

Natural language to ADQL

Can small open models make archive queries accessible?

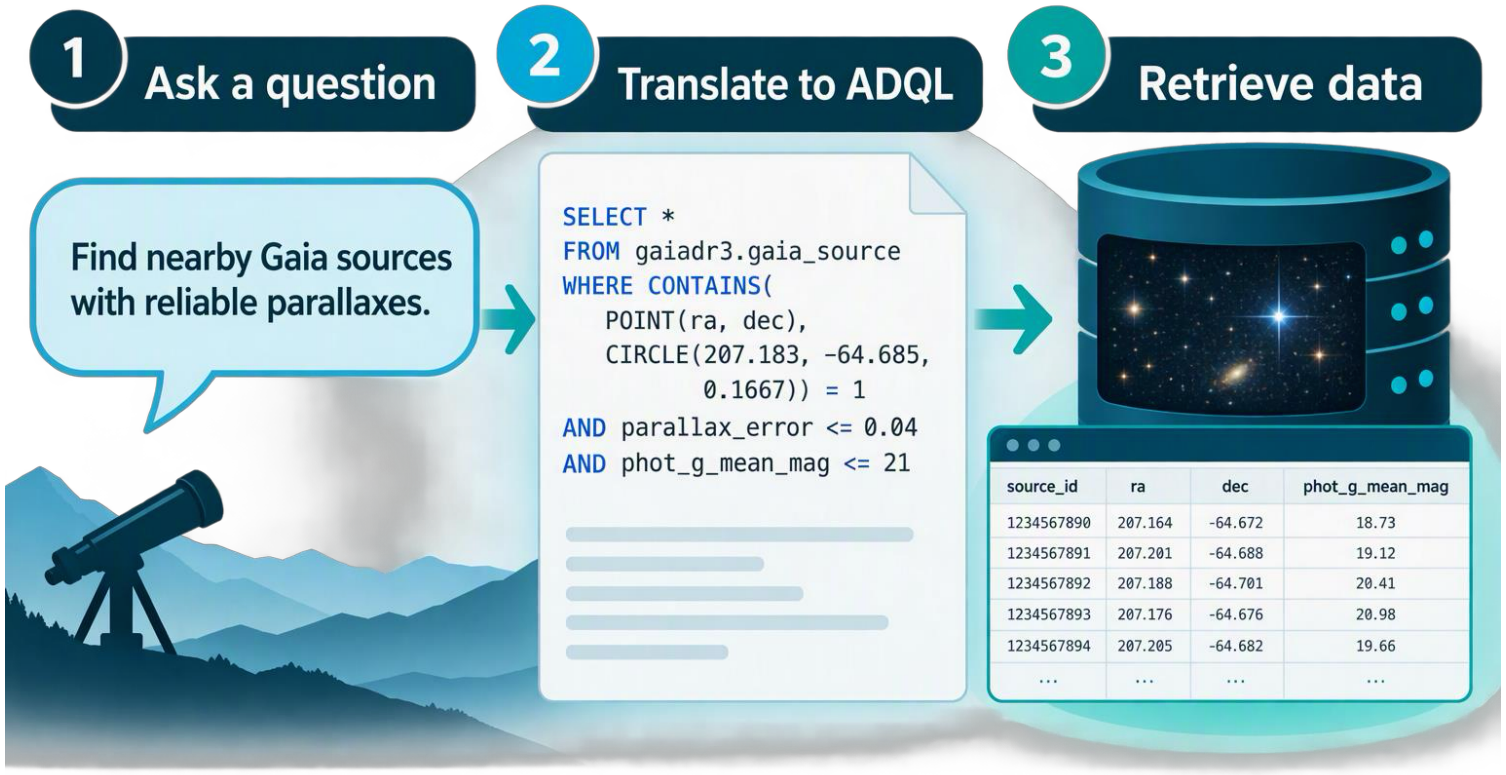
Roman Macháček (University of Bern), **Sandor Kruk (ESA/ESAC)**

AI as a tool for data discovery and data management · ADASS XXXVI · Perth · November 2026

EUROPEAN SPACE AGENCY

Why can’t we just ask the archive a question?

