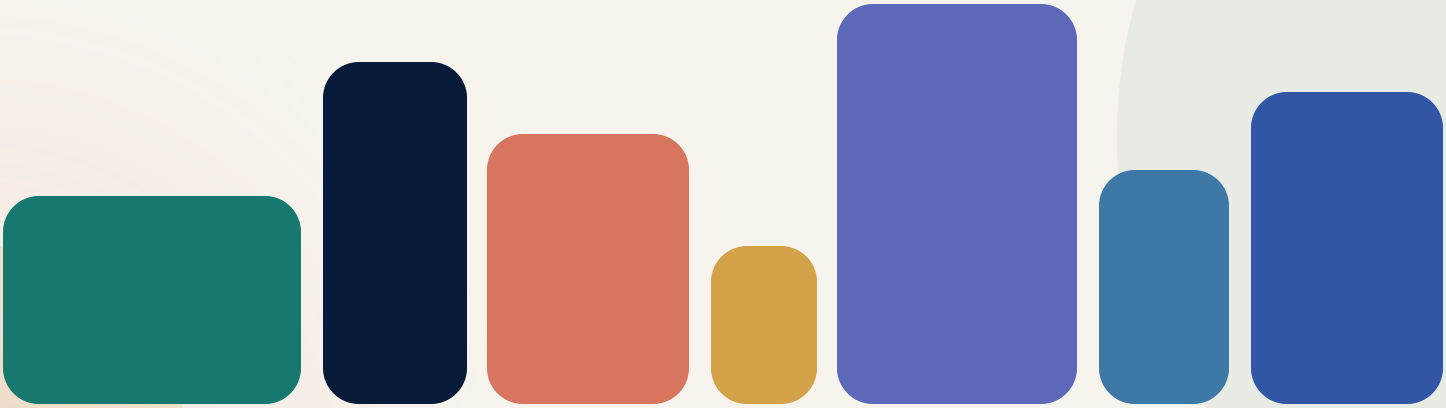
The Practical Work Benchmark

Analyzing the seven leading models today for how they perform doing real, practical work.



elvex

EXECUTIVE BRIEFING

All of these models are great. Their costs are not the same.

The practical baseline has moved

Any model in this report would have looked exceptional by last year's standards. On a shared business-document benchmark, all seven sit inside an 85 to 100 relative-rating band. **The buying decision today is entirely about price: intelligence output per dollar spent.**

Open weights are becoming daily drivers

Practitioner reports increasingly describe Kimi, GLM, and DeepSeek as serious working models, not fallback options. Users are moving the model inside existing workflows and reporting strong practical results at much lower API prices.

General quality, subtle differences

Claude is preferred for some prose. Sol leads finished editable documents. Kimi excels at detailed specifications. GLM offers remarkable workflow economics. None of those results makes another model broadly incapable.

The six jobs in The Practical Work Benchmark

- **Automate repetitive reporting and workflows**
- **Analyze data for decisions**
- **Draft and refine professional communication**
- **Write documentation**
- **Measure performance and ROI**
- **Integrate systems, troubleshoot problems**

HOW WE KNOW

The report screened 82 recent public source records: shared tests, practitioner comparisons, failure reports, and official product facts. Performance evidence had to concern the exact model and be observed from May 11 through August 9, 2026. Artificial Analysis, Composio, and exact-model practitioner reports provide the clearest open-weight adoption signals.

Six jobs knowledge workers do with LLMs

You will notice that, generally, the everyday work being done with LLMs are frequently variations of producing written materials (sales emails to documentation), querying data to support strategic decisions making, and troubleshooting how to solve problems and use tools.

01

Automate repetitive reporting and workflows

Run recurring, multi-step reporting and operational work with less manual effort.

Sample: Every Friday, pull sales, flag exceptions, and post a summary to the team.

02

Analyze data for decisions

Find, reconcile, transform, and explain the information needed for a decision.

Sample: Reconcile ERP exports and explain the variance behind this month's margin.

03

Draft and refine professional communication

Create or improve workplace writing with the right message, tone, and level of detail.

Sample: Turn rough meeting notes into a concise client update in your firm's voice.

04

Write documentation

Create durable reference material that helps people understand and execute work.

Sample: Draft a month-end close SOP with owners, checks, exceptions, and escalation paths.

