FILING-EVIDENCE SECTORS · RESULTS PER MODEL

GLM-4.6V-Flash

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

Scorecard

76.9%
correct with a checkable quote, plain setup (6,923 of 9,002)

977
quietly wrong answers (confident-wrong) in 10,000, plain setup

95.8%
correct "not found" when the number was absent

42 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	77.4%	74.6–80.1	9.2%	78.7%	9.4%
Energy (power and utilities)	83.2%	80.6–85.5	7.8%	84.7%	7.8%
Healthcare	77.1%	74.3–79.7	11.2%	78.2%	11.4%
Insurance	75.6%	72.6–78.3	11.8%	76.6%	12.4%
Media	77.4%	74.6–80.1	8.1%	79.0%	8.4%
Metals and mining	77.0%	74.1–79.6	11.7%	78.6%	11.8%
Oil and gas	72.5%	69.5–75.3	14.7%	74.1%	14.8%
Real estate	76.2%	73.3–78.9	9.0%	77.8%	9.4%
Technology	77.1%	74.3–79.7	8.2%	78.0%	8.4%
Telecom	75.4%	72.5–78.1	6.0%	77.4%	6.3%
All sectors	**76.9%**	**76.0–77.8**	**9.8%**	**78.3%**	**10.0%**

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)	6,923 → 7,048	< 0.001	977 → 1,001	< 0.001
Plain → thinking (subset)	673 → 743	< 0.001	102 → 64	< 0.001
Plain → Markdown tables, strict quote check (R1)	673 → 391	< 0.001	102 → 44	< 0.001
Plain → Markdown tables, pipes ignored (exploratory)	673 → 639	0.01	102 → 94	1.00
Plain → alternative prompt (R2)	673 → 674	1.00	102 → 117	0.38

Speed

Plain setup: median 4.6 s per answer (p95 7.3 s) at the runner's fixed concurrency, 11,996 answers per hour, 50 output tokens per answer on average. Thinking: median 10.7 s per answer, 392 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/glm-4.6v-flash/`.

© 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.