FILING-EVIDENCE SECTORS · RESULTS PER MODEL

gpt-oss-20b

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

Scorecard

78.6%
correct with a checkable quote, plain setup (7,078 of 9,002)

1,028
quietly wrong answers (confident-wrong) in 10,000, plain setup

94.1%
correct "not found" when the number was absent

24 Wh
GPU energy per 1,000 answers (₹0.19 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	80.1%	77.4–82.6	8.6%	82.8%	8.7%
Energy (power and utilities)	84.6%	82.0–86.8	6.5%	86.9%	6.8%
Healthcare	79.6%	76.8–82.1	11.0%	81.4%	11.6%
Insurance	77.1%	74.3–79.7	12.2%	78.7%	12.6%
Media	78.4%	75.6–81.0	9.5%	80.1%	9.6%
Metals and mining	74.7%	71.7–77.4	13.7%	76.6%	13.9%
Oil and gas	74.6%	71.7–77.3	16.0%	75.5%	16.2%
Real estate	80.4%	77.7–82.9	8.9%	82.4%	9.2%
Technology	80.3%	77.6–82.8	8.2%	82.9%	8.6%
Telecom	76.4%	73.6–79.1	8.2%	78.3%	8.6%
All sectors	**78.6%**	**77.8–79.5**	**10.3%**	**80.6%**	**10.6%**

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)	7,078 → 7,252	< 0.001	1,028 → 1,058	< 0.001
Plain → thinking (subset)	697 → 736	0.004	105 → 51	< 0.001
Plain → Markdown tables, strict quote check (R1)	697 → 499	< 0.001	105 → 71	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	697 → 645	< 0.001	105 → 96	1.00
Plain → alternative prompt (R2)	697 → 684	1.00	105 → 123	0.38

Speed

Plain setup: median 2.8 s per answer (p95 5.2 s) at the runner's fixed concurrency, 19,068 answers per hour, 104 output tokens per answer on average. Thinking: median 17.8 s per answer, 1899 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/gpt-oss-20b/`.

© 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.