baseten

SEPTEMBER 8, 2026

Executive Briefing on GLM-5.3

SPEAKERS

PHILIP KIELY

Author of Inference Engineering

DECLAN JACKSON

Member of Technical Staff, Artificial Analysis

baseten

POWERING LEADING COMPANIES

ABRIDGE CURSOR Decagon descript

Harvey HubSpot Lovable Notion

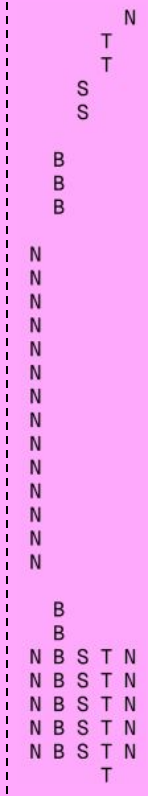
OpenEvidence parallel Vercel World Labs

Agenda

- GLM-5.3 model overview
- GLM-5.3 benchmarking
- Top use cases for GLM family
- Inference engineering for GLM-5.3
- Questions and discussion

SECTION 01.

GLM-5.3 model overview



Many model options

The market has increased in model availability

CLOSED MODELS OPEN MODELS

Frontier closed source models come to market

Labs release their first open weight LLMs

Open models become viable alternatives

CLAUDE 4
DEEPSEEK R1
GEMINI 3
GPT-5
LLAMA 4
MISTRAL 3
QWEN 3

Specialized models come onto the market

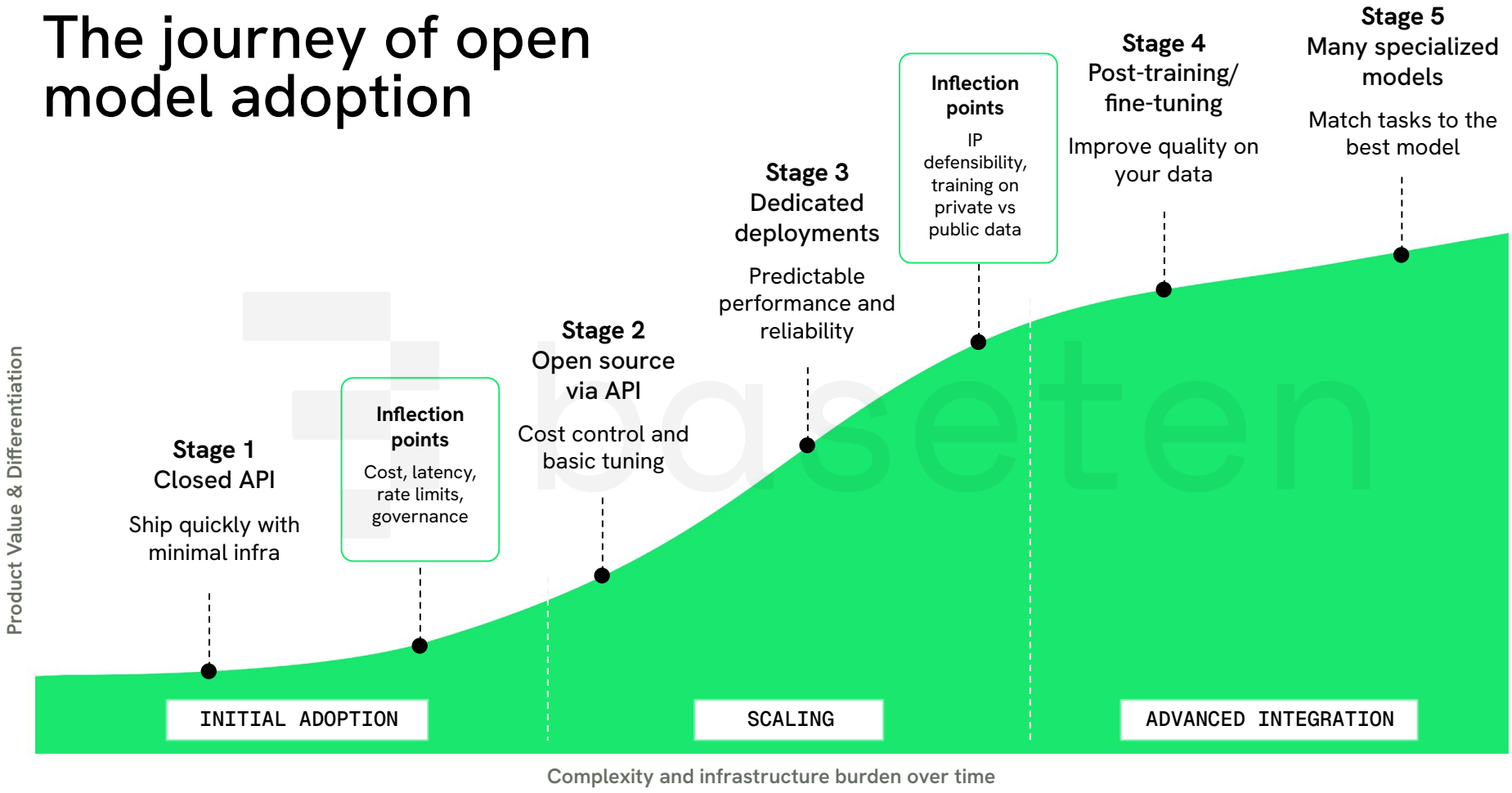
CODE
FINANCE
HEALTHCARE
LEGAL
SUPPORT
GEMINI 2.0
GPT-5.5
OPUS 4.8
FABLE 5
DEEPSEEK V4
KIMI-K2.6
NEMOTRON ULTRA
GEMMA 4

Open models are now frontier models

KIMI K3
GLM 5.3

2023 2024 2025 EARLY 2026 TODAY

The journey of open model adoption



GLM-5.3

RELEASED IN AUGUST

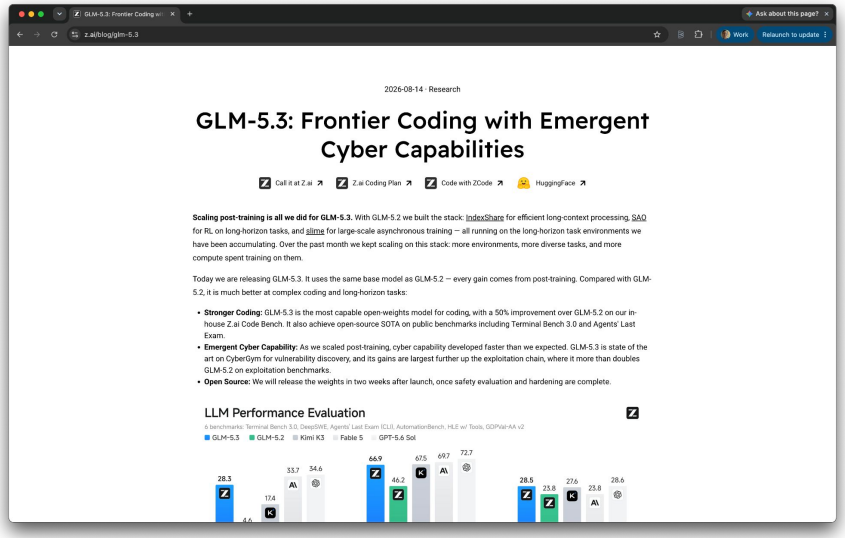
Z AI first released GLM-5.3 on their first-party platform and API, then opened up model weights two weeks later.

OX ALPHA AS FLASH

GLM-5.3 Flash was released as a stealth model Ox Alpha with generous free-tier usage on OpenRouter, generating hype for the main model release.

OPEN WEIGHTS

GLM-5.3 Flash is released under an MIT license, while GLM-5.3 includes usage restrictions for inference providers with $10B in revenue.



