FILING-EVIDENCE SECTORS · RESULTS PER MODEL

Granite-4.2-8B

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

Scorecard

75.9%
correct with a checkable quote, plain setup (6,832 of 9,002)

772
quietly wrong answers (confident-wrong) in 10,000, plain setup

98.2%
correct "not found" when the number was absent

43 Wh
GPU energy per 1,000 answers (₹0.34 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	76.1%	73.2–78.8	6.4%	78.3%	7.2%
Energy (power and utilities)	84.3%	81.8–86.6	4.8%	86.7%	5.0%
Healthcare	77.9%	75.1–80.5	7.9%	80.3%	8.6%
Insurance	72.0%	69.0–74.8	10.6%	74.8%	11.5%
Media	74.3%	71.4–77.1	7.2%	78.3%	7.9%
Metals and mining	75.0%	72.1–77.7	8.8%	77.1%	9.6%
Oil and gas	72.5%	69.5–75.3	12.6%	74.5%	14.1%
Real estate	73.3%	70.3–76.1	6.2%	77.8%	7.0%
Technology	78.1%	75.3–80.7	6.4%	79.7%	6.8%
Telecom	75.3%	72.4–78.0	6.3%	77.0%	6.7%
All sectors	**75.9%**	**75.0–76.8**	**7.7%**	**78.4%**	**8.4%**

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)	6,832 → 7,062	< 0.001	772 → 844	< 0.001
Plain → thinking (subset)	649 → 721	< 0.001	90 → 34	< 0.001
Plain → Markdown tables, strict quote check (R1)	649 → 111	< 0.001	90 → 17	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	649 → 642	1.00	90 → 68	0.007
Plain → alternative prompt (R2)	649 → 629	0.34	90 → 86	1.00

Speed

Plain setup: median 4.8 s per answer (p95 7.5 s) at the runner's fixed concurrency, 11,339 answers per hour, 48 output tokens per answer on average. Thinking: median 179.7 s per answer, 2718 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/granite-4.2-8b/`.

© 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.