

Text-to-SQL with an OKF Knowledge Wiki

Agent + OKF evaluated on the BIRD **mini_dev** benchmark (500 questions, 11 databases), graded with the official BIRD evaluator.

Model: Claude Opus 4.8 · Reasoning effort: xhigh · Grading: bird-bench evaluation_ex.py (unmodified) on SQLite

74.0

EXECUTION ACCURACY
(EX)

500 questions · SQLite set-
equality

77.5

SOFT-F1

lenient cell-level match

+11.0

EX VS BEST
LEADERBOARD ENTRY

TA + GPT-4o = 63.0

4.47

AVG WIKI READS /
QUESTION

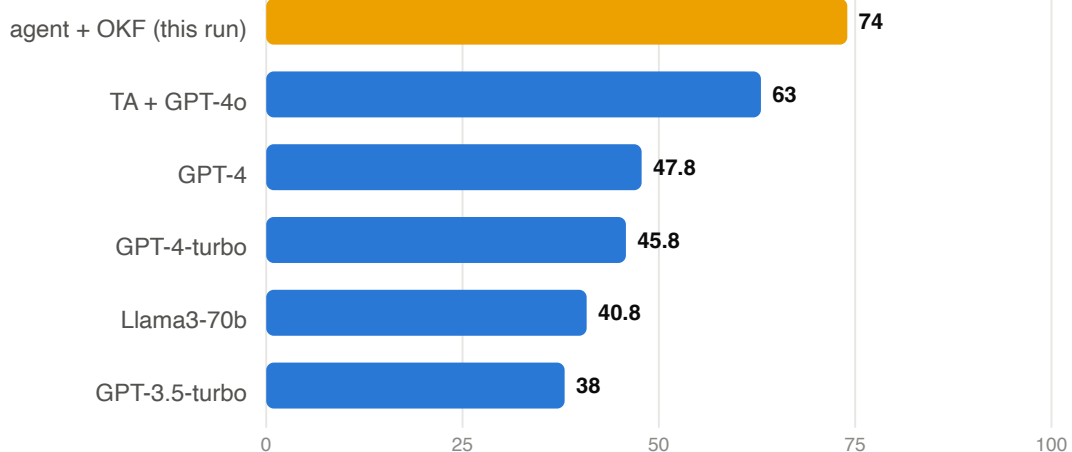
500 agents · 2,234 calls

1 · Executive summary

An autonomous agent answered all 500 BIRD mini_dev questions with **no access to the database schema**. Its only knowledge of each database came from an **Open Knowledge Format (OKF) bundle** — LLM-authored markdown docs — read on demand over an MCP server. Every other condition matches the official BIRD leaderboard (SQLite dialect, expert "evidence" hint provided, chain-of-thought enabled), and grading uses BIRD's own unmodified evaluator on the real SQLite databases.

The agent scored **EX = 74.0**, above every published mini_dev leaderboard entry (best prior: TA + GPT-4o at 63.0). Because it reconstructed all table, column, join and value semantics from the wiki alone — in about 4.47 reads per question — the result shows the OKF bundle preserves the schema knowledge required for text-to-SQL, with no measurable accuracy penalty for the substitution.

2 · Result vs published leaderboard



Execution Accuracy (EX), % — this run (highlighted) vs published BIRD mini_dev SQLite-EX entries. Same 500 questions, same grader, same evidence + CoT + SQLite conditions.

Comparability. Identical questions, grader, and prompt conditions as the leaderboard. Two things differ by design: the knowledge source (OKF bundle vs raw schema) and the base model. This is therefore "agent + OKF as a system," not a same-model ablation of the wiki in isolation.

3 · Methodology

3.1 Task & grading

BIRD mini_dev is 500 curated text-to-SQL questions over 11 relational databases. For each question the agent emits one SQLite SELECT. A prediction is **correct (EX = 1)** iff its result set equals the gold query's, as unordered row sets — computed by BIRD's own `evaluation_ex.py` (unmodified) against the real SQLite databases, 30 s/query timeout. We also report BIRD's Soft-F1 (lenient cell-level match).

3.2 Conditions — matched to the leaderboard

Question set	BIRD mini_dev, all 500
SQL dialect	SQLite (matches grader & leaderboard)
External Knowledge (evidence)	Provided — same as leaderboard (<code>use_knowledge=True</code>)
Chain-of-thought	Enabled — same as leaderboard (<code>cot=True</code>)
Knowledge source	OKF bundle only (the single deliberate swap vs the leaderboard's CREATE-TABLE schema)
Grader	bird-bench <code>evaluation_ex.py</code> / <code>evaluation_f1.py</code> , unmodified

3.3 LLM & agent configuration

Model	Claude Opus 4.8
Reasoning effort	xhigh
Agents	500 — one independent agent per question (no shared state)
Concurrency	~14 agents in parallel (pool cap)
Knowledge access	Live OKF consumption MCP server (streamable-HTTP, AgentCore)
Tools available	All 9: <code>list_domains</code> , <code>list_declared_domains</code> , <code>search_domains</code> , <code>list_directory</code> , <code>read_page</code> , <code>glob</code> , <code>grep</code> , <code>get_backlinks</code> , <code>semantic_search</code>
Run outcome	500/500 generated, 0 errors, ~20M tokens, ~51 min wall-clock

3.4 Agent task prompt (abridged)

Each agent received this instruction. The evidence hint is used (as on the leaderboard); the schema is not provided — the agent derives it from the OKF bundle over MCP. The gold SQL is held in a separate file the agent never opens.