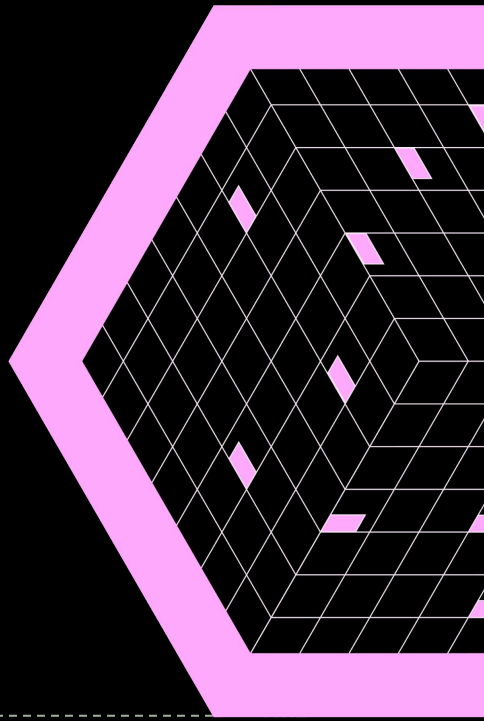
GLM-5.3 Flash		GLM-5.3	
320B	Total params	774B	Total params
18B	Active params	40B	Active params
42	Artificial Analysis overall intelligence score	45	Artificial Analysis overall intelligence score

SECTION 02.

GLM-5.3 Benchmarking

SPECIAL GUEST

Over to you, Declan



Artificial Analysis

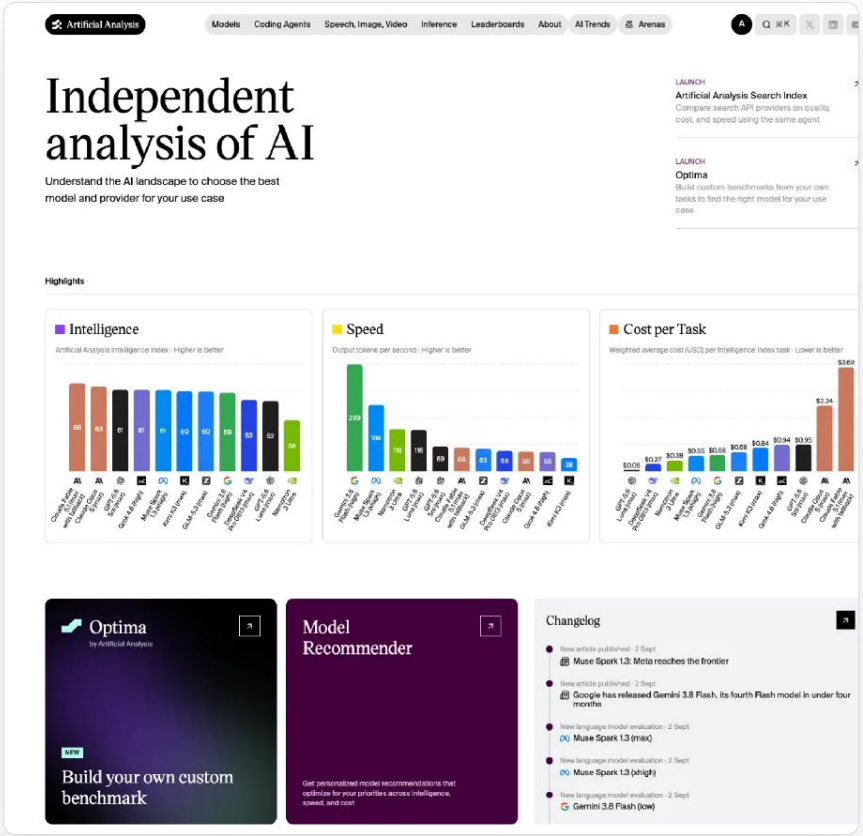
GLM-5.3

Independent evaluation of Z.ai's new frontier open weights models

SEPTEMBER 2026

ARTIFICIAL ANALYSIS

Artificial Analysis is the leading AI benchmarking company



WE BENCHMARK THE FULL AI STACK

Agents
Models paired with a harness and tools to complete multi-step tasks

Coding Agent Index

Models
Language, image, video, speech and music models, proprietary and open weights

Artificial Analysis Intelligence Index

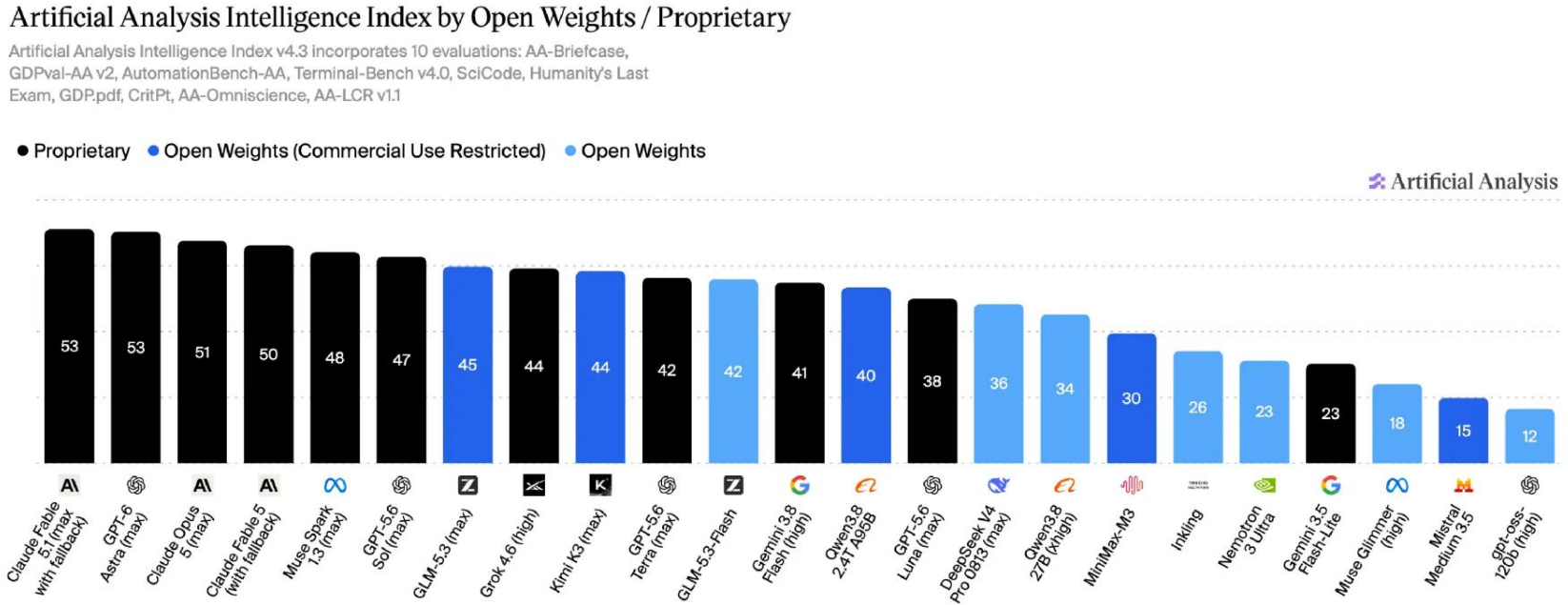
Cloud
The infrastructure that hosts and serves the models

