FILING-EVIDENCE SECTORS · RESULTS PER MODEL

Ministral-3-8B

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

Scorecard

71.9%
correct with a checkable quote, plain setup (6,476 of 9,002)

652
quietly wrong answers (confident-wrong) in 10,000, plain setup

99.1%
correct "not found" when the number was absent

27 Wh
GPU energy per 1,000 answers (₹0.21 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	72.1%	69.1–74.9	5.5%	74.6%	5.8%
Energy (power and utilities)	78.7%	75.9–81.2	3.6%	80.9%	4.0%
Healthcare	74.2%	71.3–77.0	7.7%	77.2%	8.1%
Insurance	71.8%	68.7–74.6	8.1%	73.6%	8.7%
Media	73.4%	70.5–76.2	5.8%	76.8%	6.3%
Metals and mining	68.9%	65.8–71.8	8.5%	72.2%	9.3%
Oil and gas	66.0%	62.8–69.0	11.2%	68.0%	12.0%
Real estate	70.8%	67.7–73.7	6.6%	74.2%	6.9%
Technology	72.8%	69.8–75.6	4.1%	75.1%	4.6%
Telecom	70.8%	67.7–73.7	4.1%	72.3%	4.6%
All sectors	**71.9%**	**71.0–72.9**	**6.5%**	**74.5%**	**7.0%**

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)	6,476 → 6,705	< 0.001	652 → 703	< 0.001
Plain → Markdown tables, strict quote check (R1)	648 → 203	< 0.001	52 → 19	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	648 → 547	< 0.001	52 → 47	1.00
Plain → alternative prompt (R2)	648 → 633	0.75	52 → 65	0.29
Plain → identical repeat run (R3)	648 → 639	0.28	52 → 56	1.00
Full precision → 4-bit build (exploratory)	648 → 639	0.75	52 → 55	0.94
Repeat run: 3.4% of questions changed outcome.				

Speed

Plain setup: median 3.6 s per answer (p95 5.9 s) at the runner's fixed concurrency, 15,551 answers per hour, 66 output tokens per answer 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/ministral-3-8b/`.

© 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.