05

Measure performance and ROI

Measure results, explain the drivers, and communicate value or return.

Sample: Combine campaign, revenue, and cost data into an ROI readout for leadership.

06

Integrate systems, troubleshoot problems

Connect business systems and solve the problems that prevent data and tools from working together.

Sample: Map CRM fields to billing, identify why records fail, and route exceptions for review.

SECTION 02 • THE DECISION VIEW

A strong field, with different specialties

The labels are relative to today's leading models. Writing includes tone, voice, and how much the response feels like a recognizable model rather than the intended author.

Incredible · Great · Good · Just OK

Evidence depth: ●●● high • ●●○ medium • ●○○ low

Model	Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
DeepSeek-V4-Pro	Just OK ●●○	Great ●●○	Good ●●○	Good ●●○	Good ●●○	Good ●●○
Kimi K3	Good ●●○	Great ●●●	Good ●●○	Great ●●●	Good ●●○	Good ●●○
Claude Opus 5	Great ●●●	Incredible ●●●	Good ●●●	Good ●●●	Good ●●○	Good ●●●
GPT-5.6 Sol	Good ●●●	Good ●●●	Good ●●●	Incredible ●●●	Good ●●●	Good ●●●
GLM-5.2	Great ●●●	Incredible ●●●	Just OK ●○○	Good ●●●	Great ●●○	Good ●●●
Gemini 3.6 Flash	Good ●●○	Good ●●○	Great ●○○	Great ●●●	Good ●●○	Good ●○○
Grok 4.5	Good ●●○	Great ●●●	Just OK ●●○	Great ●●●	Good ●●○	Good ●●○