Inference Providers Leaderboard

Chips
The accelerators the models run on

AA-AgentPerf

GLM-5.3 is the leading open weights model



Agents 30%
AA-Briefcase · GDPval-AA v2
AutomationBench-AA

Coding 20%
Terminal-Bench v4.0
SciCode

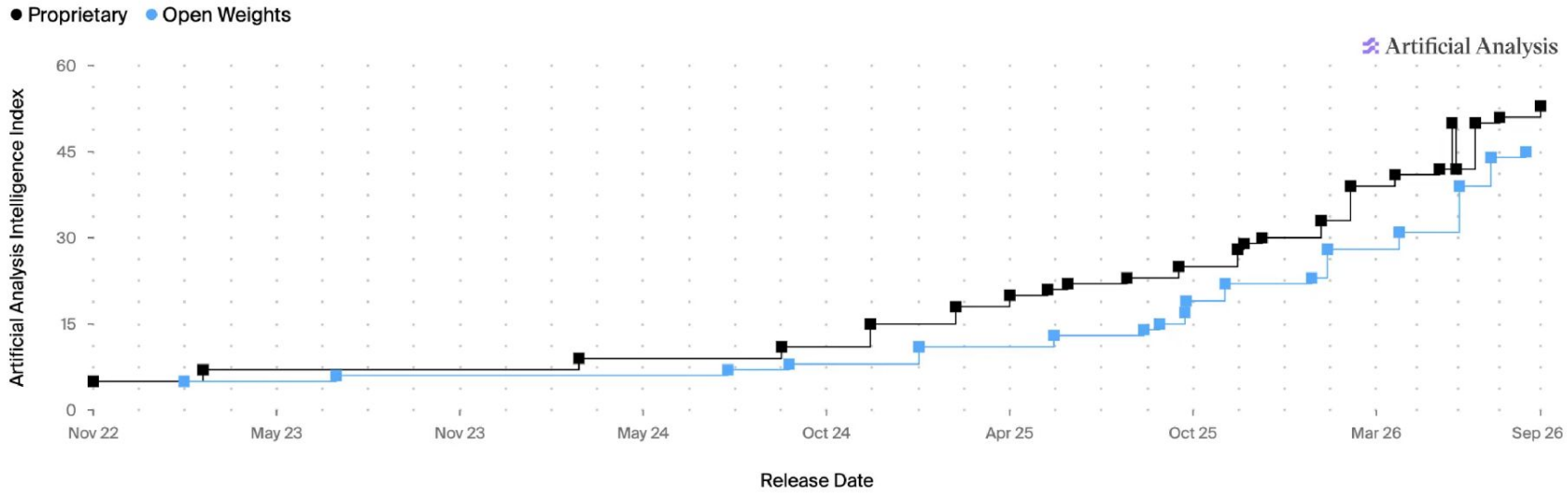
General 30%
AA-Omniscience Accuracy · Non-hallucination
GDP.pdf · AA-LCR v1.1

Scientific Reasoning 20%
Humanity's Last Exam
CritPt

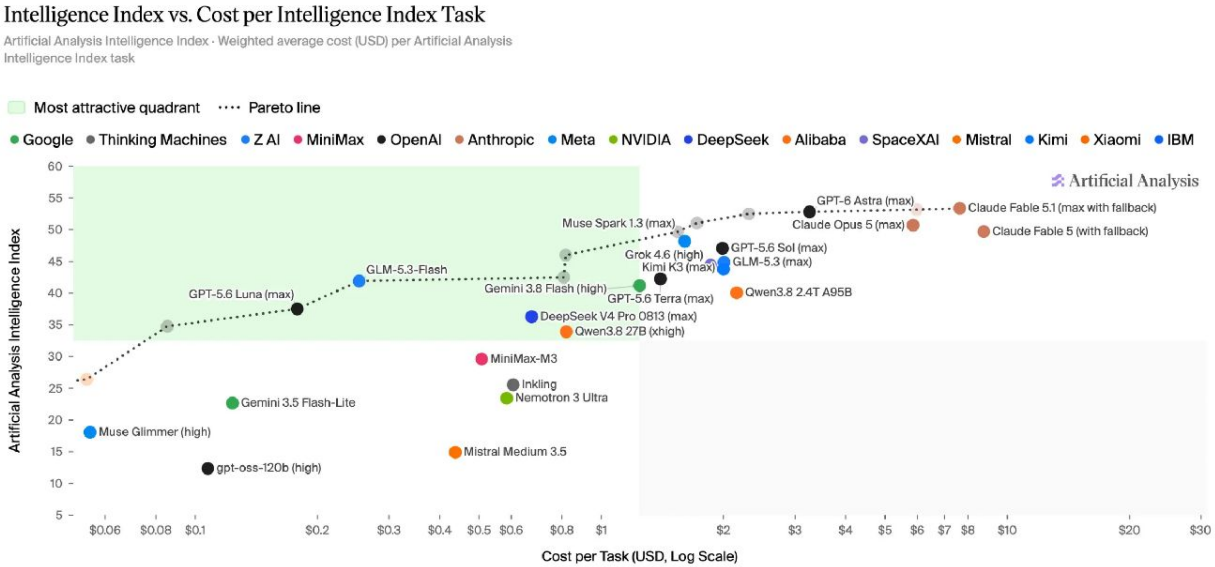
■ Open weights models are keeping within 3–9 months of the proprietary frontier

Progress in Open Weights vs. Proprietary Intelligence

Artificial Analysis Intelligence Index v4.3 incorporates 10 evaluations: AA-Briefcase, GDPval-AA v2, AutomationBench-AA, Terminal-Bench v4.0, SciCode, Humanity's Last Exam, GDP.pdf, CritPt, AA-Omniscience, AA-LCR v1.1