Archives need exact ADQL, but users think in questions



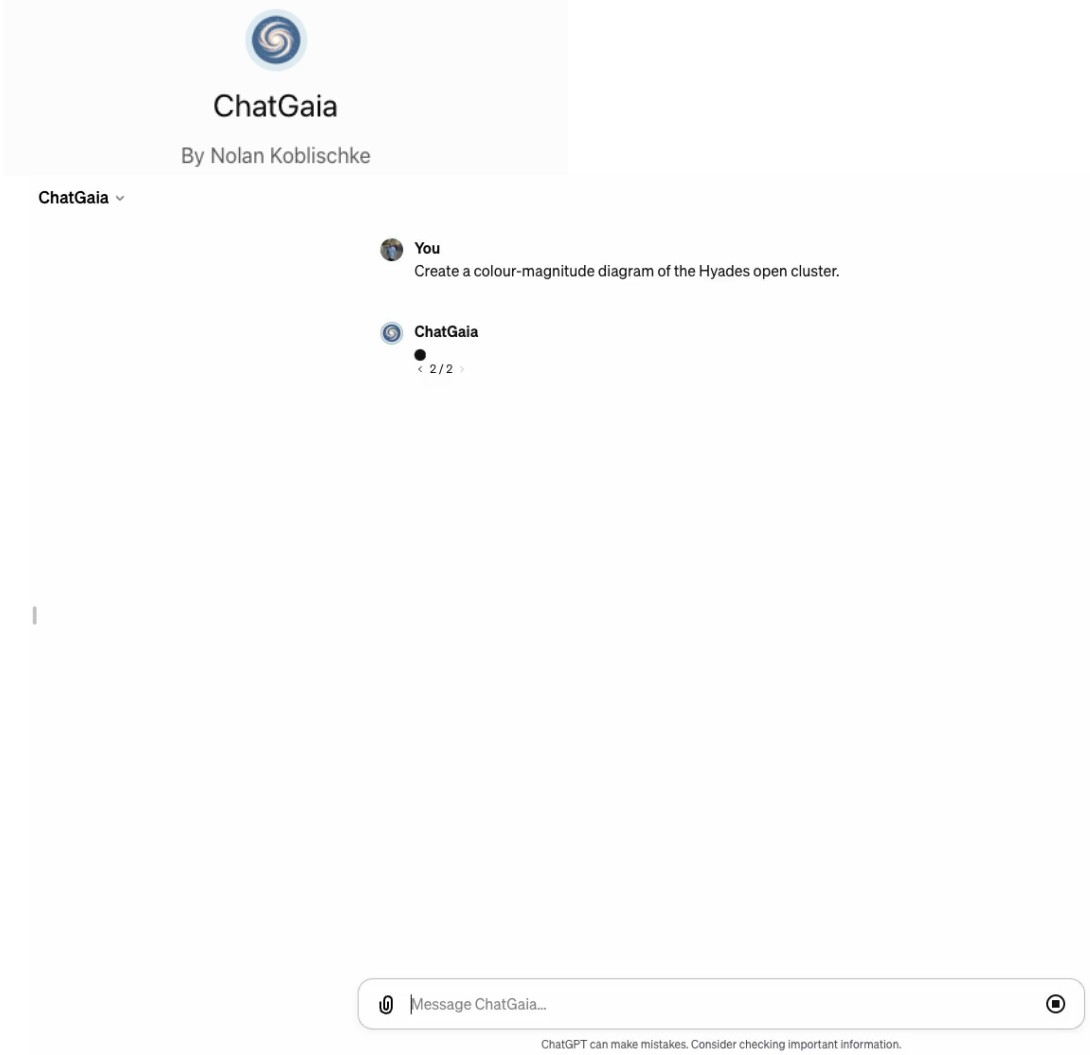
A translation problem

Users think in questions. Archives need exact instructions:

1. what to select
2. where on the sky
3. which quality limits
4. which of hundreds of tables

Hardest for students and newcomers

ChatGaia: showed that Large Language Models (LLMs) can already write ADQL



A community-built custom GPT on ChatGPT that turns questions into Gaia queries and plots. Impressive results, but:

Closed & cloud-hosted

Queries and data is not private

Paid per user

To be paid by the user through subscription

Not measured

No benchmark: how often is it right?

Hallucinations

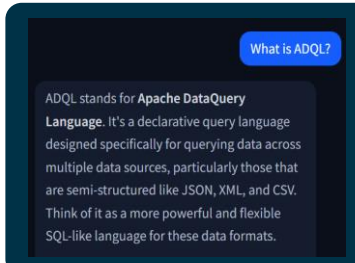
Invented tables and columns

Question: can small, open models do this?

Small language model (SLM): the same technology as ChatGPT (LLMs), but with a few billion (<15B) parameters instead of hundreds of billion and having open weights you can download, run and retrain yourself.

Model size and the hardware it needs

Llama 3.1 8B	≈16 GB · one GPU
Gemma 4 12B	≈24 GB · one GPU
Llama 3.1 405B	≈810 GB · a multi-GPU server