You are a text-to-SQL expert. Produce ONE valid SQLite query that answers the question. You will be graded by execution on the real SQLite database (BIRD execution accuracy), so the query must run on SQLite and return exactly the right rows.

STEP 1 – read your assignment (question + "evidence" as External Knowledge, and the OKF dataset name). This file contains NO answer SQL – derive the query yourself.

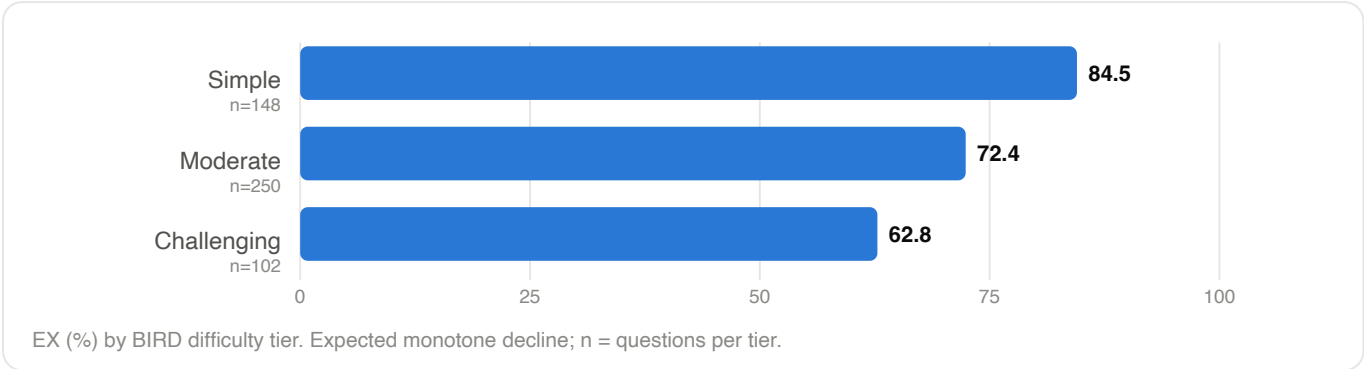
STEP 2 – learn the database SOLELY from its OKF knowledge bundle (the only knowledge source; NO raw schema is handed to you). Use the live MCP client with progressive disclosure: `list_directory` / `read_page` / `glob` / `grep` / `get_backlinks` / `semantic_search`. Read the dataset overview, the relevant tables/<t>.md (columns, types, keys, grain, value semantics, encodings), and references/ (joins, metrics, value formats). Do NOT query any database directly; do NOT read any gold/answer file.

STEP 3 – think step by step, then write ONE SQLite SELECT that answers the question.

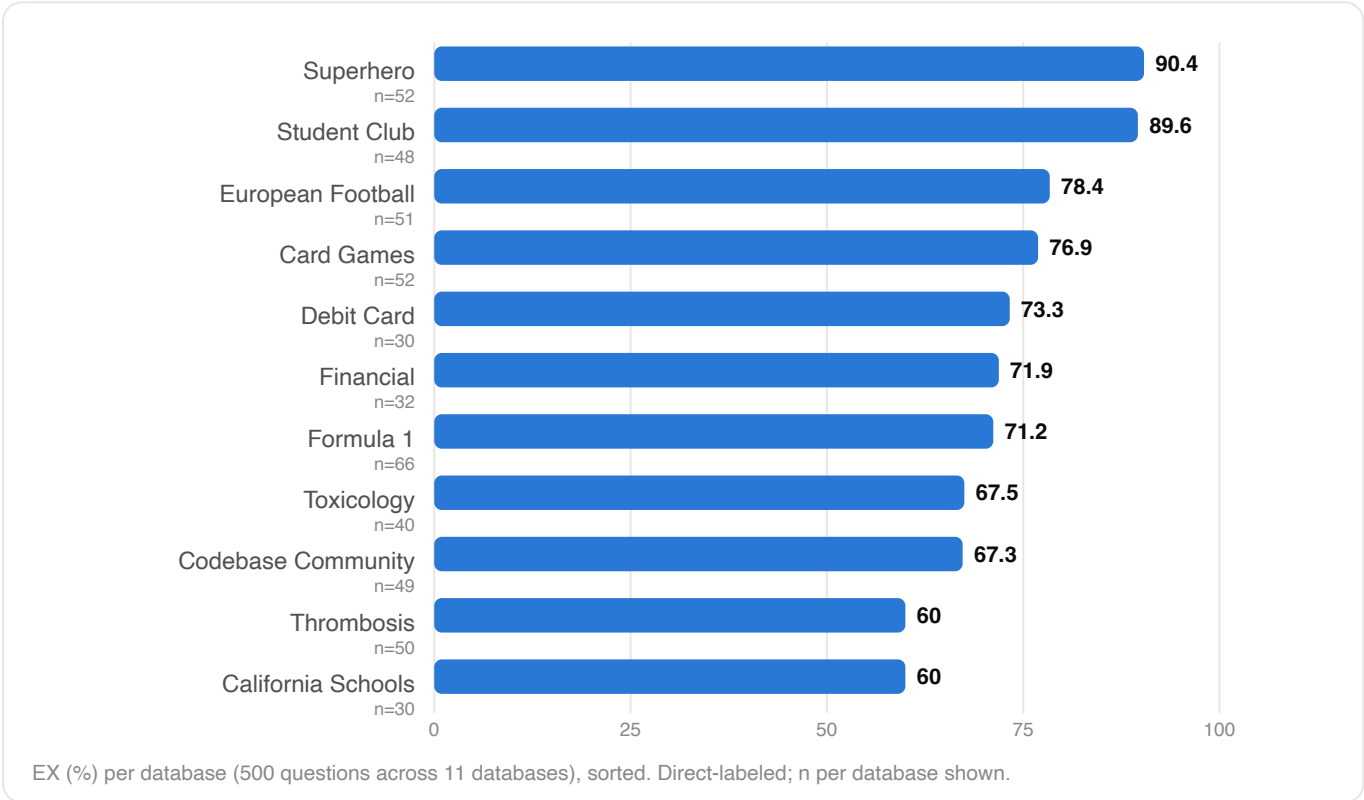
STEP 4 – write ONLY the final SQL (one statement, from SELECT; no fences/comments).

