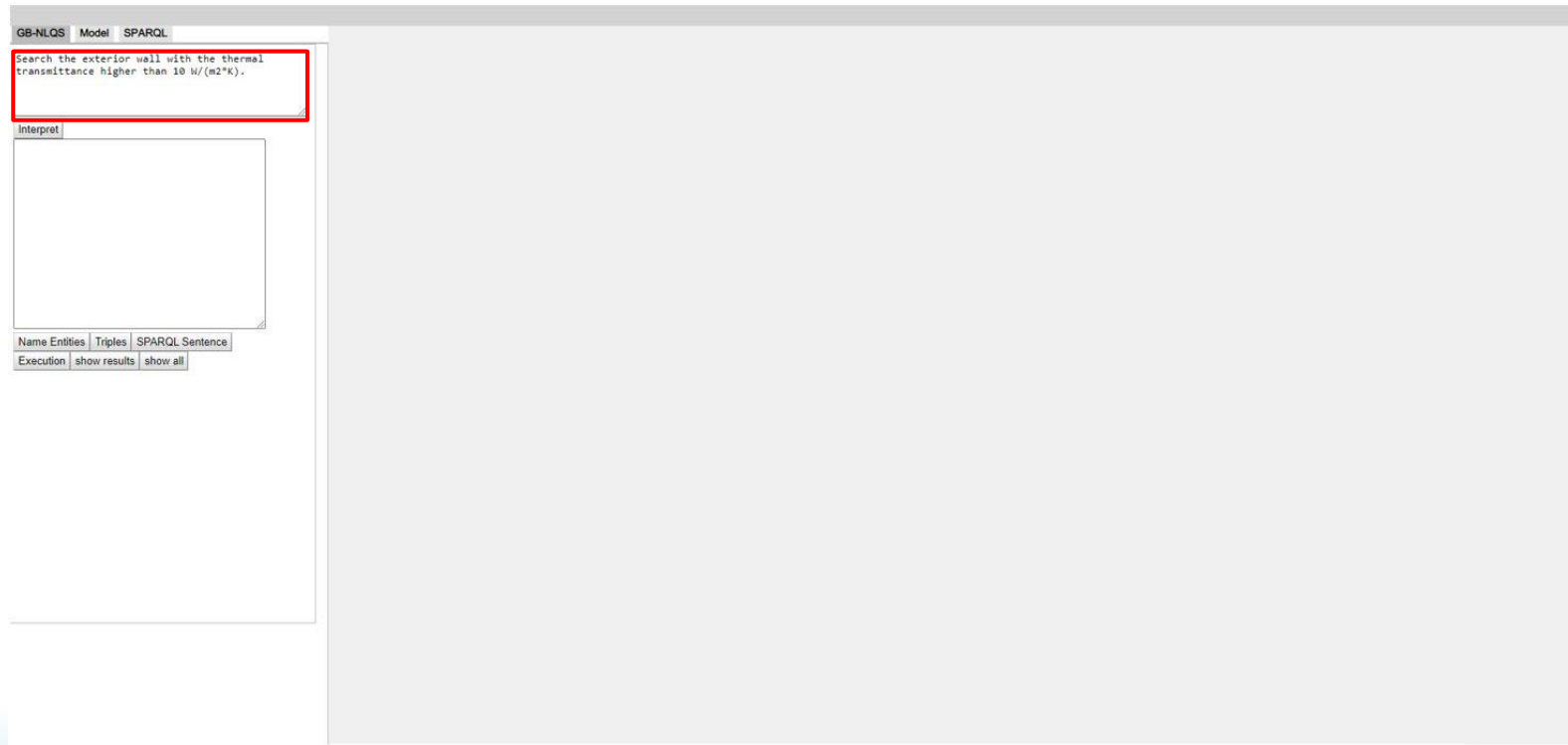


b) The rendered model graph

Demo (App in BIM):

- The Actual Procedures by Semantic Parsing:

<Under cooperation with Mengtian Yin & Prof. Llewellyn Tang in Faculty of Architecture.>



1) Users generate instructions

Demo (App in BIM):

- The Actual Procedures by Semantic Parsing:

<Under cooperation with Mengtian Yin & Prof. Llewellyn Tang in Faculty of Architecture.>