Out of the box, **Gemma-4 12B calls ADQL "Apache DataQuery Language".**

The four open models we compare

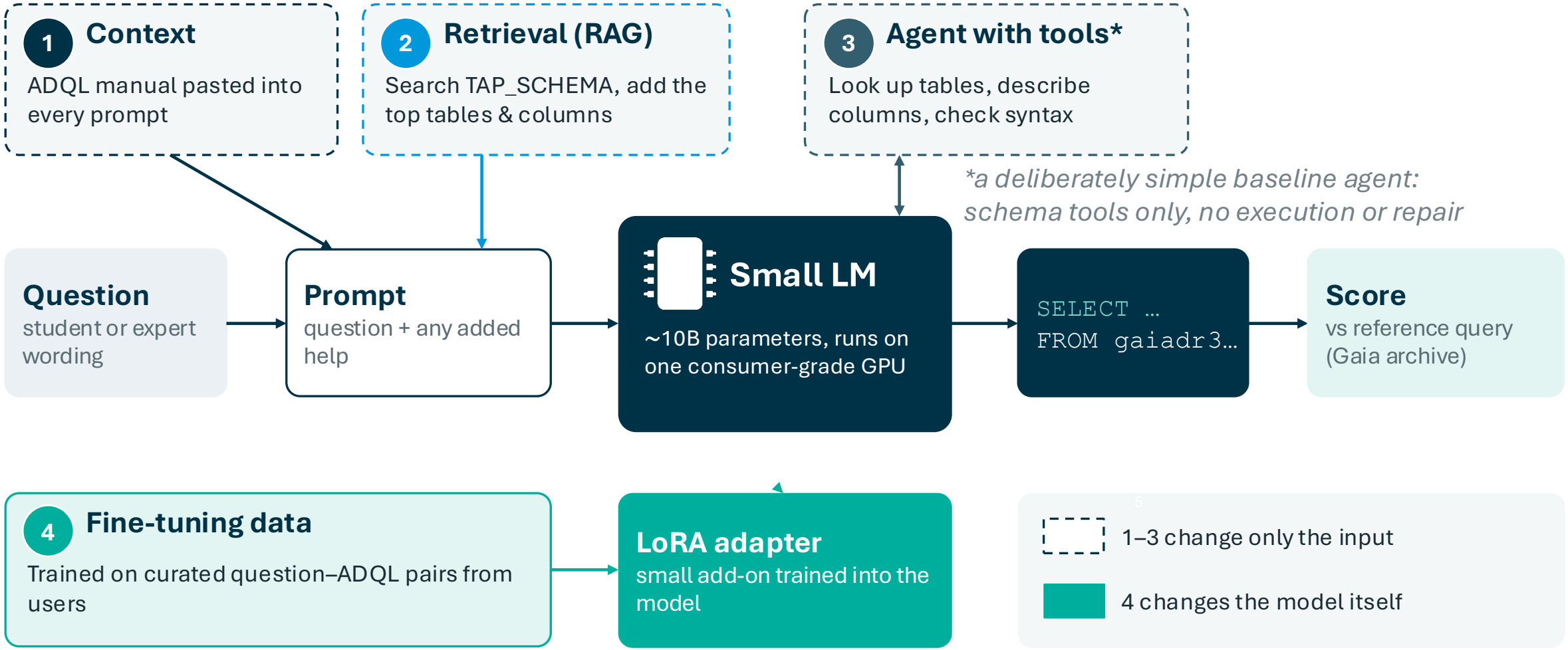
12B — **Gemma 4 12B**
Google DeepMind

9B — **Qwen 3.5 9B**
Alibaba Qwen

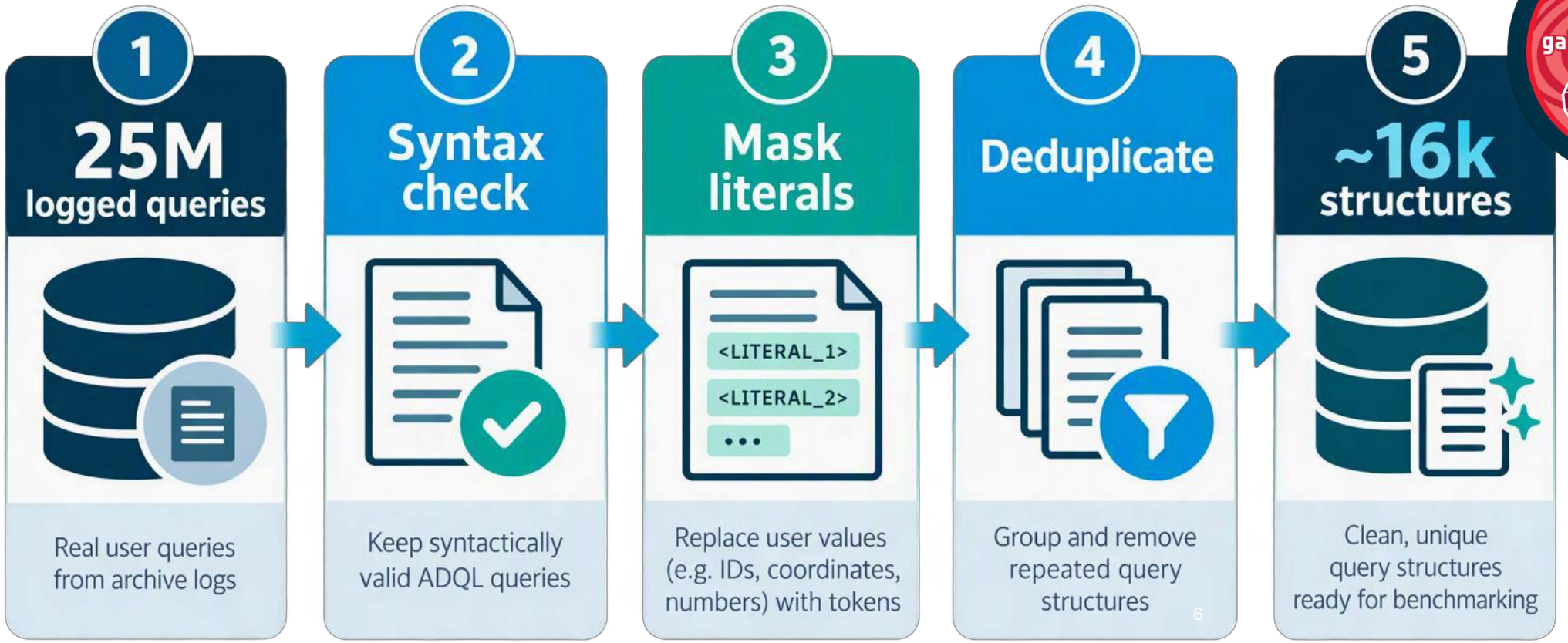
8B — **Llama 3.1 8B**
Meta

8B — **Ministral 3 8B**
Mistral AI

Should a model memorise the domain, retrieve it, or interact with it?



Data: real query logs are not yet training data



Values are masked (IDs, coordinates, thresholds → placeholders): for **grouping queries that share a structure**, also protects users searches.

Benchmark: LLM-written questions, validated by humans

Two levels of description detail (student, expert) and four levels of query complexity (Q1 lookup → Q4 multi-join).