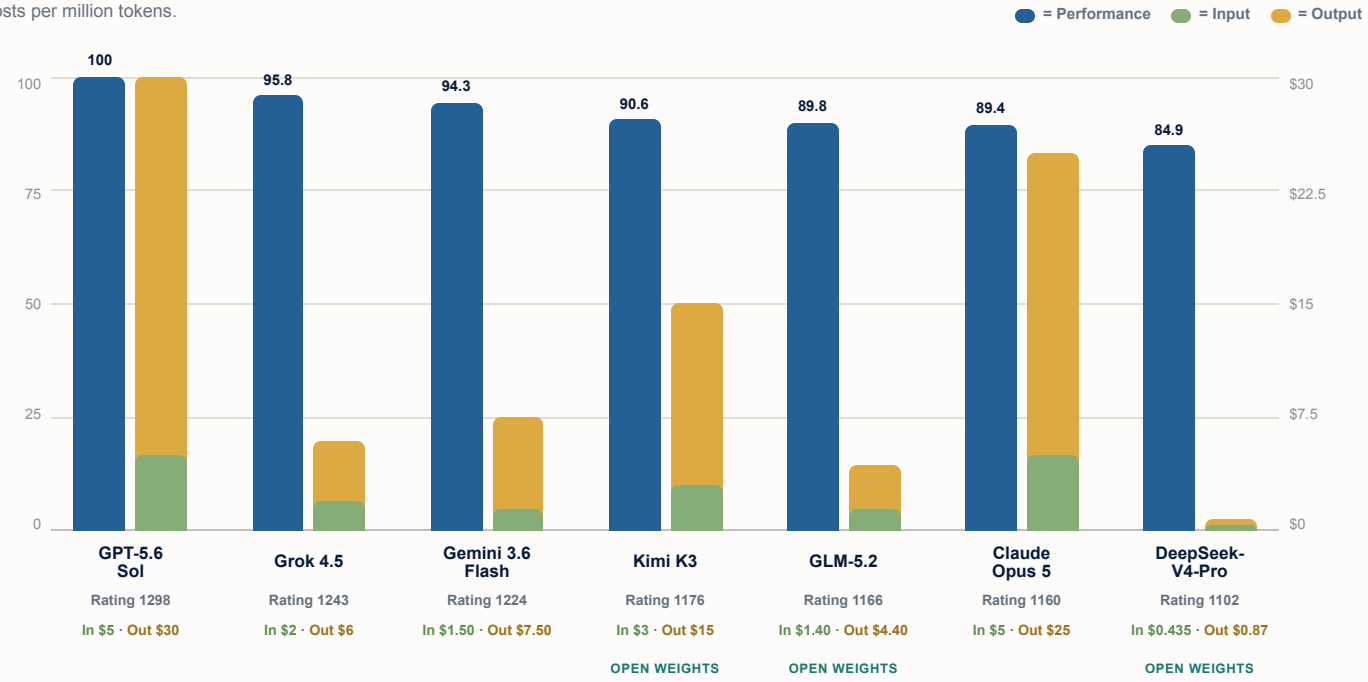
Today's curve is steep

Even “Just OK” is relative to an unusually capable group. “Good” often means the model is very useful but the evidence is mixed. Low evidence is not a failure. The matrix is a routing guide, not a claim that any model here is broadly weak. Dots show how much recent public evidence supports the rating.

Similar practical capability. Very different cost.

Capability and token cost for practical work

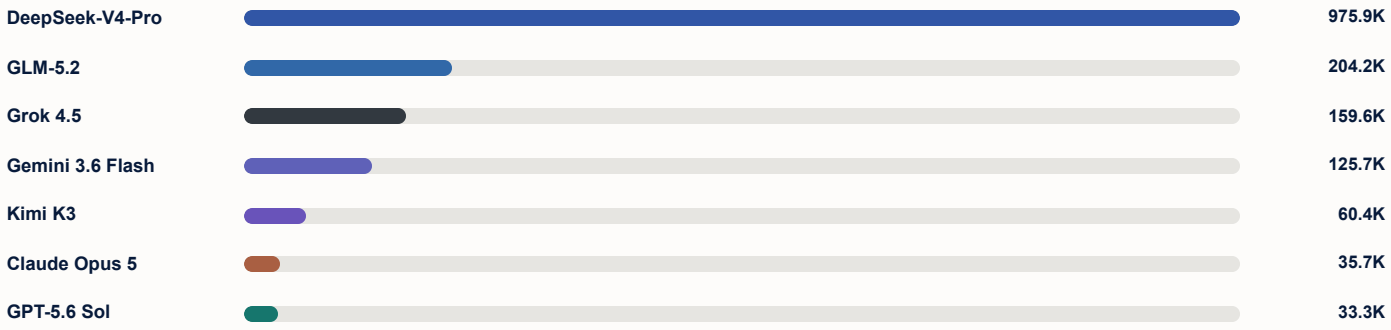
Performance is the left axis, on a graded relative scale where the leader equals 100. Price is the right axis, input/output costs per million tokens.



DocBench ratings and direct-provider list prices per 1M tokens, frozen August 9, 2026.

Intelligence output per dollar spent

Output-token allowance per $1, multiplied by each model's displayed DocBench index.



Formula: output tokens per $1 × (DocBench rating ÷ 1298). A rough proxy for how much intelligence you get per dollar spent.

elvex

SECTION 03 • MODEL PROFILE

DeepSeek-V4-Pro

DeepSeek • Open weights, MIT • Text-only input and output

The lowest-cost model in this group and a capable choice for supervised analysis, extraction, and writing that must follow a supplied style.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Just OK ●●○	Great ●●○	Good ●●○	Good ●●○	Good ●●○	Good ●●○

Best suited for

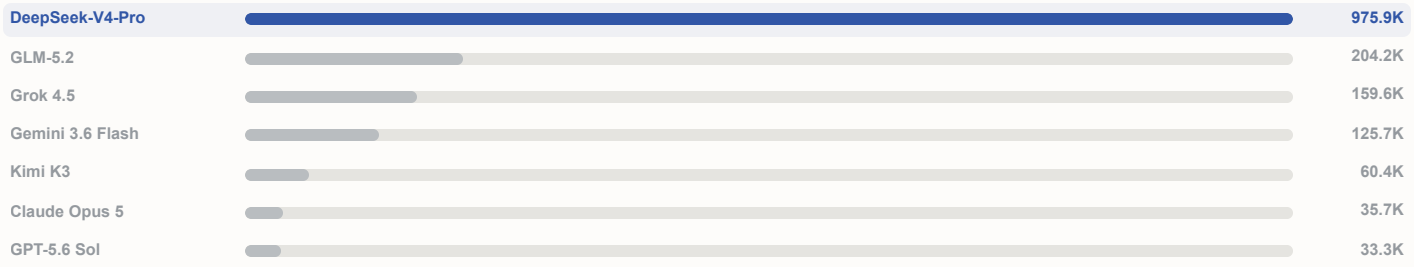
- Low-cost analysis and clean-text extraction.
- Grounded financial or regulatory research.
- Writing that follows the tone and structure of supplied source material.