The screenshot displays a software interface with a SPARQL query editor on the left and a 3D architectural model of a building on the right. The query is as follows:

```
SELECT ?exterior3_0 ?wall4_0 ?
thermaltransmittance7_0 WHERE{
?thermaltransmittance7_0
thermaltransmittance7_0
bim:isAPropertyOfSublevelValue ?
thermaltransmittance7_0value .FILTER
( ?thermaltransmittance7_0value>10.0)
. ?thermaltransmittance7_0 a
IfcWallIfcProperty . /
thermaltransmittance7_0
IfcWallName_IfcProperty/expr:hasString
"ThermalTransmittance" . ?wall4_0
schm:hasProperty ?
thermaltransmittance7_0 . ?
exterior3_0
bim:isAPropertyOfBooleanValue ?
```

Below the query, there are tabs for 'Name Entities', 'Triples', and 'SPARQL Sentence'. The 'Execution' tab is active, showing a list of results:

```
exterior3_0 wall4_0
IfcPropertySingleValue_3762 IfcWallStandardCase_5547
IfcPropertySingleValue_3762 IfcWallStandardCase_4046
IfcPropertySingleValue_3762 IfcWallStandardCase_22640
IfcPropertySingleValue_3762 IfcWallStandardCase_3368
IfcPropertySingleValue_3762 IfcWallStandardCase_4771
IfcPropertySingleValue_3762 IfcWall_5348
IfcPropertySingleValue_3762 IfcWallStandardCase_4871
IfcPropertySingleValue_3762 IfcWallStandardCase_22730
```

The 3D model on the right shows a multi-story building with a red rectangular box highlighting a specific section of the exterior wall.

- 1) Users generate instructions
- 2) Graphix-PLM translates them into BIMQL via semantic parsing.
- 3) Execute BIMQL in the system.

Graphix on Architecture

Application on Building Information System (BIM) : Graphix also achieve success in an interdisciplinary research collaboration with School of Architecture in HKU:



a) The real photograph that user desires

Semantic Parsing



Execute BIMQL



b) The rendered BIM model

- 1) Users generate instructions
- 2) **Graphix-PLM** parse them into BIMQL via semantic parsing
- 3) Generate results by executing **BIMQL** in the BIM.

Demo:



a) Type user queries



b) Generate and execute BIMQL via semantic Parsing

Our Contributions to DCCG

Benchmarking

- BIRD-SQL NeurIPS 2023 DB Spotlight
- BIRD-CRITIC (Bench) NeurIPS 2025 Main
- BIRD-Interact ICLR 2026 Oral
- Tapilot-Crossing (Bench) ICML 2025
- Others Generalist Environment

Methodologies

- Graphix-T5 AAAI 2023 Oral
- TA-SQL ACL 2024 Findings
- Tapilot-Crossing (Agent) ICML 2025
- SHARE ACL 2025 Main
- Bird-RL NeurIPS 2026 Submission
- ZeroEA VLDB 2024