1 LLMs describe each query

Gemma 4 31B (chosen) · Qwen 3.5 27B · GLM 4.7 Flash · Nemotron 3 Nano 30B

2 LLM judges filter

Natural Language – ADQL pairs scored 1–5;
kept if semantic alignment ≥ 4

3 Human experts validate

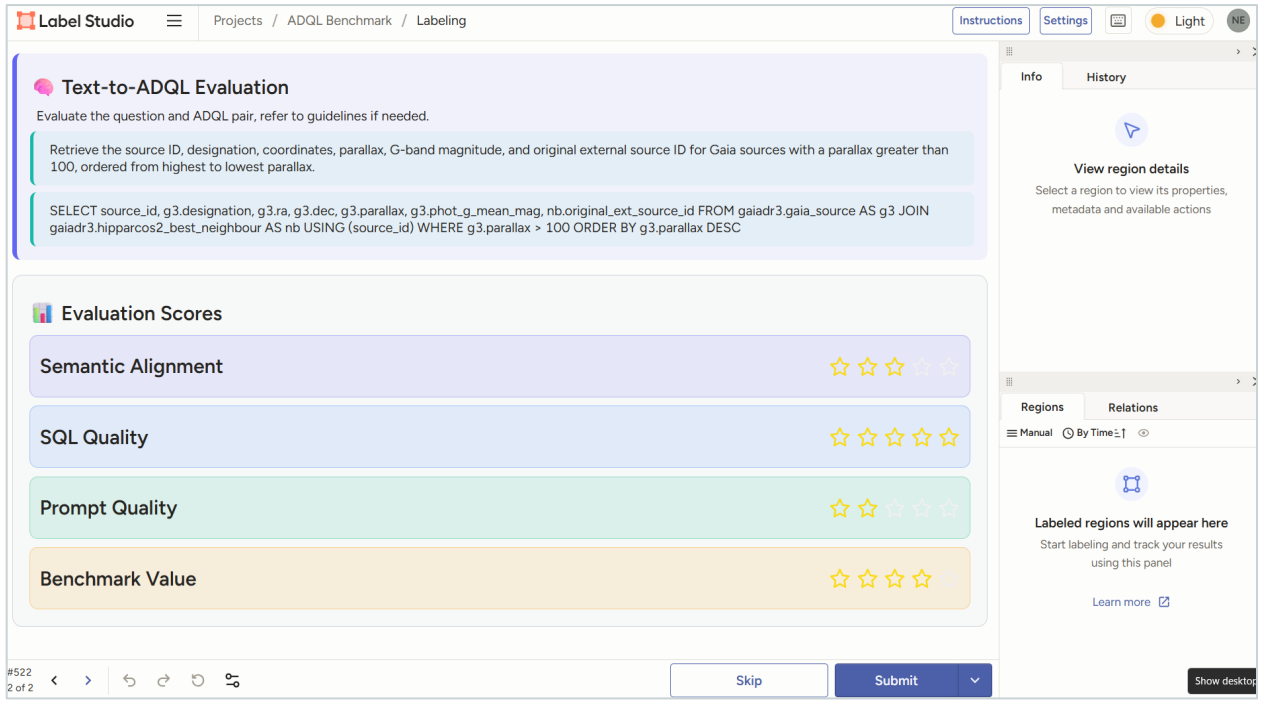
Experts rated the pairs based on: semantic alignment, SQL, prompt quality and benchmark value→ dataset of 1000 pairs

Student-style

"Find right ascension and declination of Gaia sources brighter than magnitude 15."

Expert-style

"Retrieve source id, ra, and dec from the Gaia DR3 source catalogue where mean photometric G magitude < 15."



Annotation interface. Question–ADQL pair rated 1–5 for semantic alignment, ADQL quality, prompt quality and benchmark value.

Metrics: three metrics used for every query

1 Syntax validity

Does it use correct grammar?

Parses as valid ADQL query.

2 Accuracy

Does it match the reference?

Also accepts harmless rewrites: aliases, formatting, order of conditions.