Watchouts

- You cannot paste in an image or scanned page and have the model read it.
- Not yet convincing for unattended, long-running operational work.
- Ranked seventh in a small finished-document test and lacks a shared proofreading result.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

RETRIEVAL QUALITY

“You choose it for quality and pay for it in serving cost.”

Kelsus sovereign architecture test • June 2026

WRITING

“DeepSeek V4 Pro ... wrote the single best scene in the whole sample”

Foreverse blind writing test • July 17

VALUE

“For the same amount of money, it's better.”

Exact-model practitioner • June 10

Key evidence: Kelsus, Foreverse, DocBench, and STOCKTAKE.

SECTION 03 • MODEL PROFILE

Kimi K3

Moonshot AI • Open weights, custom license • Text and image input

A deep, detail-oriented model that excels at analysis and technical plans when a person controls the final presentation and system changes.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Good ●●○	Great ●●●	Good ●●○	Great ●●●	Good ●●○	Good ●●○

Best suited for

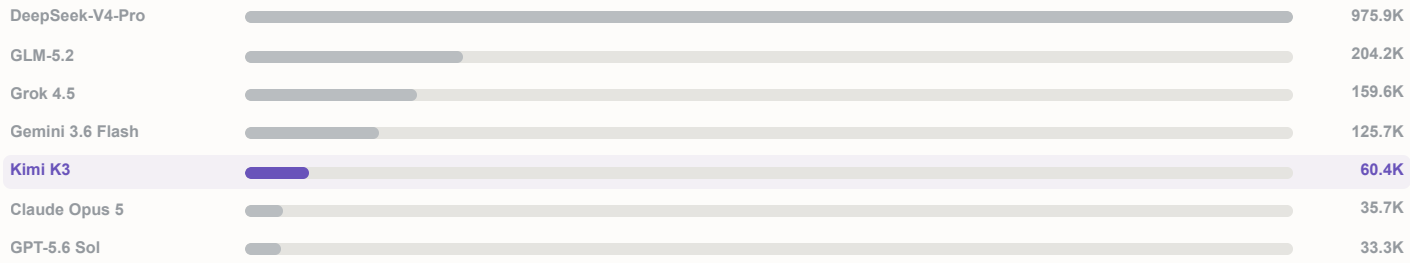
- Deep analysis of long documents and complex constraints.
- Detailed specifications, schemas, and implementation plans.
- Work where completeness matters more than executive brevity.

Watchouts

- Repeated enterprise tasks were much less reliable than one successful run.
- Default writing can be long and heavy with headings.
- Review any system change before it is allowed to take effect.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

PRACTICAL KNOWLEDGE WORK

“Kimi K3 achieves an AA-Briefcase Elo of 1543, the second-highest score overall”

Artificial Analysis • July 21

SPECIFICATIONS

“Kimi is the ultimate systems engineer.”

Tom's Guide prompt trial • July 29

WRITING WORKFLOW

“K3 is the first open model I would put into a writing workflow without hesitation.”

The Augmented Educator • July 23

Key evidence: AA-Briefcase, Toloka, Tom's Guide, and DocBench.

elvex

SECTION 03 • MODEL PROFILE

Claude Opus 5

Anthropic • Hosted model • Text and image input

An exceptional reader, reasoner, and prose writer whose excellent final work can coexist with a verbose, recognizably Claudey working voice.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Great ●●●	Incredible ●●●	Good ●●●	Good ●●●	Good ●●○	Good ●●●