4 · Results

4.1 By difficulty



4.2 By database



5 · Operational KPIs

5.1 Wiki (MCP) tool usage

Reads per question against the OKF wiki. Total across the run: **2,234** calls.

PER-AGENT CALLS	VALUE
Min	1
Average (mean)	4.47
Median	4.0
Max	11

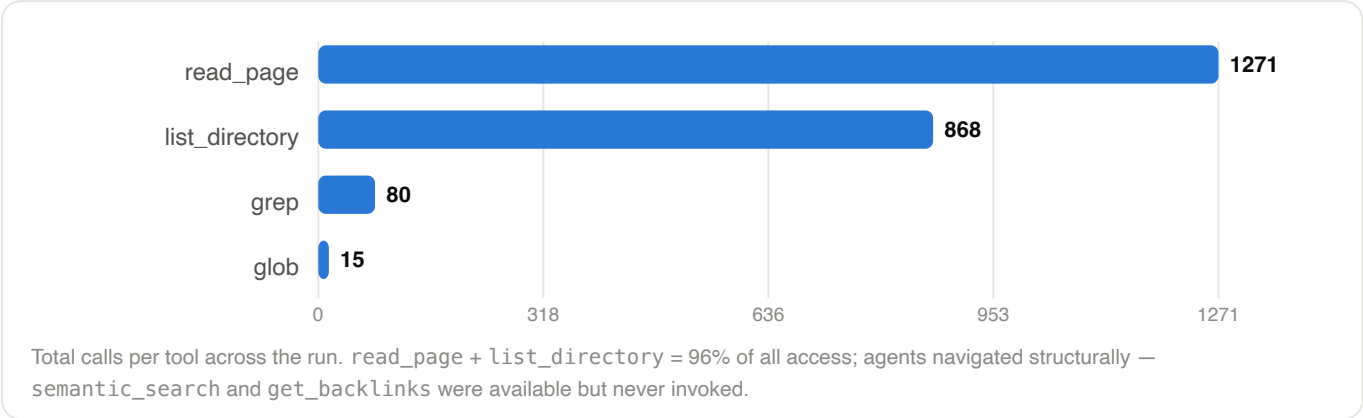
5.2 Agent execution duration

Per-agent wall-clock (s). Real elapsed ~51 min at ~14-way concurrency.

PER-AGENT SECONDS	VALUE
Min	17.8
Average (mean)	83.0
Median	63.6
p90	137.6

PER-AGENT CALLS	VALUE	PER-AGENT SECONDS	VALUE
Std dev	1.46	p95	189.3
		Max	499.1

5.3 Most-used wiki tools



TOOL	CALLS	SHARE
read_page	1,271	56.9%
list_directory	868	38.9%
grep	80	3.6%
glob	15	0.7%
get_backlinks	0	0.0%
semantic_search	0	0.0%

6 · Validity & integrity

- **Gold isolation.** Agents read a question file containing only the question, evidence, and database name — never the gold SQL. Verified: **500 / 500** agents used the MCP wiki path; byte-identical-to-gold predictions were 56/500 (11.2%), the expected base rate of canonical short queries, not copies.
- **Grader fidelity.** Feeding the gold SQL as predictions through this same evaluator yields EX 99.6 (two queries exceed BIRD's own 30 s timeout), confirming the grader is byte-for-byte the official one.
- **No schema leakage.** The agent never received CREATE-TABLE statements; all schema knowledge was reconstructed from the OKF bundle.