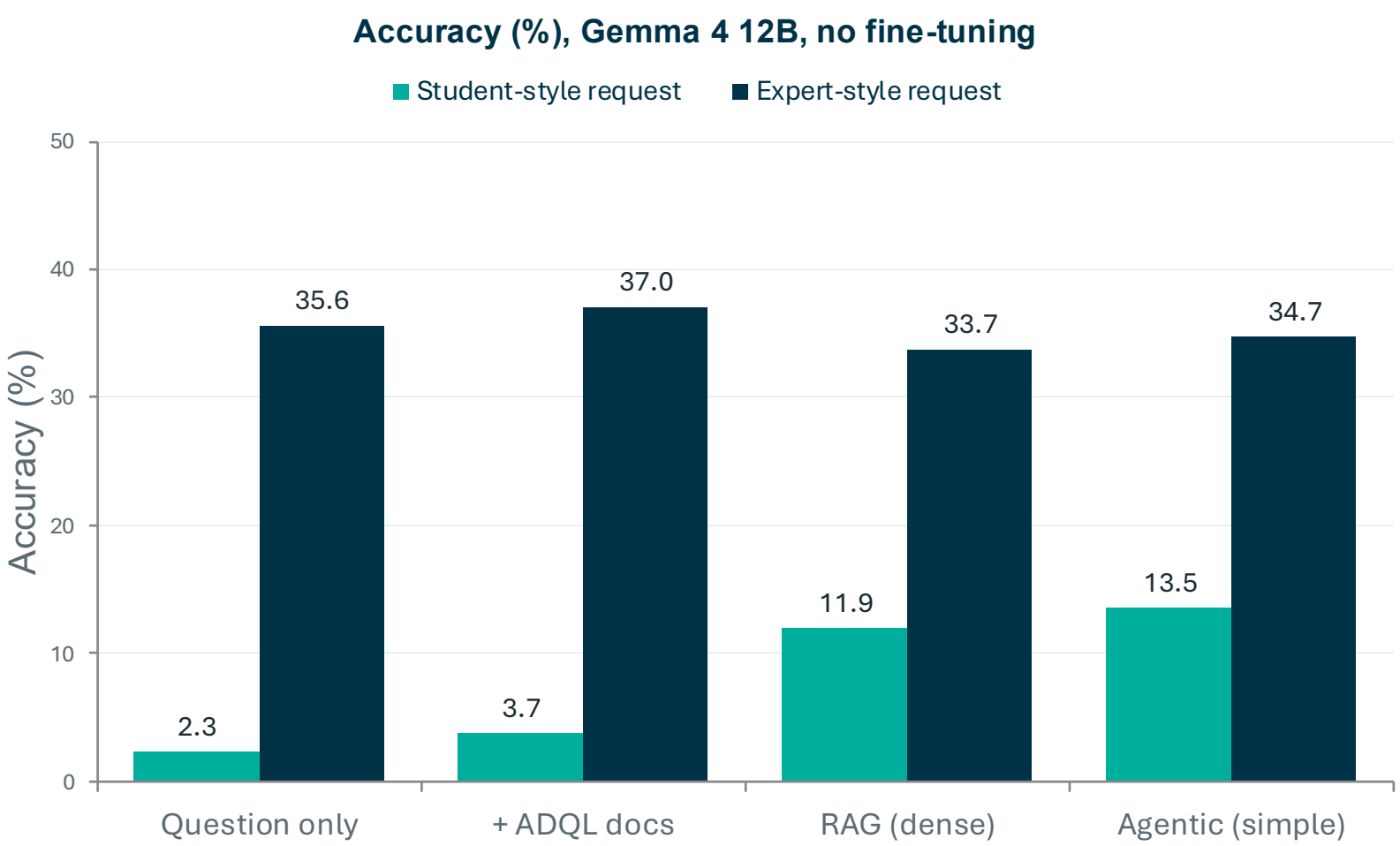
3 Executability

Does it run on the archive?

Executed on the Gaia TAP service without error.

Caveat: Syntax validity and accuracy scores are not enough; execution is important. And even then, a query that runs and returns a correct number of columns and/or rows can be wrong.

Result: help at question time cannot replace a good description



+33 pts

describing the table and columns (2% student → 36% expert)