Best suited for

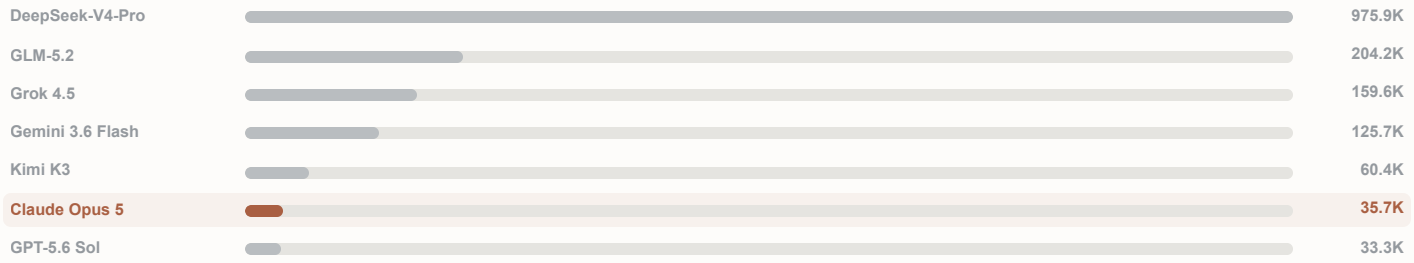
- Reading and reasoning across complex documents and images.
- Executive narrative, trade-offs, and open-ended professional prose.
- Fast, accurate editing when turnaround matters.

Watchouts

- Some experienced users report recognizing its long, narrated working style.
- Short emails and updates often need explicit length limits.
- Its prose preference did not translate into a top finished-document rank.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

EVERYDAY ADOPTION

“After spending serious time with Claude Opus 5, I've quietly switched over for most of my everyday tasks.”

Tom's Guide practitioner review • July 31

BUSINESS REASONING

“Opus 5 won … because it treated the simulation better than the others did.”

FoodTruck Bench • July 31

CAUTION: VERBOSE WORKING VOICE

“Claude Slop is legible enough to keep reading and verbose enough to make the experience miserable.”

ChatPRD / How I AI • July 25

Key evidence: Arena Writing, FoodTruck Bench, The Human Co, and ChatPRD. Recognizable voice is a reported risk, not a universal detection rate.

elvex

SECTION 03 • MODEL PROFILE

GPT-5.6 Sol

OpenAI • Hosted model • Text and image input

A highly capable all-around business model, especially strong when the result must become a polished document that a person can keep refining.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Good •••	Good •••	Good •••	Incredible •••	Good •••	Good •••

Best suited for

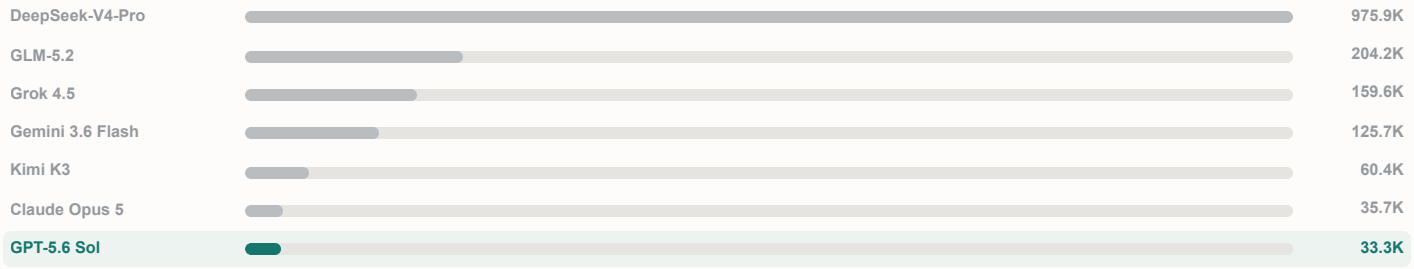
- Finished Word and PowerPoint-style business documents.
- High-completeness proofreading and document revision.
- Iterative work where a person can add context and steer the result.

Watchouts

- Large automated jobs still show meaningful completion gaps.
- Financial calculations and totals still require checking.
- Generic rewrite prompts can add headings and structure that were not requested.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

COLLABORATION

“Sol is a model that thrives on context.”

Every month-long review • July 8

DECISION SUPPORT

“ChatGPT worked out what I should do with the information.”

TechRadar connected-files test • July 29

CAUTION: CALCULATIONS

“Late in testing, Arielle caught Sol making a serious calculation error while analyzing ChatGPT usage data.”