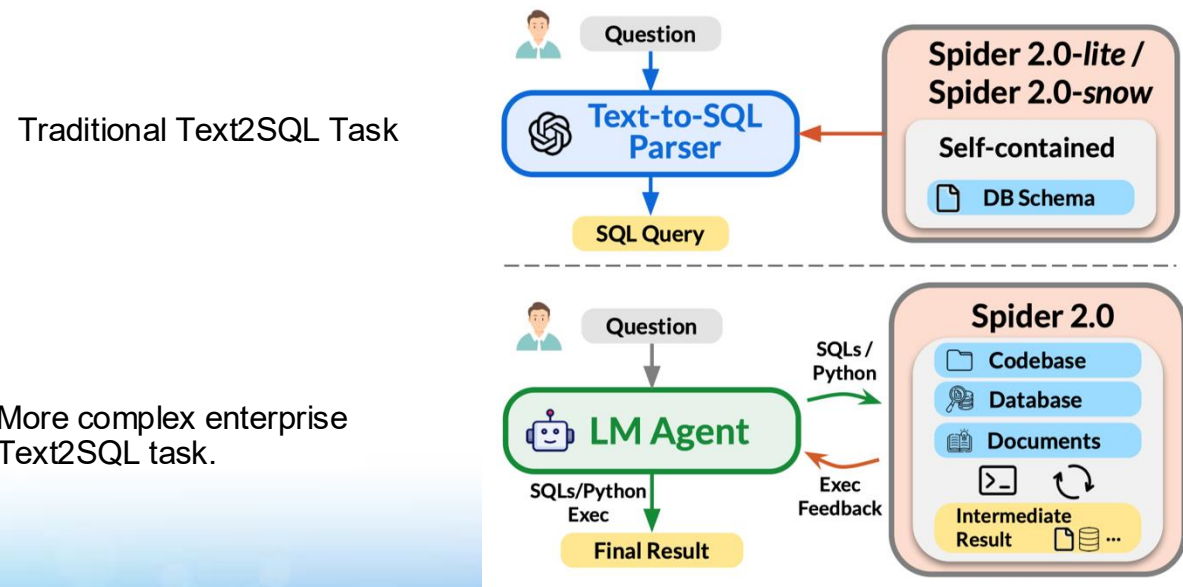
Applications and Agents

- DarGO NeurIPS 2025 Main
- SWE-SQL (Agent) NeurIPS 2025 Main
- New Compute Use

4. Future Work

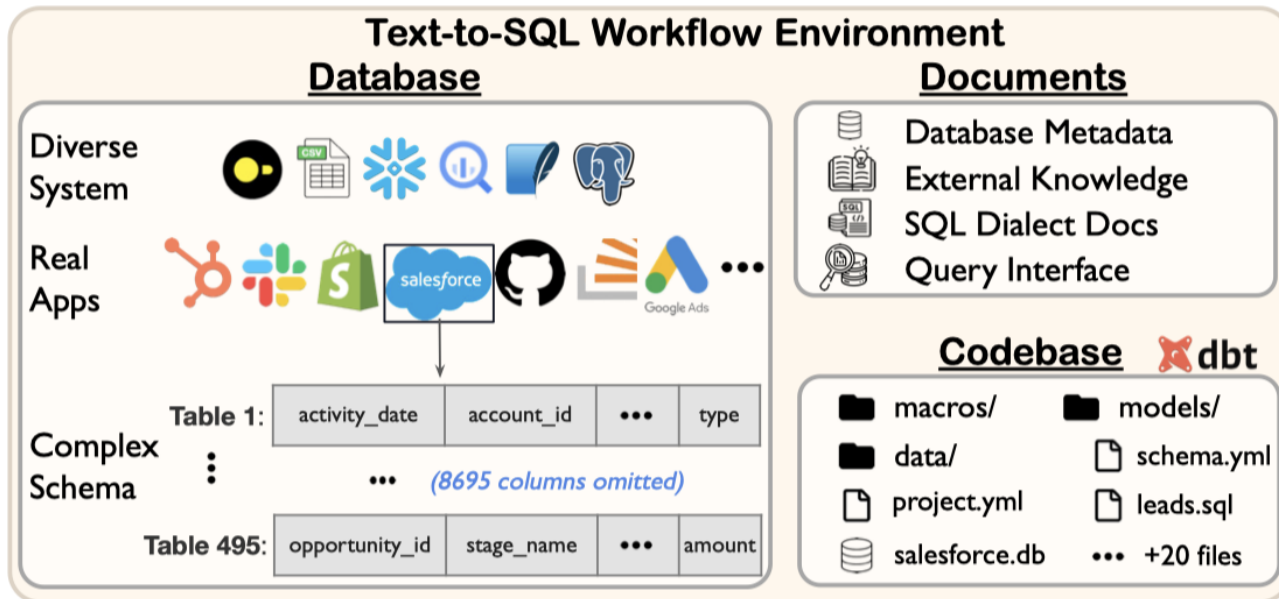
Enterprise Text2SQL

- ❑ Enterprise Text2SQL
 - Industrial applications (e.g. Google Analytics and Salesforce)
 - Include various DB systems (SQLite, DuckDB, BigQuery, Snowflake)
 - Project codebases along with documents.
- ❑ Requires LLMs to interact with environment



Enterprise Text2SQL

Q: I need a daily report on key sales activities—covering tasks completed, events held, leads generated, and the status of opportunities.



Language Models

SQL

(e.g. Google/Snowflake SQLs ...)

```
WITH opportunity AS (
  SELECT {{ dbt.date_trunc('day','close_date') }}
         AS close_date,
  CASE
    WHEN is_won THEN 'Won'
    WHEN NOT is_won AND is_closed THEN 'Lost'
    WHEN NOT is_closed AND LOWER(forecast_category)
    IN ('pipeline','best case') THEN 'Pipeline'
  END AS status
  FROM {{ var('opportunity') }}
),
... [+ 100 lines omitted]
LEFT JOIN event on ds.date_day =
salesforce_event.activity_date
LEFT JOIN opportunities_created
ON ds.date_day = opportunities_created.created_date
LEFT JOIN opportunities_closed
ON ds.date_day = opportunities_closed.close_date
```

SQLs/
Python

Exec
feedback

Future work: Enterprise Text2SQL

- A case: LLM as agent to interact with environment to solve the task.