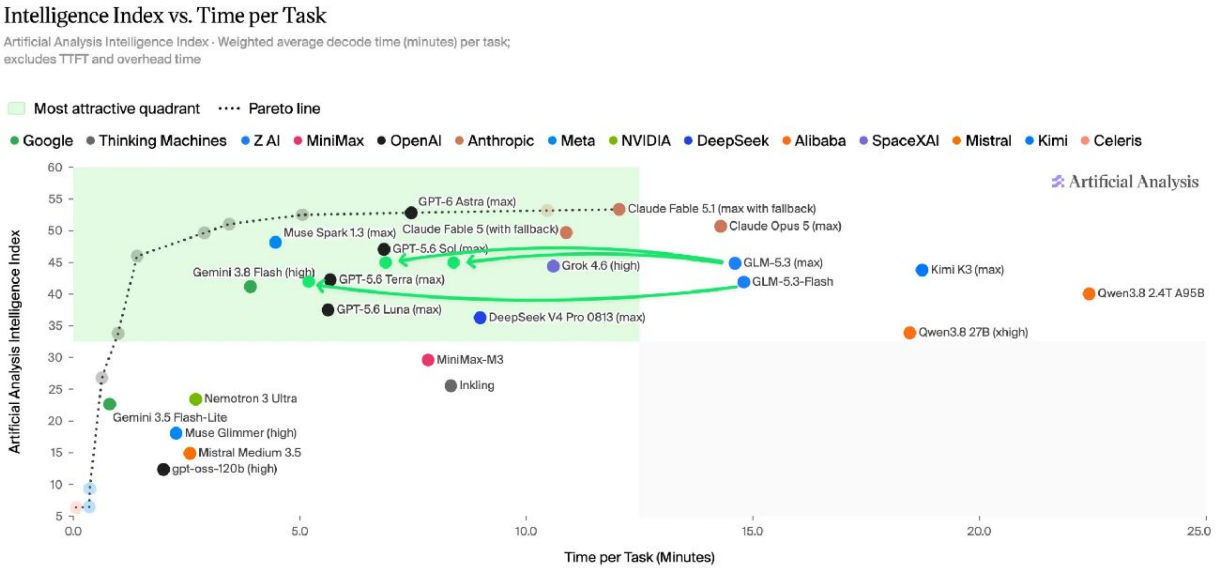


GLM-5.3-Flash is on the Cost per Task vs. Intelligence Pareto frontier



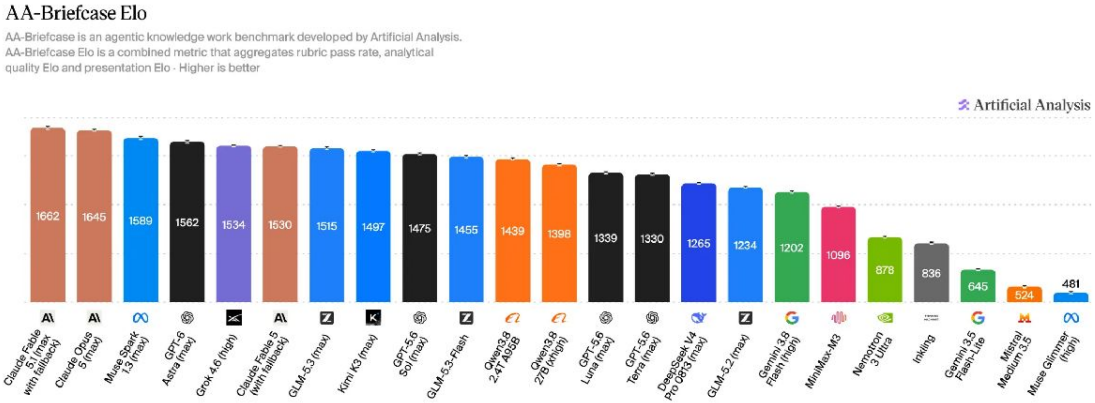
- **GLM-5.3-Flash is ~8× cheaper per task:** $0.25 per Intelligence Index task vs. $2.01 for GLM-5.3
- **Both models exhibit similar token use:** output per task is nearly identical (68.7K vs. 71.1K tokens, ~47K of it reasoning), and Flash reads more input (7.7M vs. 6.0M tokens), driven by a higher number of turns in agentic benchmarks
- **Price is the key driver of the cost per task difference:** across most providers (including Baseten) base GLM-5.3 is priced at $1.40/$4.40 per 1M in/out (same as 5.2), with an 81% cached input discount and 99.5% measured hit rate. Flash is $0.15/$0.50

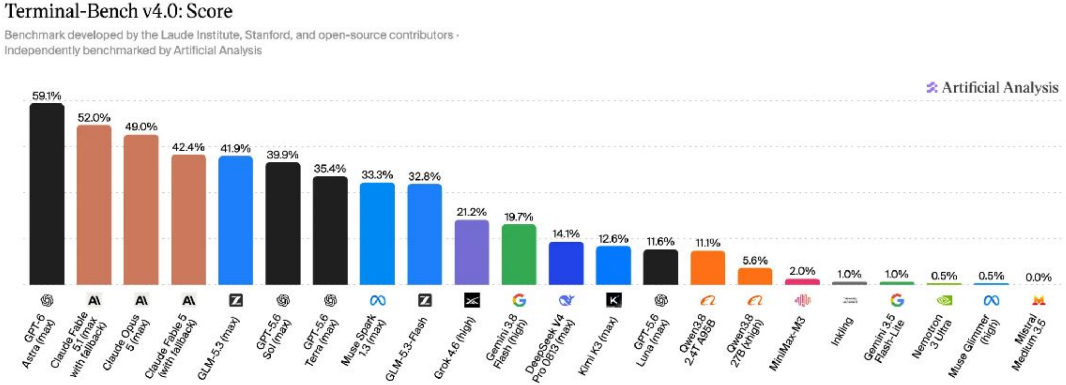
Time per Task varies nearly 15× across providers



- **Comparable Time per Task between GLM-5.3 and Flash:** GLM-5.3-Flash averages 14.8 minutes per task, while GLM-5.3 averages 14.6 minutes on the first-party endpoint, due to similar token use and endpoint speeds between the two models
- **Large variation across providers:** GLM-5.3 ranges from 2.8 to 24.1 minutes per task across 14 provider endpoints, an 8.6× spread. GLM-5.3-Flash ranges from 2.7 to 38.3 minutes across 17 endpoints, a 14.2× spread

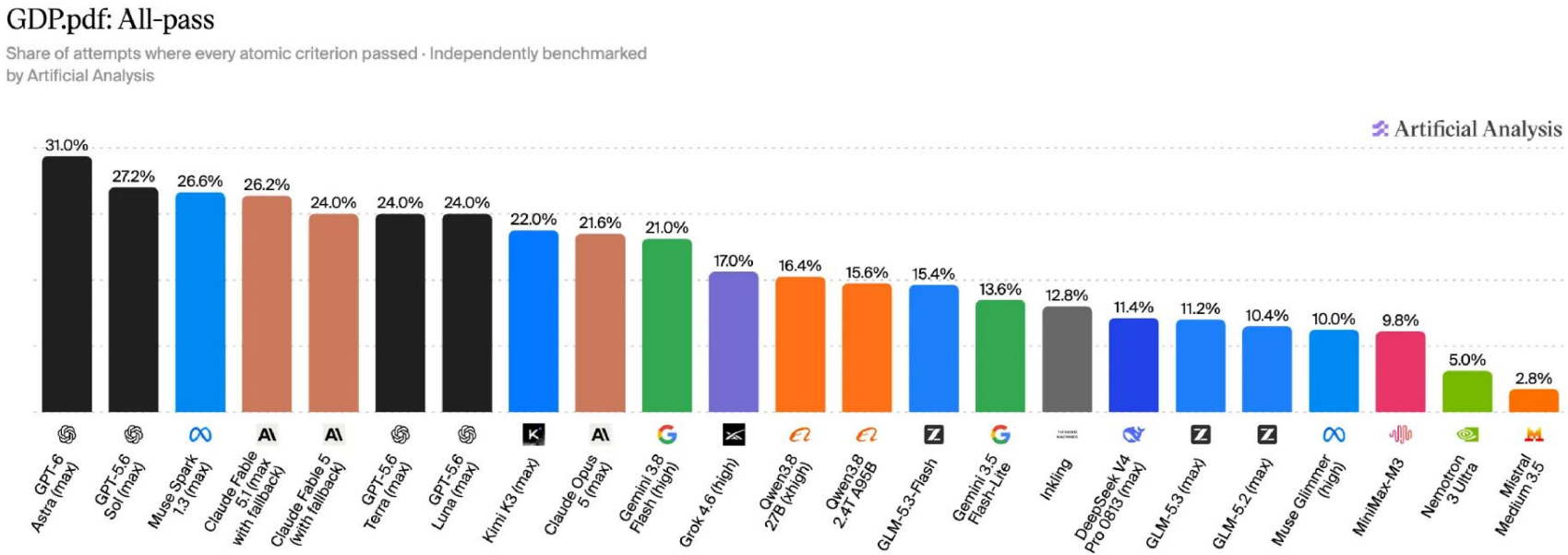
GLM-5.3 and GLM-5.3-Flash have strengths in agentic knowledge work and coding



