ALCHEMY LOOP

INDEPENDENT VERIFICATION

AI LLMQUERY · TESTING & ACCURACY

Benchmarked against BIRD-SQL

Any vendor can demonstrate software on the data it was written for. We measure AI LLMQuery against BIRD-SQL, the cross-domain research benchmark for turning questions into correct SQL — testing on its development set of eleven independent databases with 120 expert-written questions and verified answers. We publish what it scores.

11 independent databases | **75 tables · 798 columns** | **120 expert questions** | **233 automated tests** | **BIRD-SQL · DAMO · HKU · UIUC**

WHAT YOUR REVIEWER WILL ASK

The questions behind the demonstration

- *How do I know this works on my database and not just your demo?*
- *What happens on the questions it gets wrong?*
- *What does my team have to do to make it better?*
- ***Will you run this against our schema before we commit?***
- *Have you actually tested it, or am I taking your word?*
- *Where does it still lose to someone writing SQL?*

72% correct
On databases we had never seen, once set up the way a firm sets it up. Exact answers, no allowances.

85% everyday
On the day-to-day questions that make up most of what a team actually asks.

2.7× gain
Improvement in the table relationships our engine maps, driven by this testing programme.

8 seconds
Average time to a complete answer, with the query it ran shown beneath it.

THE BENCHMARK

What BIRD-SQL is, and why we chose it

BIRD — *BIg bench for laRge-scale Database grounded text-to-SQL evaluation* — is the pioneering cross-domain dataset that examines how **real database contents** affect a system's ability to turn a question into correct SQL. It is built and maintained by researchers at Alibaba DAMO Academy, the University of Hong Kong and the University of Illinois Urbana-Champaign, and it is the standard by which serious text-to-SQL systems are judged.

It matters to a law firm for one reason. Earlier benchmarks used small, tidy, invented schemas. BIRD uses **large, messy, real ones** — inconsistent naming, undocumented codes, dirty values, relationships nobody declared. That is what a firm's case management system looks like after fifteen years, and it is the only fair test of software that claims to work on it.

12,751 pairs
Unique question and verified-SQL pairs, written and checked by domain experts.

95 databases
Large, genuine schemas rather than the small invented ones earlier benchmarks used.

33.4 GB
Of real content — because the difficulty lives in the data, not only in the schema.

37+ domains
Blockchain, hockey, healthcare, education and more. Nothing resembling foreclosure.

HOW IT UNDERSTANDS YOUR DATABASE

It maps how your tables relate — then proves it against your own rows

Case management systems almost never declare how their tables connect, so AI LLMQuery works it out and then *verifies* it against your data before relying on it. We measure that engine against eleven independent research schemas that publish the correct answer, and this testing programme has driven a **2.7× gain** in the relationships it maps — while introducing no relationship that was not real.

1 It finds what nobody wrote down
On these independent schemas it identifies working relationships the authors never declared at all — the connections a firm's own developers left implicit years ago and no documentation records.

2 It refuses what the data will not support
One database formally declares a relationship its own rows contradict; joining on it would have multiplied the figures. The schema said yes, the data said no, and we believed the data.

3 Measured continuously, not once
Every release is scored against all eleven databases. Improvements are proven by measurement before they ship, on schemas nobody here designed.

4 Reproducible by you
The harness ships with the product source. Point it at these databases, or at a copy of your own, and it prints the same tables we publish.

Out of the box, and once configured

- ✓ **46% on day one** — a database it has never seen, nothing configured, no preparation of any kind.
- ✓ **72% once set up** the way a firm sets it up: the database prepared, your terms written down, your columns explained.
- ✓ **85% of everyday questions**, which is the bulk of what a team asks between one month-end and the next.
- ✓ **Complex questions more than double** — the multi-table ones that used to need an analyst are where setup pays most.

What one hour of your time is worth

Settle what your terms mean, once
The single largest gain we measured did not come from the software. It came from supplying the firm's own definitions — what counts as an open file, which date is the deadline — exactly what the Fine-tuning session collects.

- ↑ **+26 points overall** against the same questions on the same databases, measured.
- ↑ **+30 points** on the analytical questions a partner asks at month-end.
- ↑ **More than double** on the hardest multi-table questions in the set.

Settled once, applied to every question afterwards — and recorded against the name of whoever decided it.

It does not replace your analyst. It replaces the queue in front of them.

Writing SQL by hand is exact, and it always will be. The question that decides anything is not accuracy — it is who is allowed to ask, and how long everybody else waits.

	WRITING SQL BY HAND	AI LLMQUERY
Who can ask	One or two people in the firm	**Anyone**
Time to an answer	Hours to days, in a queue	**8 seconds**
Cost per question	Analyst time	**About 2p**
Shows its working	If the analyst saves the query	**Every answer, always**
Questions never asked at all	Most of them	—

Every answer arrives with the query that produced it and the assumptions it rested on. The one person in your firm who reads SQL moves from **writing every query** to **checking the ones that matter**.

HOW THIS COMPARES

Where 60% sits on a benchmark this hard

BIRD is deliberately difficult — large real schemas, dirty values, undocumented codes. The numbers below are published by the benchmark's own maintainers, and they are the context that makes ours meaningful.

SYSTEM	EXECUTION ACCURACY	WHAT IT IS
Human experts	**92.96%**	Data engineers and database students, working by hand
Best research system	**81.95%**	A laboratory pipeline — agents, retrieval, many candidate queries, oracle knowledge. Not a product you can install.
AI LLMQuery, configured	**72%**	**One question, one query, no retries. Set up the way a firm sets it up.**
AI LLMQuery, out of the box	**46%**	Nothing configured at all, on a database never seen before
GPT-4, the published baseline	**46.35%**	The reference result the benchmark's authors published

Read the middle of that table carefully. The systems scoring above us are research pipelines: they generate many candidate queries, retrieve schema and values, run agents that retry and self-correct, and in places are handed knowledge in advance. AI LLMQuery writes *one query, once*, in eight seconds, on a workstation, with no data leaving the building — and configured, it clears the published GPT-4 baseline by twenty-six points, sitting within ten of a laboratory pipeline built for this benchmark alone.

Verified on every release

233 automated tests covering safety controls, query validation, access tiers and the answering pipeline end to end. Run twice — once against our source, once against the protected file you receive. Several exist because a control behaved differently in the shipped build.

Your data never leaves

The AI provider receives your table and column names and the question typed — never a value from any row. The query runs locally and the results stay on that machine. Columns you mark sensitive are withheld entirely, and questions needing one are refused rather than worked around.

The test that decides it

Every figure here is about somebody else's data. During evaluation we will run the same harness against a copy of your own schema and show you exactly what it found — before you commit to anything.

ALCHEMY LOOP

Measured 19 August 2026 against release 1.26.0 · BIRD-SQL development set (Mini-Dev), Alibaba DAMO Academy · University of Hong Kong · UIUC, CC BY-SA 4.0 · 120 questions stratified across 11 databases and 3 difficulty bands · Test harness available on request