FILING-EVIDENCE SECTORS · RESULTS PER MODEL

Qwen3.5-9B

10,000 questions over 200 filings of 100 companies in ten sectors, answered on one RTX 4090

Scorecard

86.4%
correct with a checkable quote, plain setup (7,780 of 9,002)

304
quietly wrong answers (confident-wrong) in 10,000, plain setup

99.0%
correct "not found" when the number was absent

44 Wh
GPU energy per 1,000 answers (₹0.35 at ₹8/kWh)

By sector

Plain setup (C1) and check + retry (C2), all 10,000 questions. Accuracy (Acc.) is the share of answerable questions answered correctly with a checkable quote; confident-wrong (CW) is the share of all questions answered wrongly with a quote that passes every check.

Sector	Acc. C1	95% CI	CW C1	Acc. C2	CW C2
Banks	87.0%	84.6–89.0	2.6%	87.2%	2.7%
Energy (power and utilities)	89.8%	87.6–91.6	1.7%	90.1%	1.7%
Healthcare	88.9%	86.7–90.8	2.9%	89.7%	2.9%
Insurance	82.2%	79.6–84.6	4.6%	82.7%	4.7%
Media	86.1%	83.7–88.2	3.2%	86.9%	3.3%
Metals and mining	88.3%	86.1–90.3	2.9%	89.1%	2.9%
Oil and gas	85.1%	82.7–87.3	5.6%	85.7%	5.6%
Real estate	89.0%	86.8–90.9	2.7%	89.6%	2.9%
Technology	86.0%	83.6–88.1	1.5%	86.6%	1.5%
Telecom	81.8%	79.1–84.2	2.7%	82.1%	2.7%
All sectors	**86.4%**	**85.7–87.1**	**3.0%**	**87.0%**	**3.1%**

Setups and checks

Each comparison is paired (the same questions) with an exact McNemar test; p is Holm-corrected within the model or the check. CW = confident-wrong. Thinking and the robustness checks use the 1,000-question subset.

Comparison	Correct	p	CW	p
Plain → check + retry (10,000)	7,780 → 7,828	< 0.001	304 → 309	0.06
Plain → thinking (subset)	743 → 689	0.001	34 → 15	0.007
Plain → Markdown tables, strict quote check (R1)	743 → 268	< 0.001	34 → 8	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	743 → 749	1.00	34 → 32	1.00
Plain → alternative prompt (R2)	743 → 715	0.02	34 → 45	0.38
Plain → identical repeat run (R3)	743 → 740	0.75	34 → 35	1.00
Full precision → 4-bit build (exploratory)	743 → 711	< 0.001	34 → 42	0.35
Repeat run: 0.6% of questions changed outcome.				

Speed

Plain setup: median 5.4 s per answer (p95 8.2 s) at the runner's fixed concurrency, 10,654 answers per hour, 53 output tokens per answer on average. Thinking: median 72.4 s per answer, 3734 output tokens on average.

← All models · Every answer of this model, with its raw reply, is in the dataset (https://huggingface.co/datasets/pramana-research/filing-evidence-sectors) under `results/qwen3.5-9b/`.

© 2026 Pramana Research · Text, data and figures: CC BY 4.0 · Code: Apache 2.0 · Legal

Model outputs are not investment advice. Check every figure against its source.

Pramana Research is an independent research group, not affiliated with any other company of a similar name.