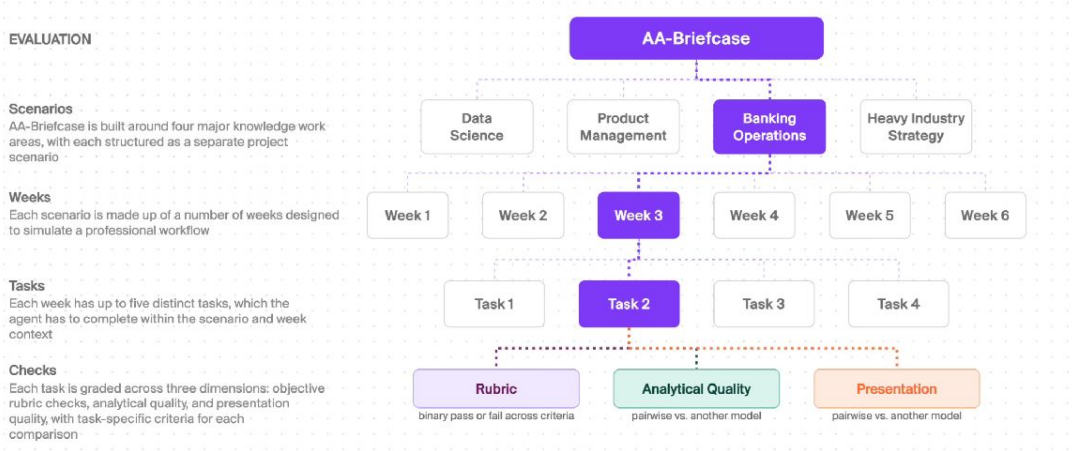


- **Strong performance on agentic knowledge work:** GLM-5.3 and Flash rank 6th and 7th of 130 models on GDPval-AA v2 (1676 and 1669 Elo). GLM-5.3 leads all open weights models on GDPval-AA v2, AA-Briefcase and AutomationBench-AA
- **GLM-5.3 excels for coding use cases:** 41.9% on Terminal-Bench v4.0, 1st among open weights models, vs. 32.8% for Flash. 59.0% on SciCode vs. 51.6%. These 9.1pp and 7.4pp gaps are the widest of any index component
- **Weaker on knowledge recall and document work:** AA-Omniscience accuracy is the lowest-ranked component for both models (33.9% and 27.5%), with non-hallucination helped by abstaining on roughly half of questions. Also weaker on GDP.pdf, AA-LCR and CritPt

■ GLM-5.3-Flash's multimodality makes it a stronger choice for visual reasoning tasks



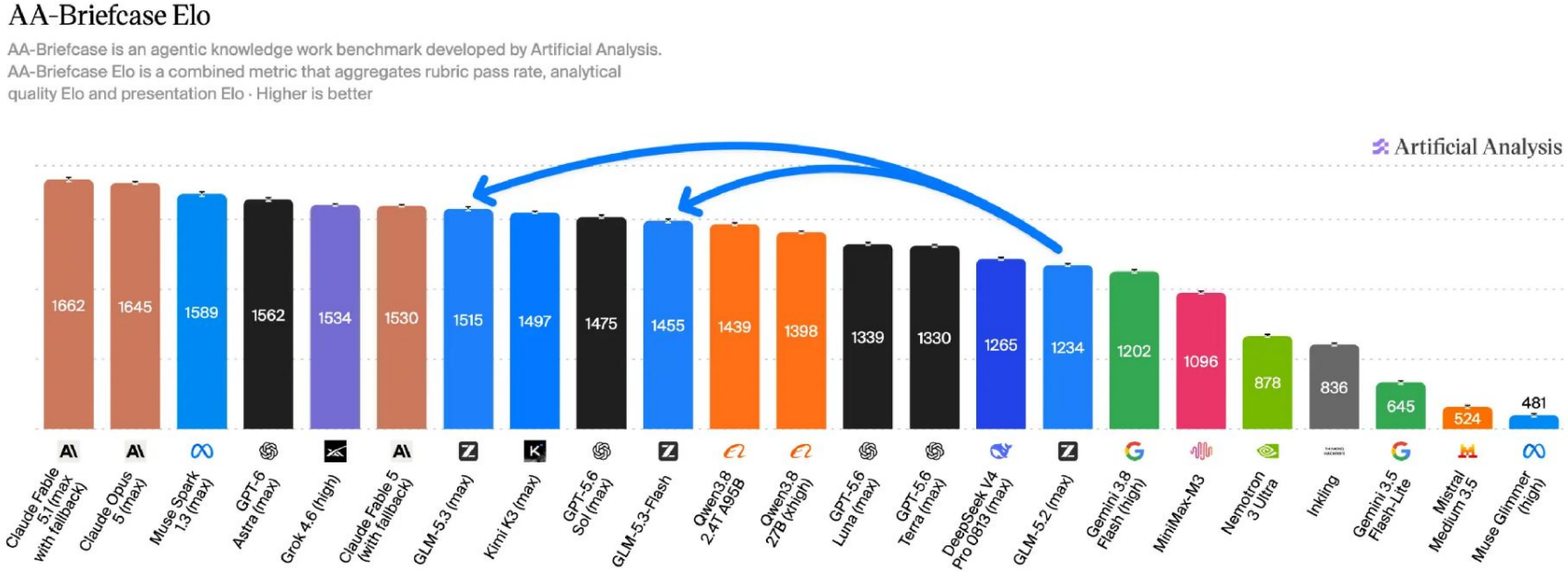
What is AA-Briefcase?



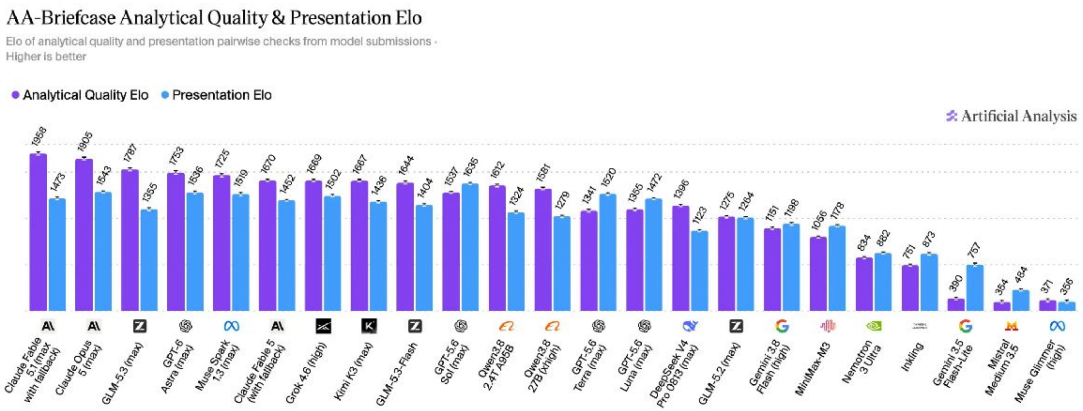
- **Realistic, long-horizon projects:** multi-week projects with real deliverables like financial models and board decks. Leading models spend 20+ minutes per task
- **High volumes of fragmented context:** ~2,000 source files including 3,500+ emails and 25,000+ Slack messages — fragmented and contradictory, like real work
- **Fully private dataset:** all 91 tasks, inputs and rubrics are private. A public sample, AA-Briefcase Lite, is on Hugging Face
- **Composite rubric and pairwise grading:** binary rubric checks for correctness, plus pairwise grading on analytical quality and presentation
- **Built by industry experts:** tasks built over months by experts from Google, McKinsey & Company and BCG
- **Reflects real-world complexity:** the top model meets every rubric criterion on just 3% of tasks; on 31 of 91, no model exceeds 50%

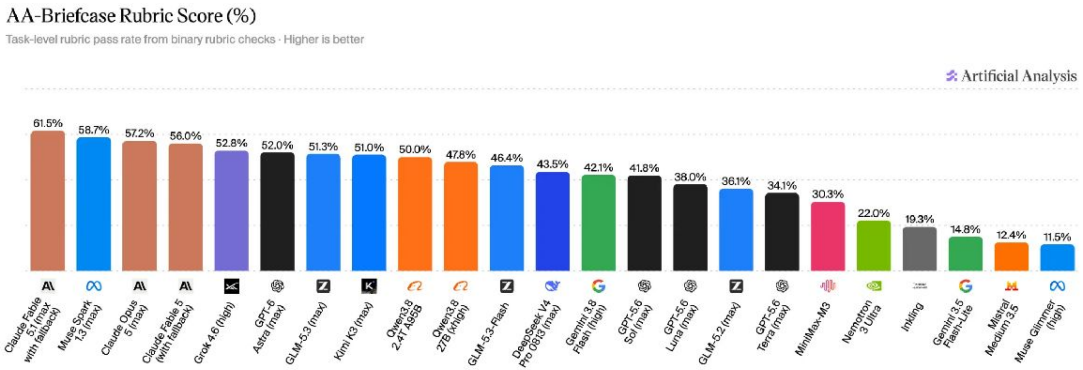
Every model runs in **Stirrup**, our open-source agent harness, with up to 500 turns per task. A panel of judges scores each submission on all three dimensions, combined into a single **AA-Briefcase Elo**.