Best question-time help

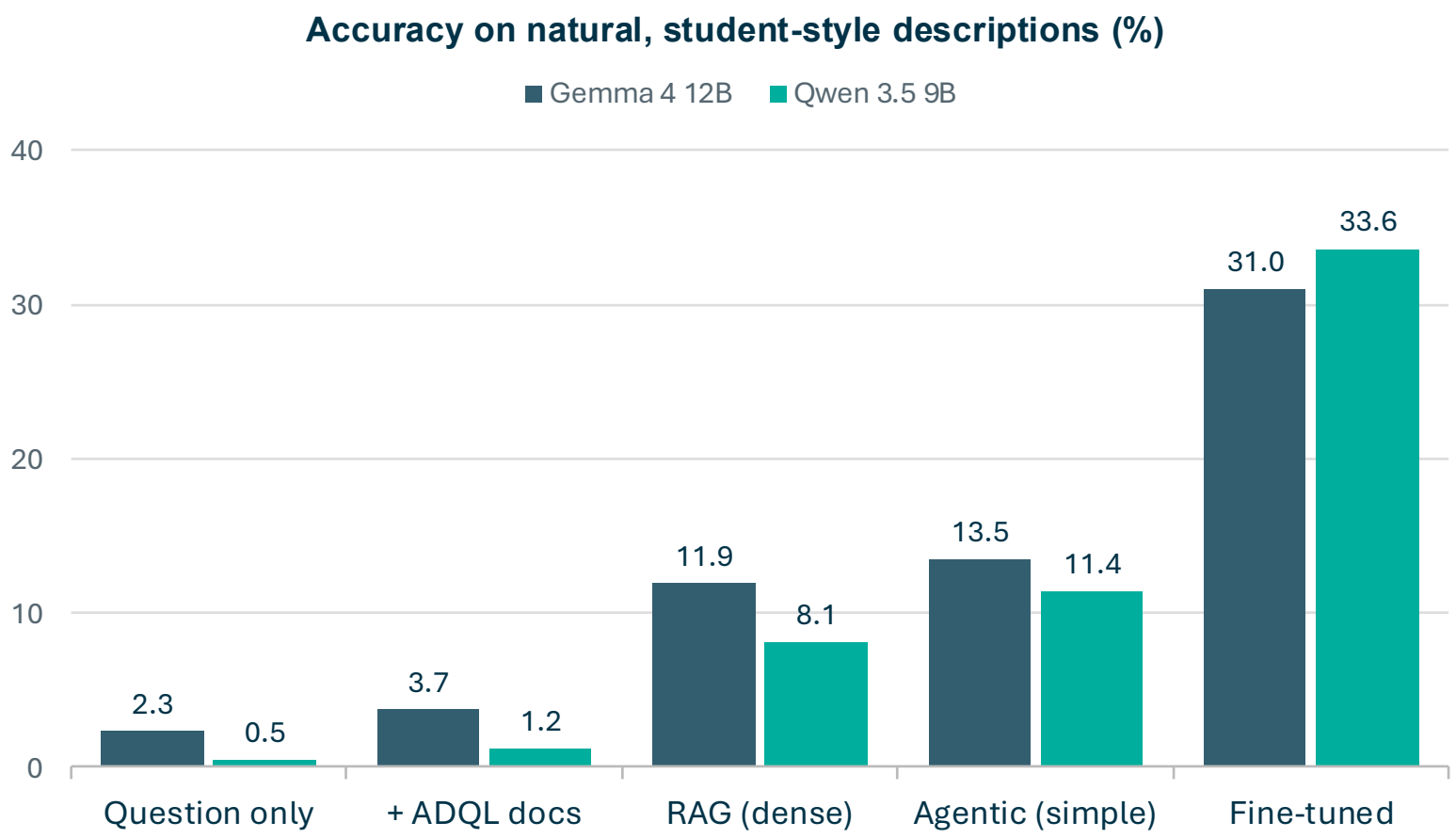
+10 pts

Retrieval and agent tools on student request

Retrieval finds tables, but less often columns

Table	**92%**
Columns	**34%**

Result: fine-tuning gives by far the largest improvement



~2% → ~34%

Student-style requests after fine-tuning

Expert requests

36% → 67%

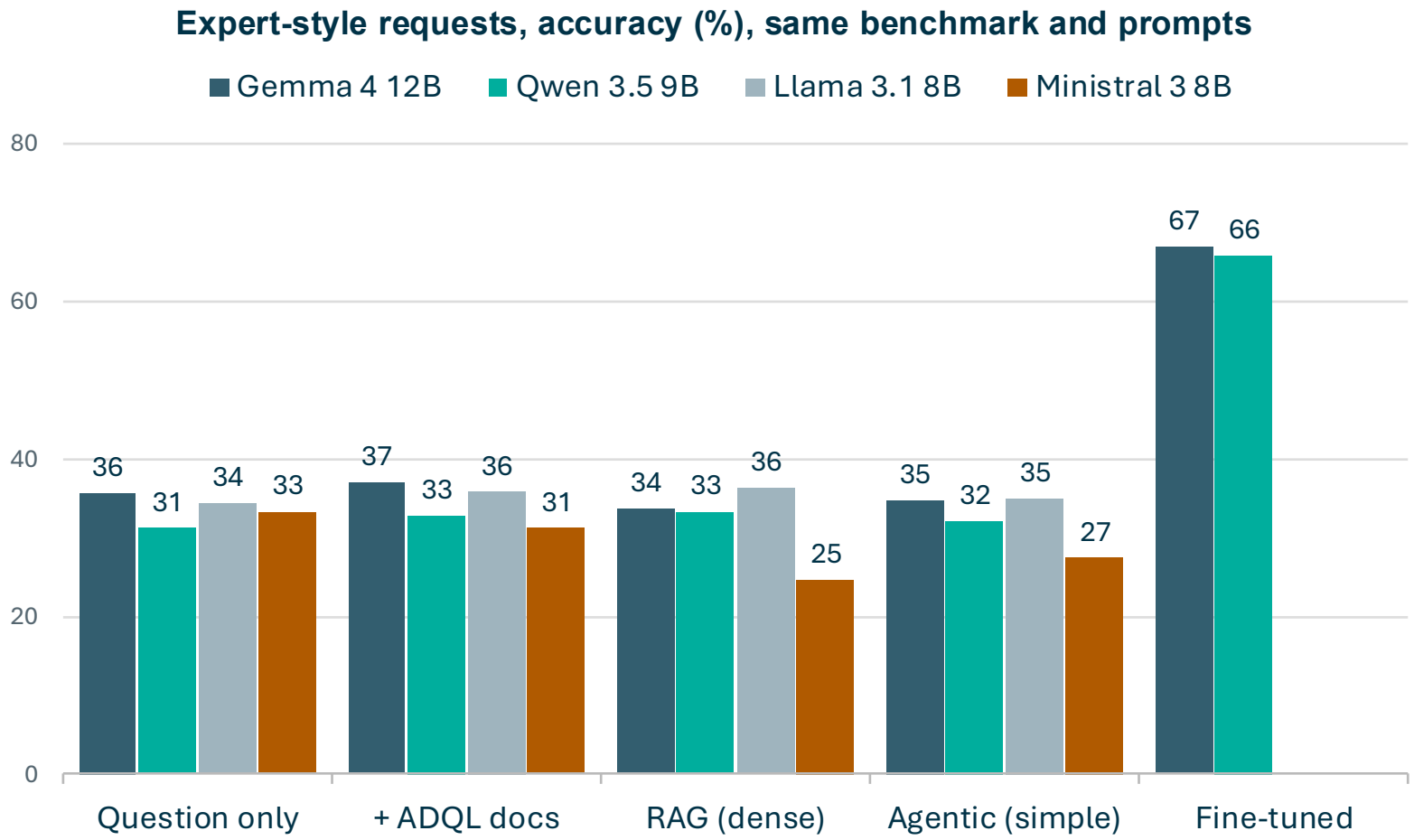
Gemma 4 12B; Qwen 3.5 9B: 31% → 66%

Executability on Gaia TAP

3% → 73%

of generated queries run without error after fine-tuning. For Gemma, student requests; Qwen 1% → 69%