Every month-long review • July 8

Key evidence: DocBench, ErrataBench, Every, and Arena Agent.

elvex

SECTION 03 • MODEL PROFILE

GLM-5.2

Z.AI • Open weights, MIT • Text-only input and output

A remarkably capable low-cost model for structured data, high-volume text work, and connected workflows that include a review step.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Great ●●●	Incredible ●●●	Just OK ●○○	Good ●●●	Great ●●○	Good ●●●

Best suited for

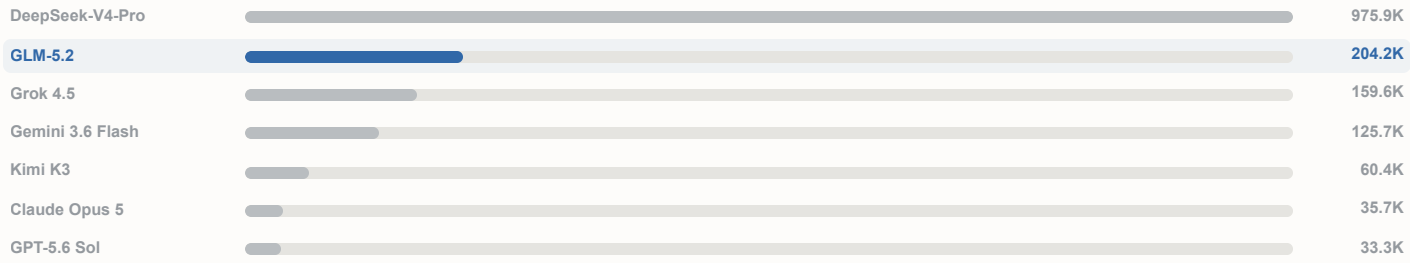
- Structured extraction and grounded retrieval from clean text.
- Efficient, high-volume workflow and data tasks.
- Low-cost document revision when review is available.

Watchouts

- You cannot paste in an image or scanned page and have the model read it.
- Proofreading was inexpensive but missed more errors than peers.
- Validate service reliability and prompt quality before large deployments.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

EVERYDAY WORKFLOWS

“For most normal agent workflows in this run, I’d use GLM 5.2.”

Composio live SaaS test • July 14

CONSISTENT WRITING

“GLM 5.2 is the only model in the top three of both tables”

Foreverse blind writing test • July 17

PRACTICAL MANAGEMENT

“I am blown away. I've been using GLM-5.2 to manage my local LLM setup and it's been extremely thorough and perfectly on point.”

Exact-model practitioner • June 17

Key evidence: Kelsus, Composio, ErrataBench, and Foreverse.

elvex

SECTION 03 • MODEL PROFILE

Gemini 3.6 Flash

Google • Hosted model • Text, image, audio, video, and PDF input

A fast, economical choice for structured work and documents, with broad input support and a particularly natural fit inside Google products.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Good ●●○	Good ●●○	Great ●○○	Great ●●●	Good ●●○	Good ●○○

Best suited for

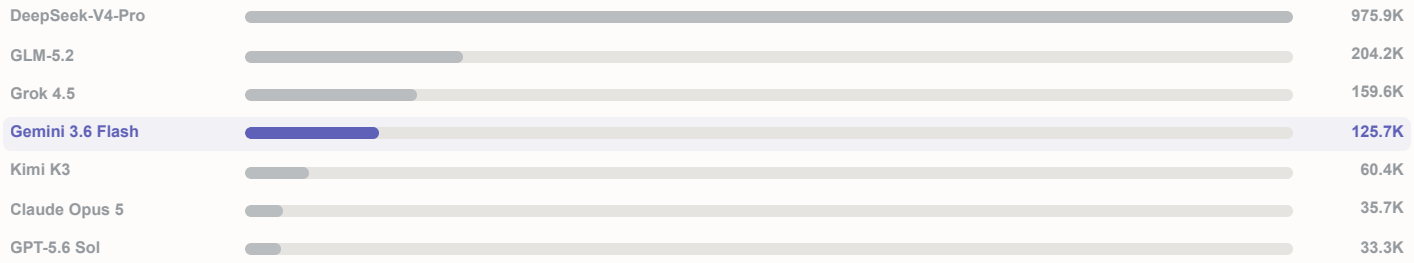
- Fast structured output and document work.
- Work that benefits from image, audio, video, or PDF input.
- High-volume tasks inside Google and API-connected workflows.

