FILING-EVIDENCE SECTORS · RESULTS PER MODEL

OLMo-3-7B

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

Scorecard

19.2%
correct with a checkable quote, plain setup (1,726 of 9,002)

849
quietly wrong answers (confident-wrong) in 10,000, plain setup

98.0%
correct "not found" when the number was absent

40 Wh
GPU energy per 1,000 answers (₹0.32 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	19.0%	16.6–21.7	7.9%	19.1%	8.1%
Energy (power and utilities)	21.8%	19.2–24.6	11.0%	21.8%	11.1%
Healthcare	21.0%	18.5–23.8	8.0%	21.0%	8.0%
Insurance	17.1%	14.8–19.7	9.8%	17.1%	9.8%
Media	20.8%	18.3–23.5	7.4%	20.8%	7.5%
Metals and mining	22.2%	19.6–25.1	7.8%	22.3%	7.9%
Oil and gas	16.4%	14.1–19.0	9.7%	16.4%	9.9%
Real estate	17.9%	15.5–20.5	6.1%	18.0%	6.1%
Technology	17.4%	15.1–20.1	9.5%	17.4%	9.5%
Telecom	18.1%	15.7–20.8	7.7%	18.2%	7.8%
All sectors	**19.2%**	**18.4–20.0**	**8.5%**	**19.2%**	**8.6%**

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)	1,726 → 1,730	0.12	849 → 857	0.02
Plain → Markdown tables, strict quote check (R1)	181 → 56	< 0.001	75 → 37	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	181 → 145	0.01	75 → 51	0.06
Plain → alternative prompt (R2)	181 → 89	< 0.001	75 → 159	< 0.001
Plain → identical repeat run (R3)	181 → 181	1.00	75 → 74	1.00
Full precision → 4-bit build (exploratory)	181 → 191	0.75	75 → 81	0.94
Repeat run: 1.6% of questions changed outcome.				

Speed

Plain setup: median 5.2 s per answer (p95 8.1 s) at the runner's fixed concurrency, 10,778 answers per hour, 41 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/olmo-3-7b/`.

© 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.