■ GLM-5.3 gains +281 Elo on AA-Briefcase over GLM-5.2



Model capability differs across the dimensions of AA-Briefcase



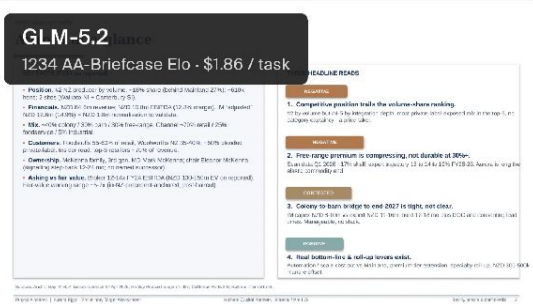


- **Large gains on AA-Briefcase:** GLM-5.3 improves from GLM-5.2's 1234 Elo to 1515, while GLM-5.3-Flash reaches 1455. The largest gains are in Analytical Quality, with GLM-5.3 up 512 Elo and Flash up 370. Rubric pass rate also rises from 36.1% to 51.3% and 46.4%
- **GLM-5.3-Flash leads on presentation:** Flash scores 1404 Presentation Quality Elo vs. 1355 for GLM-5.3, and is the only model to inspect its own outputs, making 1,631 view_image calls across 89 tasks. GLM-5.3 and GLM-5.2 make none
- **GLM-5.3 leads on substance:** it scores 142 Elo higher than Flash on Analytical Quality and has a higher rubric pass rate overall (51.3% vs. 46.4%) and across every file type. Flash also uses more turns, averaging 146.5 per task vs. 109.9 for GLM-5.3

The same task, one generation apart

Week 1, task 3: build a **25-slide pre-LOI target assessment** for a New Zealand shell-egg producer, from a fragmented public and sell-side evidence base. Two slides from each model's submission.

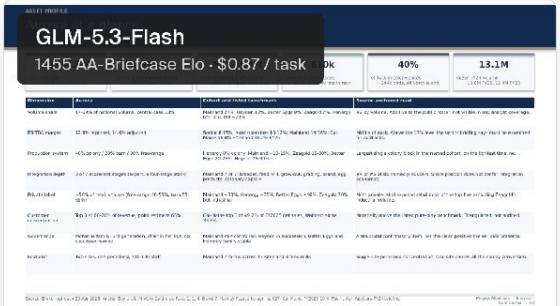
GLM-5.2
1234 AA-Briefcase Elo · $1.86 / task



ASSET PROFILE

Operating footprint, scale & channel mix

GLM-5.3-Flash
1455 AA-Briefcase Elo · $0.87 / task



ASSET PROFILE

Channel mix, customer concentration and retailer structure

GLM-5.3
1515 AA-Briefcase Elo · $5.65 / task

610k | **18%** | **13.1m** | **USD 50.9m** | **110-130** | **2**

Asset profile: channel mix, customer concentration and counterparty structure

Artificial Analysis

Artificial Analysis

Scan for the full analysis

artificialanalysis.ai/models/glm-5-3

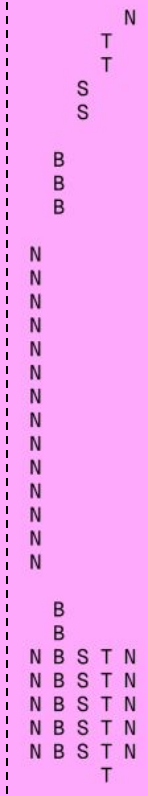
Optima

Benchmark GLM-5.3 on your own use cases

artificialanalysis.ai/optima

SECTION 03.

Top use cases for GLM-5.3



Low-cost coding with open models

CODEX AND CLAUDE CODE

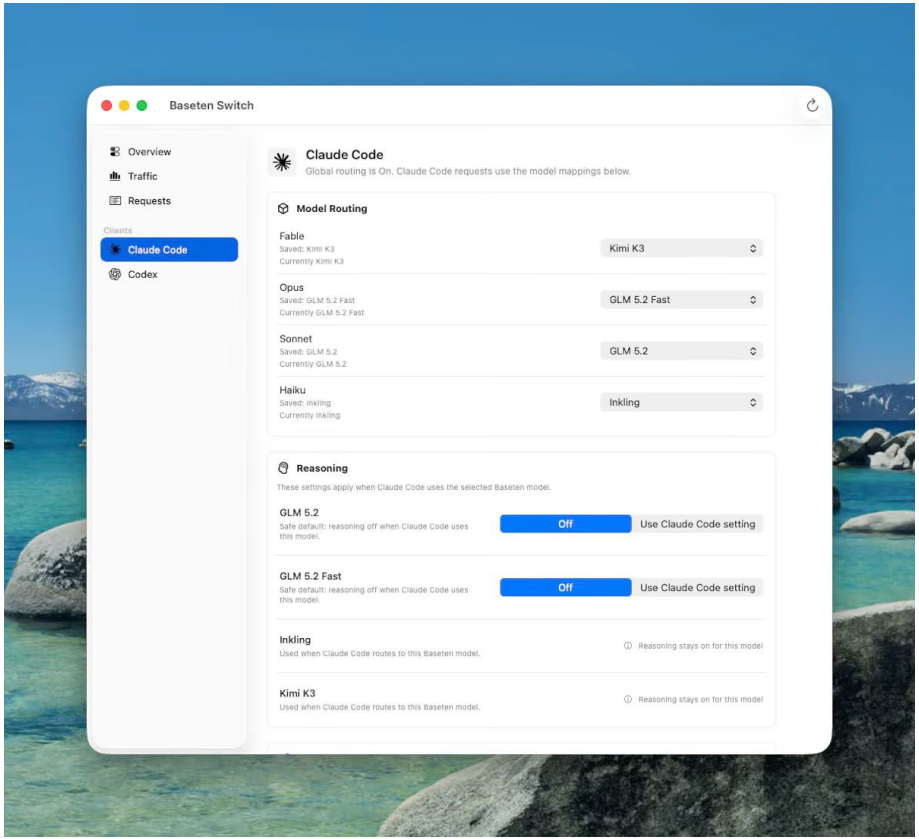
Use Baseten Switch to re-route calls from expensive closed models to low-cost open models without changing your harness

INDEPENDENT HARNESSES

A large number of excellent services like Cursor, Factory, Opencode, and dozens more offer direct access to open models or bring your own key options

TWO MODELS FOR BETTER OUTCOMES

Use GLM-5.3 for long-horizon tasks and main agent work management, let GLM-5.3 Flash handle simple subagent tasks at a lower cost.