Watchouts

- Cross-source comparisons can miss relationships and priorities.
- Long-form testing found repetition that worsened over time.
- Current workplace writing evidence is favorable but still thin.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

NATURAL LANGUAGE

“Gemini's language was natural and even evocative in places.”

TechRadar same-prompt test • August 2

OPERATING ECONOMICS

“Gemini 3.6 Flash helped make Paul's daily operation more cost-effective.”

Google practitioner case study • July 28

CAUTION: SYNTHESIS

“Gemini told me what was in my files.”

TechRadar connected-files test • July 29

Key evidence: DocBench, Google practitioner case, ErrataBench, and TechRadar.

elvex

SECTION 03 • MODEL PROFILE

Grok 4.5

SpaceXAI • Hosted model • Text and image input

A strong document maker and broad-recall assistant that works best when important outputs can be checked before they are used.

Automate repetitive reporting & workflows	Analyze data for decisions	Draft & refine professional communication	Write documentation	Measure performance & ROI	Integrate systems, troubleshoot problems
Good ●●○	Great ●●●	Just OK ●●○	Great ●●●	Good ●●○	Good ●●○

Best suited for

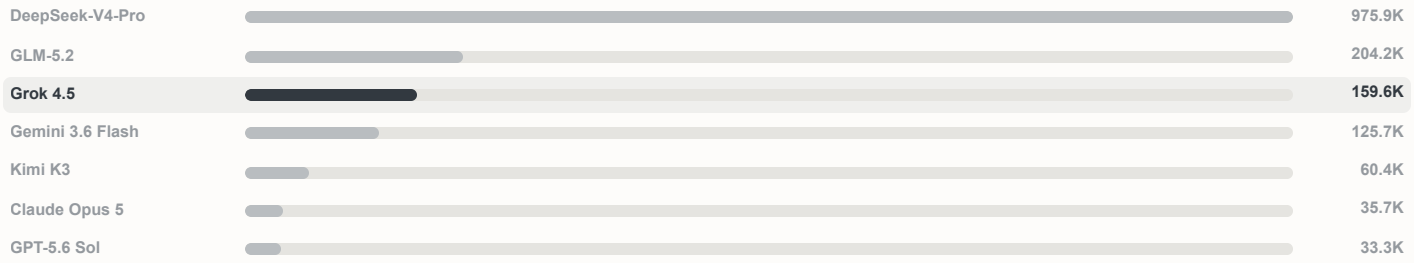
- Finished business documents and broad professional analysis.
- Recall-heavy work across connected applications.
- Checked outputs where breadth matters more than subtle voice.

Watchouts

- Long-running operating decisions underperformed a simple rule.
- Current writing reports describe flat voice and repetition.
- Used more time and tokens than GLM in a shared SaaS test.

Intelligence output per dollar spent

Output-token allowance per $1, adjusted by DocBench



Formula: output tokens per $1 × (DocBench rating ÷ 1298).

PROFESSIONAL WORK

“Overall, Grok 4.5 demonstrated the strongest overall performance.”

Snorkel AI professional-work test • July 8

CONNECTED RECALL

“Use Grok 4.5 when recall is the scary part.”

Composio live SaaS test • July 14

CAUTION: CONTINUITY

“Grok 4.5 hit repetition loops in both a fantasy epic and a palace-intrigue novel.”

Foreverse long-form test • July 16

Key evidence: Snorkel AI, Composio, DocBench, STOCKTAKE, and Foreverse.

What went into this report

This is a structured review of what other people tested and reported. It is not a new benchmark run and it does not treat provider marketing as independent proof of performance.

Shared tests	Common-task comparisons such as DocBench, ErrataBench, AA-Briefcase, Arena, STOCKTAKE, and live business-application evaluations provide the main quantitative comparisons.
Practitioner reports	Current users add evidence about everyday adoption, output quality, recognizable writing habits, workflow friction, and the amount of editing still required.
Official facts	Provider documentation establishes the exact model, API pricing, available inputs and outputs, context limits, and whether downloadable weights exist.
Failure reports	Negative findings remain visible. Strong average performance does not erase a repeated workflow failure, a calculation error, or a model-specific writing habit.

The review screened 82 source records and converted the admitted and retained findings into 168 traceable claims. The conclusion is a current business guide, not a permanent ranking.