Result: (small) model choice matters far less than the method



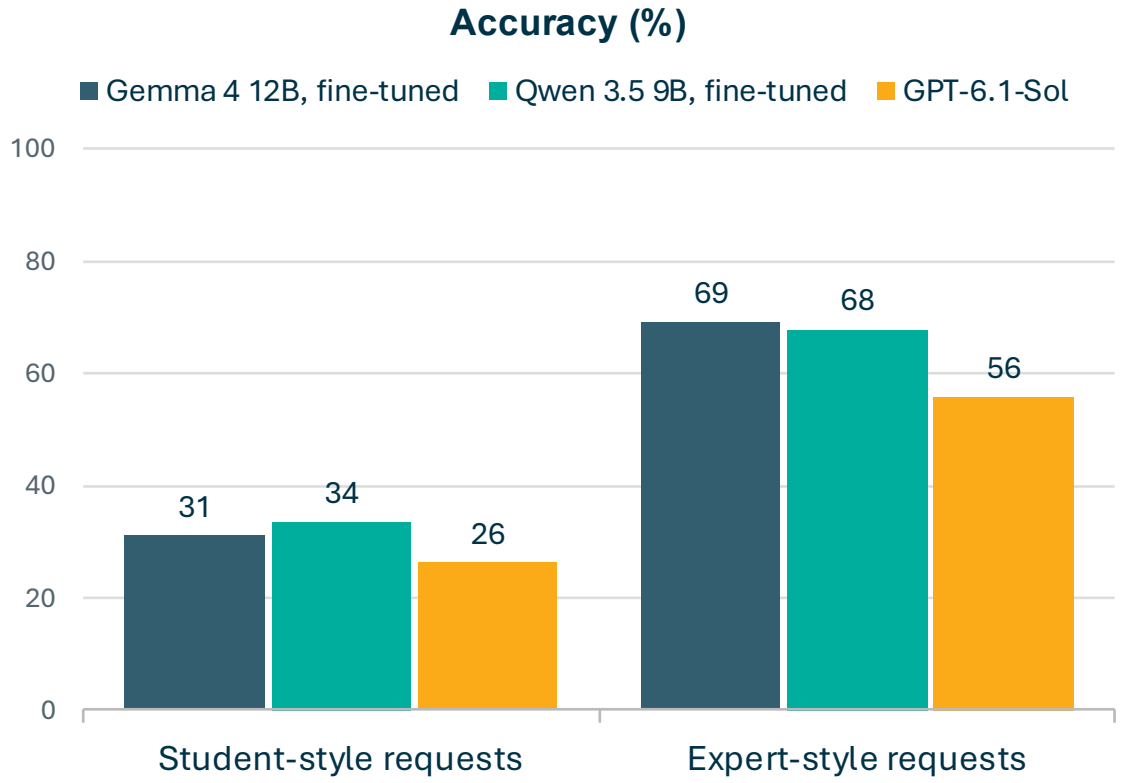
≤ 6 pts

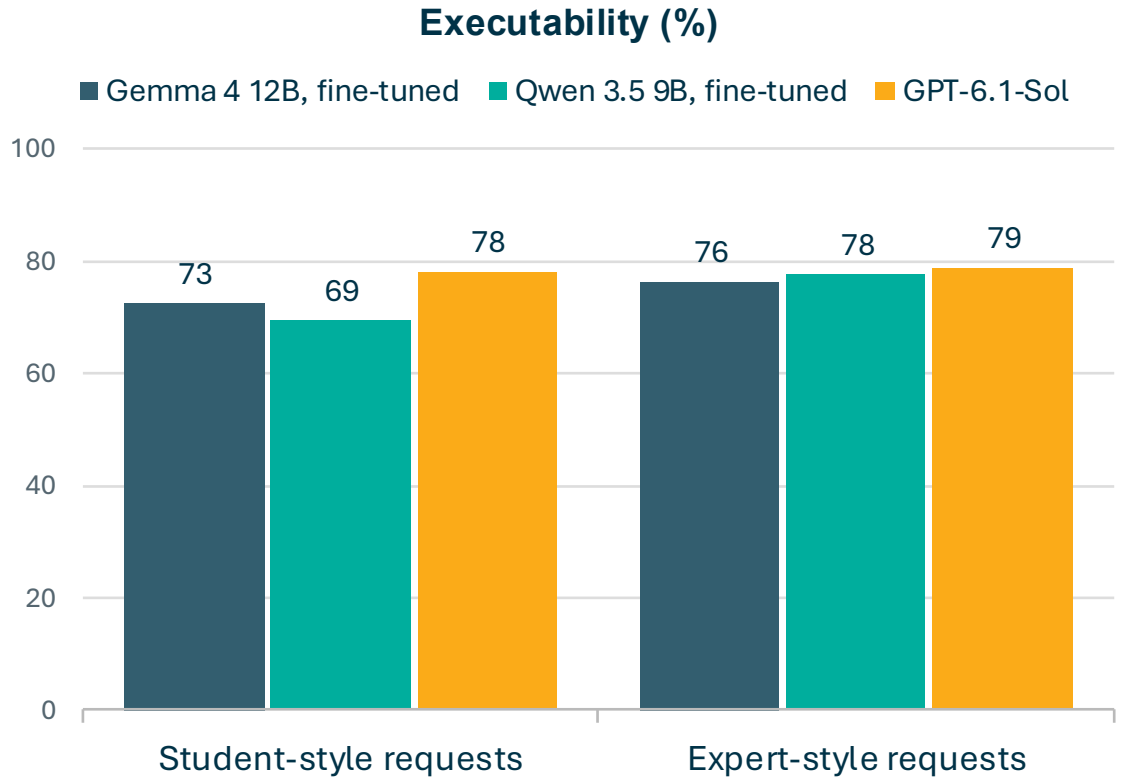
between the four models with the same prompt (expert 31–37%)

+31–35 pts

from fine-tuning

Result: fine-tuned small models vs frontier GPT model*





More accurate: fine-tuned 9–12B models match the reference more often than GPT-6.1-Sol (expert 68–69% vs 56%).

Slightly less runnable: GPT's queries run more often (78% vs 69–78%).
Syntax validity: fine-tuned models parse 99–100%, so the gap is schema or server errors, not grammar.

**GPT-Sol 6.1 model used for comparison, in inference mode only. Could not compare against ChatGaia as no API was available.*

What real queries look like: and how often we get them right

Q1 Simple table lookup

One row from the Hubble Source Catalogue.

```
SELECT TOP 1 *
FROM public.hubble_sc
```

Fine-tuned: **student 57–61%** · **expert 97%**

Q2 Filtering and sorting

Sources with G magnitude < 6 by descending RA.

```
SELECT TOP 2000 *
FROM gaiadr3.gaia_source
WHERE phot_g_mean_mag < 6
ORDER BY ra DESC
```

Fine-tuned: **student 48–50%** · **expert 76–77%**

Q3 Joining two tables

Coordinates of Gaia sources with a two-body orbit solution.

```
SELECT g.source_id, g.ra, g.dec
FROM gaiadr3.gaia_source AS g
JOIN gaiadr3.nss_two_body_orbit AS n
  USING (source_id)
```