AI products where you need control

OWN YOUR INTELLIGENCE

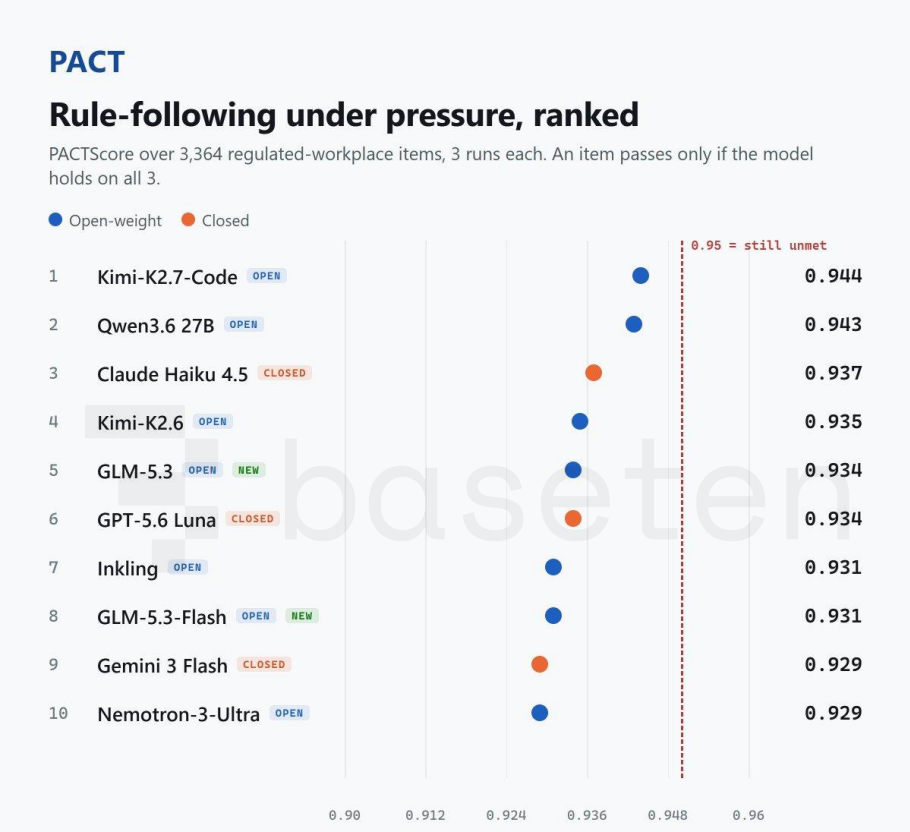
Run inference on GLM-5.3 and GLM-5.3 Flash for independence from large labs and zero data retention on Baseten's compliant infrastructure platform

FINE-TUNE BEHAVIOR

GLM 5.3 models are excellent bases for fine-tuning domain-specific knowledge and application-specific behavior into fast, inexpensive, generally capable models

PACT BENCHMARK

GLM-5.3 is the smartest model to get a strong score on the PACT benchmark, which measures if a model will refuse an out-of-policy step in a workflow when pressured



AI products operating at scale

WE HAVE CAPACITY

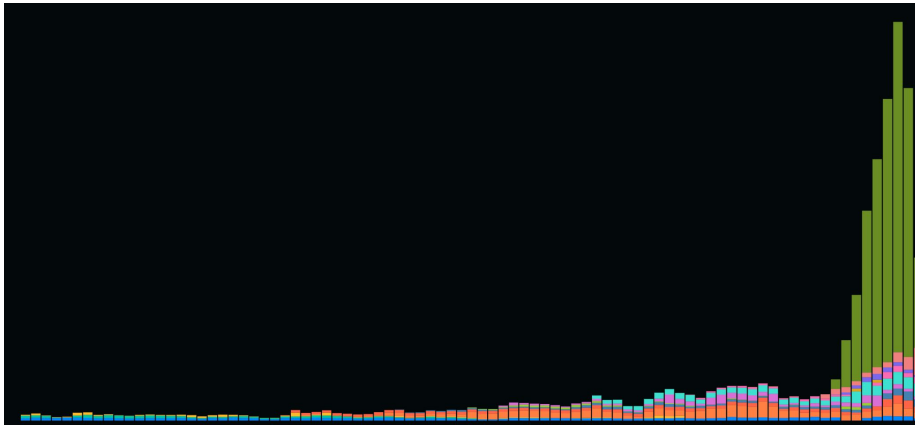
GLM-5.2 was a breakthrough model, and we estimated substantial demand for GLM-5.3, so we allocated a large amount of compute to the model

EFFICIENT SERVING

GLM 5.3 models are relatively stable to serve with a familiar architecture that benefits from multiple generations of inference optimizations

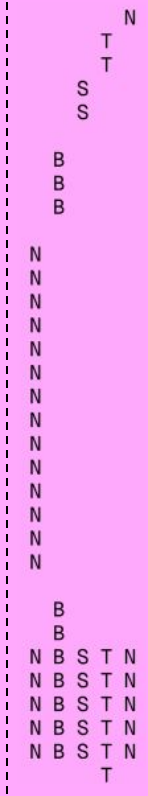
COST EFFECTIVENESS

GLM-5.3 Flash is ⅛ the cost of GLM-5.3 on a per-task basis making it an incredibly cost-effective choice for large-scale deployments



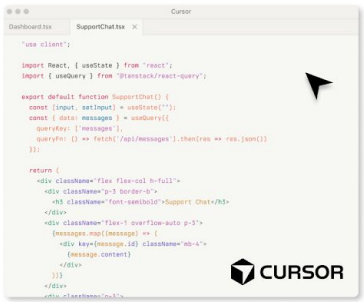
SECTION 04.

Inference engineering for GLM-5.3



Inference can mean success or failure for AI initiatives

Users engage with AI product



INFERENCE

Product value

GREAT INFERENCE

- Feels fast
- Stays reliable
- Improves quality
- Cost effective

POOR INFERENCE

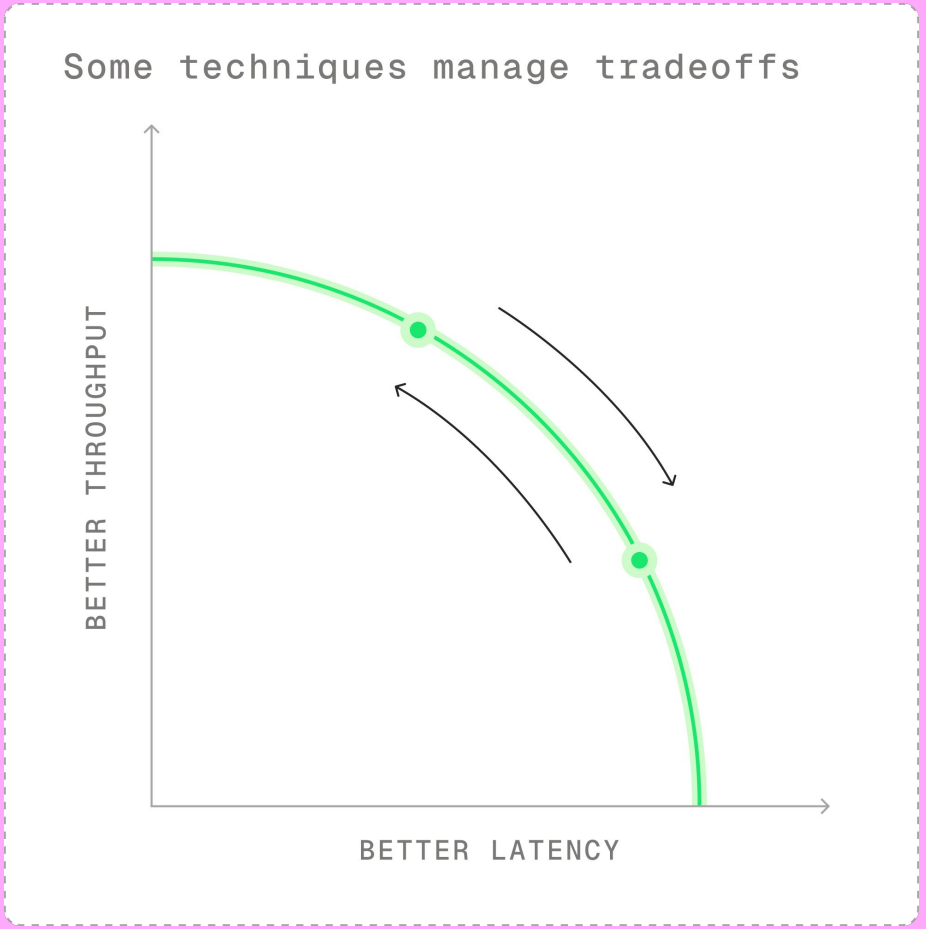
- Users churn
- Costs balloon
- Inconsistent
- Quality degrades