Action 0: Agent's action to execute "ls" cmd in bash to show files.

Action 2/3: show DB and tables

Action 4: execute sql but meet the error

Observation-4: error information

Thought-4: Agent starts to thinking and try to solve the bug

.....

Final correct SQL

Finish the task

Success case 1: about patents analysis

17 Steps

Action 0: Bash(code="ls")

Action 1: Bash(code="cat README.md")

Action 2: BQ_GET_TABLES(database_name="patents-public-data", dataset_name="patents", save_path="dataset_metadata.csv")

Action 3: BQ_GET_TABLE_INFO(database_name="patents-public-data", dataset_name="patents", table="publications", save_path="publication_info.csv")

Action 4: BQ_EXEC_SQL(sql_query=..., is_save=False)

Observation-4: Error occurred while fetching data: 400 No matching signature for function DATE Argument types: INT64 Signature: DATE(TIMESTAMP, [STRING]) Argument 1: Unable to coerce type INT64 to expected type TIMESTAMP

Thought-4: The error indicates that I'm trying to use the DATE function on an INT64 field, which is invalid. The date fields in the dataset are stored as integers in YYYYMMDD format. I need to convert these integer date fields to DATE using the PARSE_DATE function. I'll update my SQL query to properly parse the dates and include the condition for published applications.

.....

Action-14: BQ_EXEC_SQL(sql_query=..., is_save="/workspace/result.csv")

Observation-14: Error occurred while fetching data: 400 No matching signature for function DATE Argument types: INT64 Signature: DATE(TIMESTAMP, [STRING]) Argument 1: Unable to coerce type INT64 to expected type TIMESTAMP

Action-15: BQ_EXEC_SQL(sql_query=..., is_save="/workspace/result.csv")

Observation-15: Now that the results are saved to /workspace/result.csv.

Action-16: Terminate(output="result.csv")

Software Engineering Coding for Data Tasks

Bird-Critic 1.5: Benchmarking SWE Agents on Data Management Tasks



Diverse Ecosystem Landscape & Coverage

<p>1. Frappe (Python)</p> <p>Vertical Domain Focus 12+ Vertical Applications</p> <p>ERP CRM HR EDU</p>	<p>2. Sylius (PHP)</p> <p>E-commerce Framework Specialized for custom and scalable online stores.</p> <p>Core Plugins</p>	<p>3. Strapi (Node.js)</p> <p>Headless Content Management Manages content for any platform or device.</p> <p>Core API Plugins</p>	<p>4. Discourse (Ruby)</p> <p>Community Forum Platform Modern software for online discussions.</p> <p>Core Plugins</p>
--	---	---	--

Multimodal Data + SQL

🍊 Data-centric agent for multimodal reasoning (SQL queries + image understanding + tool calling)



User Question

Airbnb is running a **promotion**: for all the listings shown in the table, any nights beyond the first 7 will be charged at a 15% discount. A user has requested accommodation for their trip from **08/10/2025 to 08/24/2025**. Their group consists of **two couples and a mutual friend**, and they have also provided a **reference photo of their preferred room**. Based on their preferences, select the listing that best matches their expectations and calculate the total cost of their stay.



Options

- A. Coral Wave Residence, 5687.0
- B. Vista Azul Penthouse, 6050.0
- C. Coastal Horizon House, 5887.0
- D. Ocean Breeze Villa, 5808.0
- E. None of the above.



Database

```
CREATE TABLE airbnb (  
  property_id INTEGER PRIMARY KEY, -- Unique ID for the property  
  name TEXT NOT NULL, -- Property name  
  location TEXT NOT NULL, -- City / Country  
  view TEXT NOT NULL, -- e.g., 270° Ocean View, Bay Sunset View, ...  
  style TEXT NOT NULL, -- e.g., Modern Minimalist, Scandinavian Cozy, ...  
  rooms_baths TEXT NOT NULL, -- Number of Bedrooms/ Number of Bathrooms  
  cooking_allowed TEXT NOT NULL, -- Yes / No / Limited  
  kitchen_type TEXT NOT NULL, -- None, Indoor Open Kitchen, Outdoor Kitchen,...  
  special_features TEXT, -- Notable highlights  
  capacity INTEGER NOT NULL, -- Max number of guests  
  price_per_night REAL NOT NULL -- Price per night in USD  
);
```

property_id	name	view	style	...	price_per_night
1	Bayfront Escape	Bay Sunset View	Scandinavian Cozy	...	410
2	Ocean Breeze Villa	270° Ocean View	Modern Minimalist	...	480
...
17	Harbor Light Loft	Waterfront + Skyline	Modern Minimalist	...	390

