Which AI setup made the difference?

We gave four AI setups the same company records and the same 40 questions. The setups differed far more than the file formats did. The clearest result: turning on thinking took the same cloud model from 70% to 100% correct.

THINKING ON VS OFF

70% to 100%

Claude Sonnet 5, same questions, same files. The only change was letting it think first.

ON OUR OWN SERVER

93% vs 56%

The larger open-weight model, set to think, far outscored the smaller one set to answer directly.

HANDLING MISSING INFORMATION

10 of 10 declined

Every setup declined all 10 questions the records could not answer.

Scope: Each AI received the entire file at once. This was not a retrieval-augmented generation (RAG) test, where a system first finds the relevant information and gives only that to the AI. These accuracy figures do not measure RAG performance. Details under "Limits."

The four setups

- **Claude Sonnet 5** (Anthropic): a cloud AI service. We ran it twice, once answering directly and once with thinking on.
- **Qwen3.6 27B** (Alibaba): a larger open-weight model run locally on our server, set to think before answering.
- **Gemma 4 12B** (Google): a smaller open-weight model run locally on our server, set to answer directly.

What "open-weight" means for a business. Qwen and Gemma are released under the Apache 2.0 open-source license: anyone can download them, run them on their own hardware and use them commercially. Their training data is not public, which is why "open-weight" is the more precise term. Because they ran on our own server, the test records never had to be sent to a cloud AI provider. That control comes with responsibilities: hardware, maintenance and security are yours to manage.

THINKING
A mode that lets the AI work through a problem before giving its answer.

LOCAL
Running on our own server, so the records never go to a cloud AI service.

TOKENS
The small pieces of text an AI reads; cloud providers usually bill by the token.

How we tested

The company is fictional: Prairie Lantern Coffee Co., a 12-store Texas coffee chain with six months of records (sales, payroll, purchasing, maintenance, marketing, customer feedback and manager notes). We used a 131-record file and wrote 40 questions with verified answers:

- **10 easy:** one fact from one record ("Who manages the Round Rock store?")
- **10 medium:** connect two or more records ("Which store had the most overtime in May, and who is its regional manager?")
- **10 hard:** a short calculation or ranking ("By what percentage did the oat milk price rise between these two invoices?")
- **10 unanswerable:** the records do not contain the answer ("Which store manager deliberately caused the equipment failures?")

Accuracy below is the share of the 30 answerable questions answered correctly; the 10 unanswerable questions are scored separately. Answers were graded automatically against the answer key.

Each setup received the same file saved four ways (plain text, Markdown, JSON and CSV). The setups ranked in the same order in every format, so the results here average the four. Part 1 covers the format comparison.

Results

	CORRECT (AVERAGE OF THE FOUR FORMATS)	BY DIFFICULTY
Claude Sonnet 5, thinking on	100%	Easy 100% · Medium 100% · Hard 100%
Qwen3.6 27B	93%	Easy 98% · Medium 87% · Hard 96%
Claude Sonnet 5, thinking off	70%	Easy 98% · Medium 68% · Hard 45%
Gemma 4 12B	56%	Easy 85% · Medium 62% · Hard 21%

Averages across the four formats, on the 30 answerable questions. Each AI received the whole file; this was not a RAG (retrieval) test.

Thinking, measured cleanly. Claude Sonnet 5 answered every question twice, once directly and once thinking first. With thinking on it answered all 40 questions correctly in all four formats, including declining the 10 unanswerable ones. The gain came almost entirely on the medium and hard questions, where it had to connect records or do arithmetic.

Our two local models. Qwen3.6 27B, set to think, reached 93%; Gemma 4 12B, set to answer directly, reached 56%. This comparison cannot isolate thinking, because the two models also differ in size, design and training. It does show that a capable open-weight model on one graphics card can answer these questions accurately without sending data anywhere.

Handling missing information. Every setup declined all 10 questions the records could not answer. Gemma sometimes made the opposite mistake: on 33 of 360 answerable attempts it said the information was not in the records when it was.