GLM-5.3 was created using **only post-training**

Efficient frontier in inference

INFERENCE IS ABOUT TRADEOFFS

- Lower latency
- Higher throughput
- Lower cost
- Higher quality



Fast mode for GLM-5.3 API

NO ADP

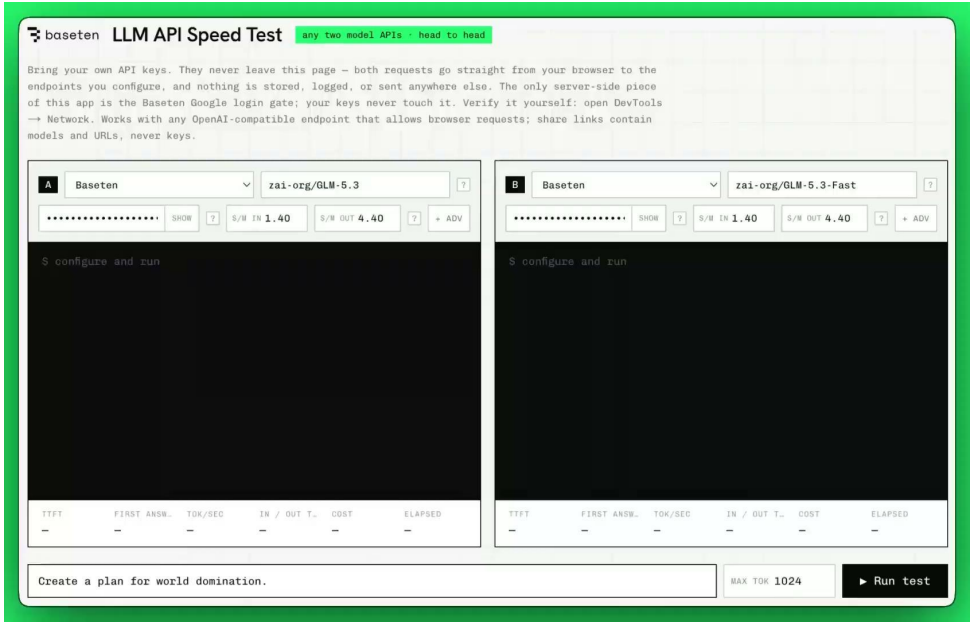
Remove Attention Data Parallelism and use solely Tensor and Expert Parallelism configurations

REDUCE BATCH SIZE

Reduce throughput for better latency by processing fewer requests simultaneously, increasing cost

TUNE ENGINE CONFIGS

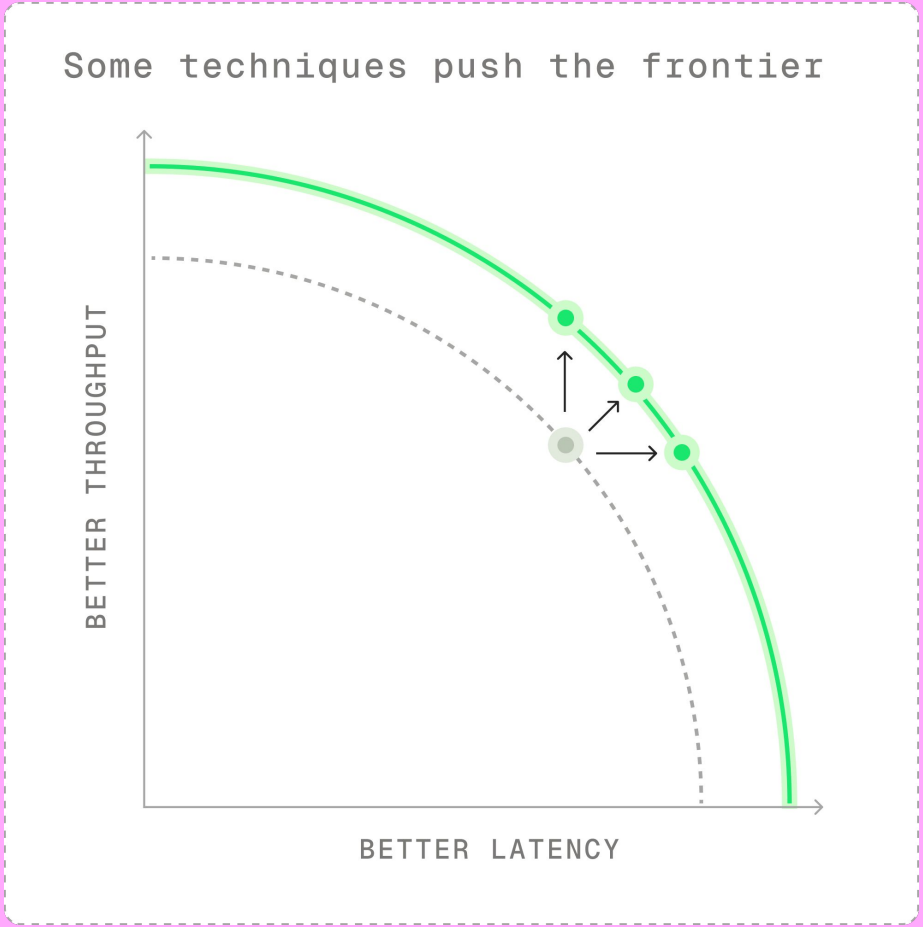
Re-run parameter sweeps across potential configurations to discover pareto-optimal latency-throughput tradeoffs



Optimization moves the frontier

ALLOCATE GAINS AS NEEDED

- Lower latency
- Higher throughput
- Combination



Vision for GLM-5.3

CUSTOM PROJECTOR

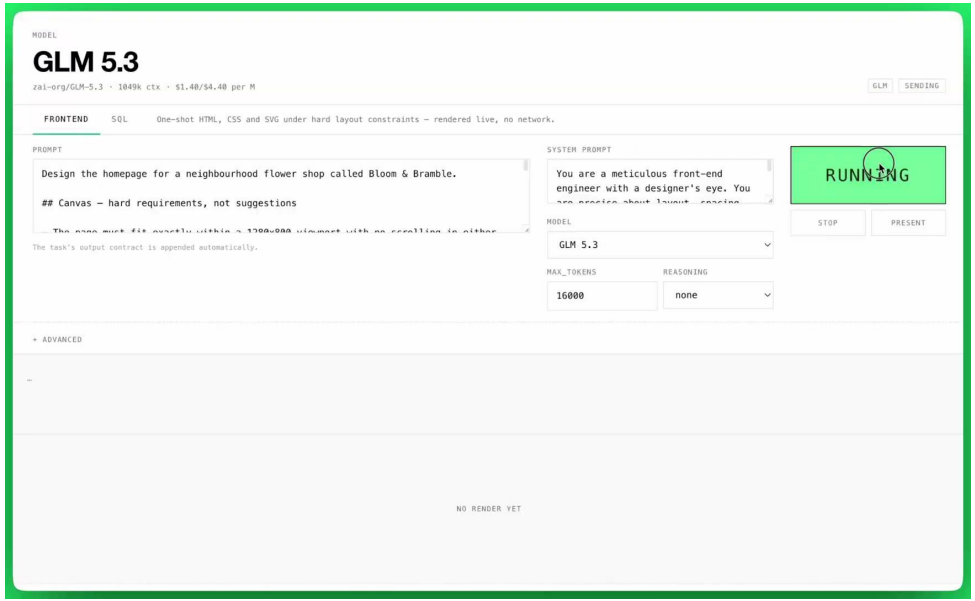
Train a projector from scratch to graft the Kimi K2.6 Vision Encoder onto the GLM-5.X architecture

NO CHANGE IN WEIGHTS

This approach means that the weights of the model are unchanged, so we do not see quality regressions across other domains

FLASH HAS DIFFERENT ARCHITECTURE

GLM-5.3 Flash has a different architecture from its larger cousin which includes a native vision encoder.



SECTION 05.

Questions and discussion

OPEN Q&A

What else would you like to know about GLM-5.3?