FILING-EVIDENCE SECTORS · RESULTS PER MODEL

# Phi-4

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

## Scorecard

**63.3%**
correct with a checkable quote, plain setup (5,695 of 9,002)

**1,488**
quietly wrong answers (confident-wrong) in 10,000, plain setup

**96.4%**
correct "not found" when the number was absent

**44 Wh**
GPU energy per 1,000 answers (₹0.36 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 | 65.6% | 62.4–68.6 | 15.1% | 68.4% | 16.4% |
| Energy (power and utilities) | 65.1% | 61.9–68.2 | 15.9% | 67.3% | 16.4% |
| Healthcare | 58.4% | 55.2–61.6 | 17.9% | 60.9% | 20.3% |
| Insurance | 62.4% | 59.2–65.5 | 16.0% | 64.6% | 17.7% |
| Media | 67.0% | 63.9–70.0 | 12.1% | 69.3% | 13.0% |
| Metals and mining | 65.7% | 62.5–68.7 | 13.6% | 67.3% | 14.9% |
| Oil and gas | 56.0% | 52.7–59.2 | 23.3% | 57.6% | 24.8% |
| Real estate | 64.0% | 60.8–67.1 | 11.5% | 67.1% | 12.5% |
| Technology | 64.8% | 61.6–67.8 | 13.6% | 66.4% | 14.3% |
| Telecom | 63.7% | 60.5–66.7 | 9.8% | 65.7% | 10.9% |
| **All sectors** | **63.3%** | **62.3–64.3** | **14.9%** | **65.5%** | **16.1%** |

## 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) | 5,695 → 5,894 | < 0.001 | 1,488 → 1,612 | < 0.001 |
| Plain → Markdown tables, strict quote check (R1) | 566 → 331 | < 0.001 | 152 → 69 | < 0.001 |
| Plain → Markdown tables, pipes ignored (exploratory) | 566 → 450 | < 0.001 | 152 → 107 | < 0.001 |
| Plain → alternative prompt (R2) | 566 → 529 | 0.03 | 152 → 123 | 0.02 |

## Speed

Plain setup: median 5.5 s per answer (p95 9.0 s) at the runner's fixed concurrency, 9,983 answers per hour, 59 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/phi-4/`.

© 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.