Scientific Data + SQL

🍊 Data-centric agent for multimodal reasoning (SQL queries + image understanding + tool calling)

User Questions

Please tell me the **proportion** of **graphene samples** observed between **March and May** that show the **defects depicted in the figure**, relative to the total number of graphene samples observed in that period.

🗣️ 🖼️ 🔍

Try Visual Search

Search with a picture instead of text

🖼️ Drag one or more images here or [browse](#)

OR

📄 Paste image or URL

📷 Take photo

Input Figures

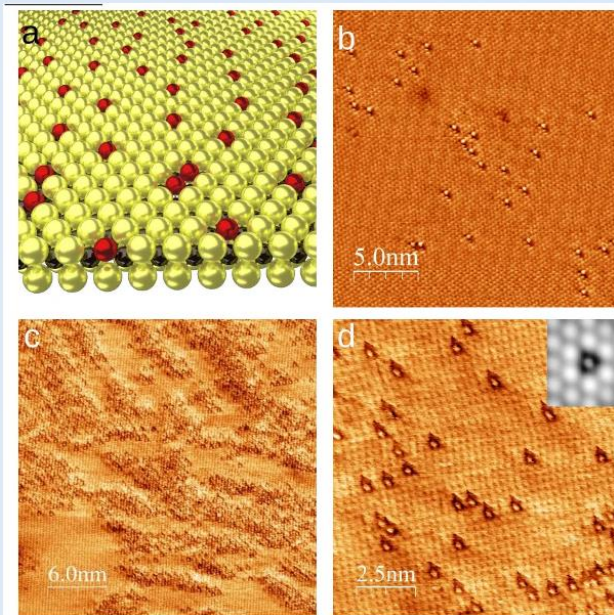
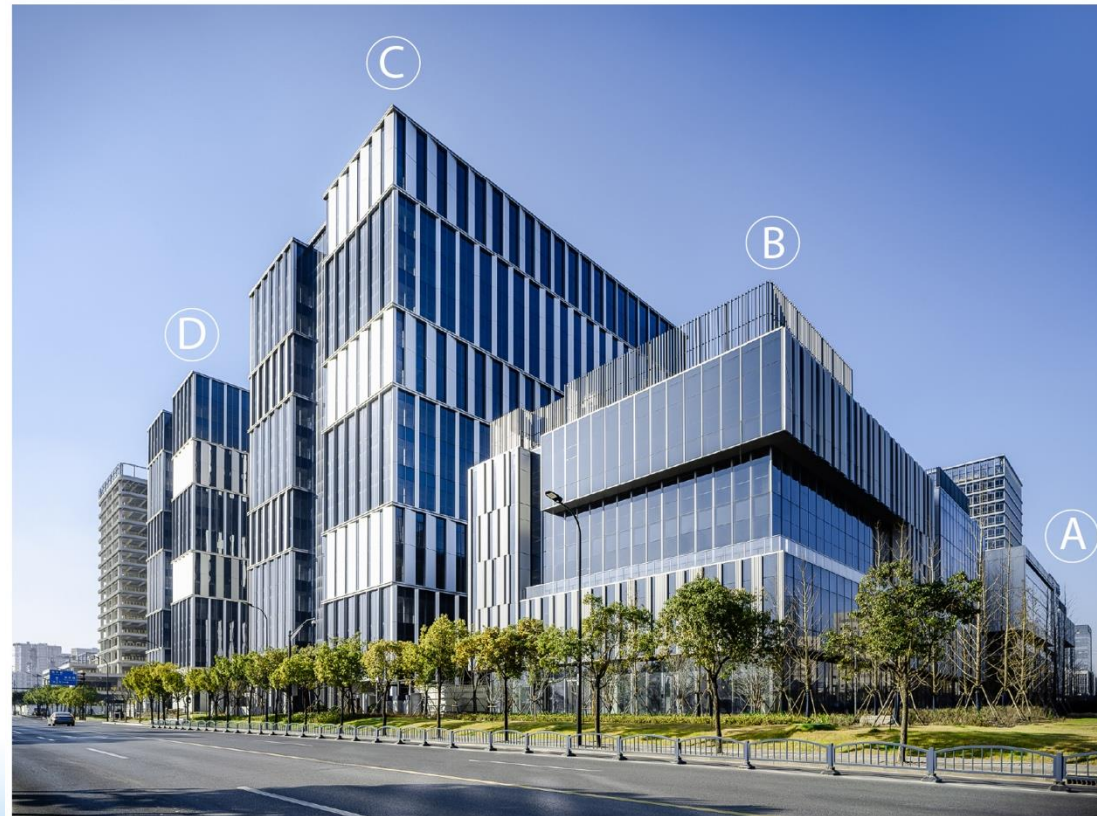
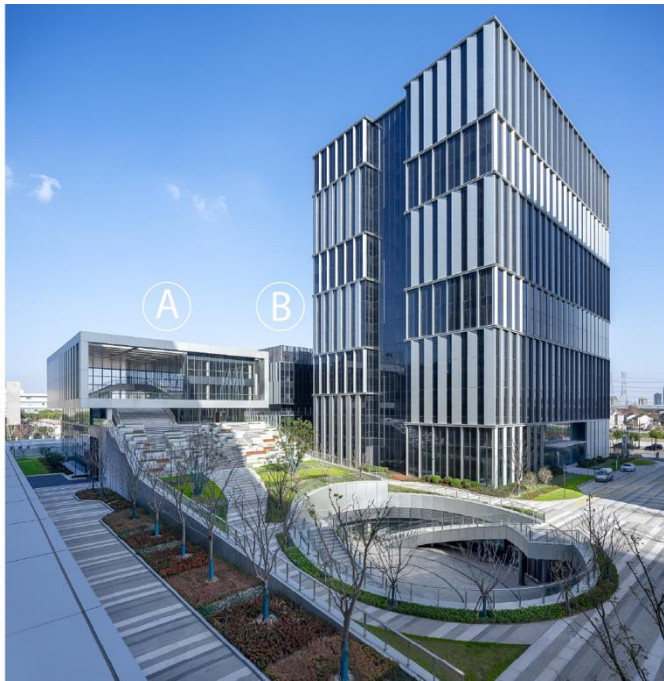