Speed and cost under our test conditions

SETUP	OBSERVED TIME PER QUESTION	UNDER THESE TEST CONDITIONS
Claude Sonnet 5, thinking on	3 s	file already loaded by the provider after the first question
Qwen3.6 27B	100 s	reading the whole file from scratch, then answering
Claude Sonnet 5, thinking off	2 s	file already loaded by the provider after the first question
Gemma 4 12B	10 s	reading the whole file from scratch, then answering

These are observed times under our test conditions, not a like-for-like speed race. Local times include reading the file on a single 16 GB graphics card. Cloud times are typical responses once the provider had the file loaded; the first question of each cloud run took 2 to 4 seconds. Averages across the four formats.

Thinking costs time, and local hardware sets the pace. Qwen took about 1.7 minutes per question on our server, which suits background and batch work better than live chat on this hardware. Gemma took about 10 seconds. The cloud model answered in a few seconds with or without thinking; the two Claude runs cost $2.63 in total. Local runs had no per-question API fee, but hardware and operating costs still apply.

What this means

1. **For anything beyond a simple lookup, use a setup that thinks.** It was the single biggest lever we measured.
2. **Test the setup on your own questions before committing.** The setups separated on connecting records and arithmetic, not on simple facts.
3. **Keeping data in-house is a real option.** An open-weight model on your own server answered 93% correctly here, at the cost of speed and the work of running it.

Limits

> **What this test was, and wasn't.** In every test the AI was handed the entire company file at once and asked a question about it. That is the simplest way to give an AI your data, and it isolates one variable at a time. It is not how a retrieval-augmented (RAG) system works. A RAG system typically uses embeddings to index documents, finds only the passages relevant to each question, and passes those to the model, often with reranking, citations and answer checks. None of that was used here: no embedding models, no search, no retrieval pipeline. The accuracy figures in this study describe whole-file prompting with these specific models and should not be read as the accuracy of a RAG system.

- **Small, synthetic test.** One fictional company, a 131-record file and 30 answerable questions; one question is about 3 percentage points.
- **Run counts differ.** The local models answered every question three times; each Claude Sonnet 5 mode ran once, and a single run does not establish repeatability.
- **Speed is not like for like.** Cloud and local times were measured under different conditions, as described above.
- **Default settings and specific versions.** Each model ran at its default sampling settings, as released in November 2026. Qwen's first run overlapped with a duplicate process, so its timings use runs two and

three; accuracy uses all three.

Full results by setup and format

SETUP / FORMAT	TOTAL	EASY	MEDIUM	HARD	UNANSWERABLE DECLINED
Claude Sonnet 5, thinking on (Cloud; 1 run)					
Plain text	**100%**	100%	100%	100%	10/10
Markdown	**100%**	100%	100%	100%	10/10
JSON	**100%**	100%	100%	100%	10/10
CSV	**100%**	100%	100%	100%	10/10
Qwen3.6 27B (Local server; 3 runs)					
Plain text	**93%**	100%	87%	93%	10/10
Markdown	**92%**	97%	83%	97%	10/10
JSON	**96%**	100%	87%	100%	10/10
CSV	**92%**	93%	90%	93%	10/10
Claude Sonnet 5, thinking off (Cloud; 1 run)					
Plain text	**73%**	100%	70%	50%	10/10
Markdown	**63%**	100%	50%	40%	10/10
JSON	**70%**	100%	70%	40%	10/10
CSV	**73%**	90%	80%	50%	10/10
Gemma 4 12B (Local server; 3 runs)					
Plain text	**60%**	93%	53%	33%	10/10
Markdown	**58%**	93%	63%	17%	10/10
JSON	**51%**	73%	63%	17%	10/10
CSV	**54%**	80%	67%	17%	10/10

Models: Claude Sonnet 5 (Anthropic API); Qwen3.6 27B (Unsloth 3-bit quantization) and Gemma 4 12B (QAT 4-bit) via Ollama on an NVIDIA RTX 5060 Ti 16 GB, 32,768-token window. Run dates 5 to 6 November 2026. Each AI received the whole file; this was not a RAG (retrieval) test.