Fine-tuned: **student 16–19%** · **expert 54–55%**

Q4 Multiple joins and selections

Gaia + 2MASS photometry via two link tables, in a Galactic sky box

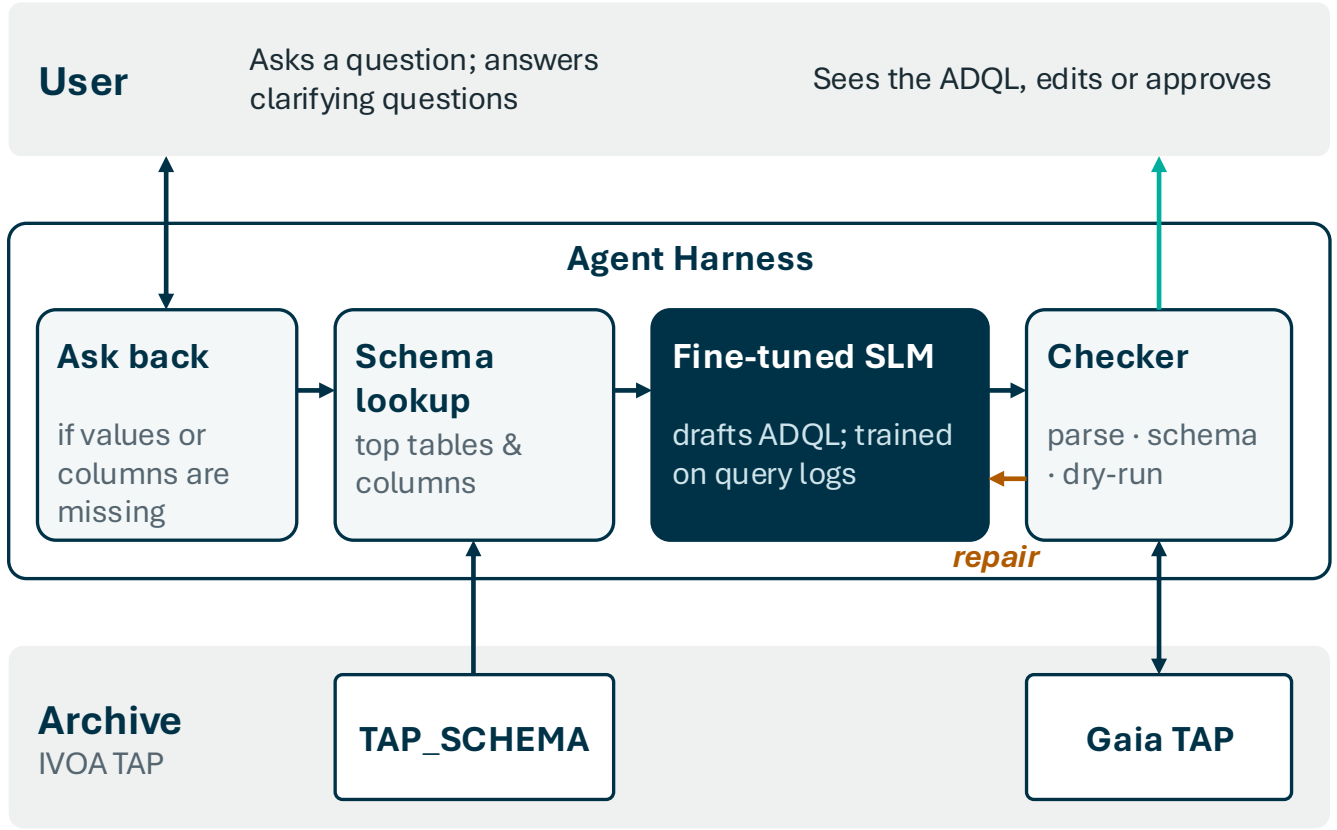
```
SELECT *, t.j_m, t.j_msigcom, t.h_m, t.h_msigcom, t.ks_m, t.ks_msigcom
FROM gaiadr3.gaia_source AS g
LEFT JOIN gaiadr3.tmass_psc_xsc_best_neighbour AS n USING (source_id)
LEFT JOIN gaiadr3.tmass_psc_xsc_join AS x USING (clean_tmass_psc_xsc_oid)
LEFT JOIN gaiadr1.tmass_original_valid AS t ON x.original_psc_source_id = t.designation
WHERE l > 277.465 AND l < 281.465 AND b > -33.672 AND b < -29.672
  AND phot_g_mean_mag < 18.0
```

Fine-tuned: **student 0–2%** · **expert 56–58%**

Proposition: fine-tuned SLM inside an agentic harness

Four parts, each fixing a problem we measured

1. **Language & style** → **Fine-tuning**
2. **Knowing the schema** → **Schema lookup**
3. **Catching mistakes** → **Checker + repair**
4. **Missing information** → **Ask back, show ADQL**



Part of wider ESAC effort: see similar use cases for other ESA missions, presented by **Marcos López-Caniego** and **Aitor Ibarra**

What we learnt

Conclusions

1. **Fine-tuning makes small models (<15B) work**
   Accuracy increases significantly after fine-tuning
2. **Method matters more than model**
   Within the same model class, performance is similar
3. **The natural language description matters**
   Missing values cap any model
4. **Retrieval finds tables, not columns**
   Right table retrieved 9 out of 10, right columns less than half
5. **Fine-tuned SLMs competitive with a frontier model** More accurate than GPT-6.1-Sol: 69% vs 56%

Lessons learnt

Columns are the hard part

- **Fine-tuning** learns real column use
- **Richer** TAP_SCHEMA descriptions help

Query logs are a training asset

Masked and deduplicated

Small models are deployable

8–12B parameters, one GPU

An assistant, not an autopilot

Ask back; show the ADQL

Find stars with good parallaxes around the Pleiades.

Two details first: which search radius?
Is “good parallax” parallax_over_error > 10?

EUROPEAN SPACE AGENCY

THANK YOU!

Sandor Kruk (ESA)

sandor.kruk@esa.int