Figure Explanations

Scanning Tunneling Microscopy (STM) figure is used to visualize the surface of materials at the atomic level. The types observed in the STM figure above includes **vacancy** and **substitution**.

HKU-CDS Shanghai T&R Site

Zhangjiang Hi-Tech Park in the Pudong New Area of Shanghai



- A: Auditorium
- B: Student Amenities & Learning Commons
- C: CDS Main Building
- D: CDS Tech Landmark

WE ARE HIRING!

We invite outstanding scholars of all academic ranks and all research fields related to CS/AI and Data Science to join our fast-growing school.



TENURE-TRACK ASSISTANT PROFESSOR (AI + X)

Application Deadline: 31 May 2026

Requirements:

- PhD in CS/AI/Data Science or related field
- Excellence in research/teaching
- Strong **cross-disciplinary** research record

APPLY NOW



TENURE-TRACK PROFESSOR/ ASSOCIATE PROFESSOR/ ASSISTANT PROFESSOR

Application Deadline: 30 June 2026

Requirements:

- PhD in CS/AI/Data Science or related field
- Excellence in research/teaching
- Strong research record
- Academic leadership and international recognition for Professor level

APPLY NOW

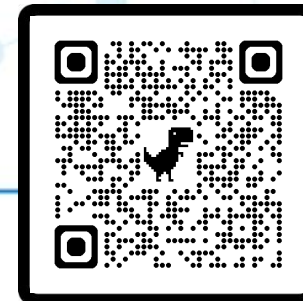


WHAT WE OFFER

- Highly competitive salary and start-up package
- Hong Kong salary tax capped at 15%
- Gratuity and retirement benefits of up to 15% of salary
- Housing benefits



The Bird Team



Bird 2025 Homepage



Reynold Cheng



Binyuan Hui



Chenhao Ma

Expert Panel



Jinyang Li



Ge Qu



Nan Huo



Xiaohan Xu



Xiaolong Li



Bowen Qin



Shipei Lin



Shuzheng Si



Ziwei Tang

Core Members



Per Jacobsson



Fatma Ozcan



Yannis
Papakonstantinou

Google Cloud
Collaborators

