yak.

RESEARCH NOTE · NOVEMBER 2026

Confidential AI for regulated finance.

Capability, trust boundaries and economics of human-supervised agentic workflows that make organizations and their people more effective — from routine tasks to complex assignments.

Three findings.

1	The strongest orchestrators are hosted and closed; the best open one now runs on a confidential API, a step behind.	→ slides 06 · 17
2	Open-weight workers are good enough for bounded work, and can run inside your boundary.	→ slides 08 · 13
3	A hybrid lets a public orchestrator coordinate the work without seeing raw confidential data.	→ slide 18

Useful AI work inside financial institutions is held back by the trust boundary. The architecture can be changed so that the data boundary, not the model vendor, sets the design.

Where this matters in a bank.

- internal document analysis
- KYC case review
- reconciliation
- report drafting
- policy and legal research
- AML investigation support
- analyst copilots

Analyst-assist, not autonomous decisions.

EBA: 92% of EU banks already deploy AI; 55% use general-purpose or agentic AI in customer-facing processes · CCAF 2026: data privacy is the top-rated AI risk for 74% of firms and 80% of regulators

The obvious architecture is one frontier model for everything. It is often the wrong one.

Economics	Frontier models are expensive overkill for bounded calls: ~$634 per billion tokens at API rates against ~$41 for a capable worker.
Deployability	The strongest models are not available as weights you can place inside your own trust boundary.
Split the work	Separate planning from execution: a strong orchestrator decomposes the work; smaller workers execute bounded tasks.

Two building blocks.

Orchestrator

Larger, smarter model

Long-context, long-running work

High autonomy, reinforced by the harness

e.g. Opus 5.5 · Fable 5.2 · Gemini 4 Neon

Worker

Smaller model, fast inference

Bounded, well-defined tasks

Less GPU, far cheaper to run

e.g. Claude Sonnet 5.5 · Gemini 4 Hydrogen · GLM-5.3-Flash · DeepSeek V4.1 Flash

The top orchestrators are still closed. The best open model is now close behind on documents.

Model	Weights	AA-Briefcase	GDPval-AA	AA Index
Claude Opus 5.5 max	closed	1797	1866	59
Claude Fable 5.2 max	closed	1783	1830	65
Gemini 4 Neon xhigh	closed	1699	—	62
Qwen4 Max xhigh	closed	1685	1809	43
Kimi K3.1 Pro	open	1672	1856	47
GPT-6.2 Astra max	closed	1643	—	62
Gemini 4 Carbon high	closed	1611	1733	60
Qwen3.8-2.4T-A95B	open	1429	1613	—
DeepSeek V4 Pro max	open	1252	1455	23

Artificial Analysis, Nov 2026 · best published effort · AA-Briefcase and GDPval-AA are Elo ratings · Intelligence Index v5, rebased: not comparable with earlier versions · — = not evaluated

Pay-as-you-go, open orchestrators are cheapest. On a flat plan, closed ones are.

Model	Weights	~$200 plan, $/BTok	API $ / BTok
Gemini 4 Neon	closed	~$8.4	~$438
Gemini 4 Carbon	closed	~$11	~$584
Claude Opus 5.5	closed	~$20	~$634
DeepSeek V4 Pro *off-peak*	open	–	~$73
GPT-6.2 Astra	closed	~$82	~$1,267
Claude Fable 5.2	closed	~$131	~$1,079
Kimi K3.1 Pro *on Phala*	open	–	~$386

BTok = billion processed tokens · 100:1 input:output, 95% cached, cache writes where billed · Carbon at its price from 1 Dec 2026 (~$292 until then) · plan closest to $200, a per-seat subscription used through the provider's own apps, not the API; capacity estimated, not a quota; newer models scaled from their predecessor's estimate by list price · DeepSeek: pay-as-you-go only; Kimi: plan terms not verified

Pay-as-you-go, DeepSeek is the cheapest worker. On a flat plan, Gemini is.

Model	Weights	~$200 plan, $/BTok	API $ / BTok	Output tok/s
Gemini 4 Hydrogen	closed	~$4.2	~$219	163
GLM-5.3-Flash	open	~$5.0	~$41	58
Claude Sonnet 5.5	closed	~$6.3	~$317	154
GPT-6.1 Terra	closed	~$10	~$158	69
DeepSeek V4.1 Flash	open	—	~$16	175
Gemini 3.8 Flash	closed	—	~$145	101

DeepSeek at off-peak rates (~$32 peak) · Gemini 3.8 Flash at introductory price to 31 Dec 2026 (~$290 from 1 Jan 2027) · output speed: Artificial Analysis, provider API, best published effort · plans as on the previous slide; GLM plan = ZCode Max off-peak

DeepSeek now leads GLM on capability and price. GLM is still the easier one to run on premises.

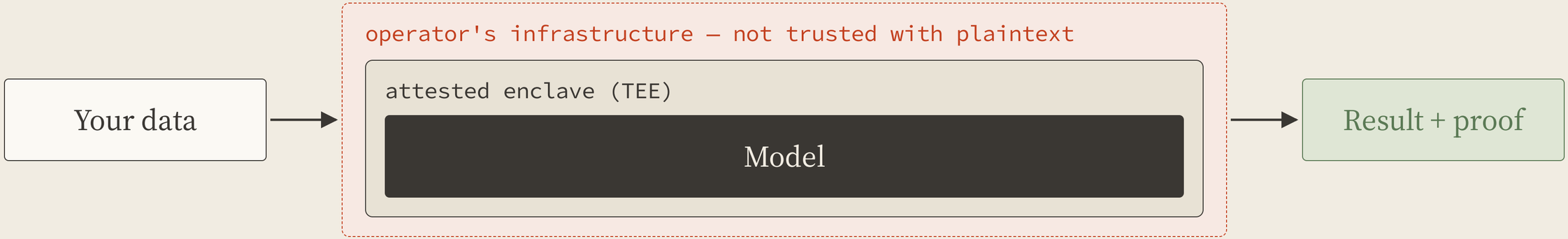
	GLM-5.3-Flash	DeepSeek V4.1 Flash
AA Intelligence Index	28	32
AA-Briefcase	1448	1570
GDPval-AA	1649	1746
API $/BTok	~$41	~$16 off-peak · ~$32 peak
Output tok/s	58	175
Parameters, total / active	320B / 18B	552B / 8B prefill, 16B decode
Memory, native checkpoint	~328 GB FP8	~510 GB FP8/FP4
Smallest private appliance †	2× DGX Spark, $9.4k	4× DGX Spark, $18.8k
Phala confidential API	$0.15/M input · ~$41/BTok	$0.30/M input · ~$32/BTok
Used as a worker in my own work	*in daily use*	not tried

same mix · Artificial Analysis, best published effort; DeepSeek scores are for its October checkpoint · † GLM served at 4-bit (W4A16) on 2× DGX Spark; DeepSeek at its native FP8/FP4 with the Engram table on NVMe, measured on the September checkpoint of the same architecture

The question is who must be trusted with data in use.

Hosted inference relies on the provider's controls and contracts to keep prompts and documents confidential while they are processed. Confidential computing replaces part of that trust with a boundary you can verify.

Confidential compute: the infrastructure operator need not be trusted with plaintext.



Managed confidential API

per token · no commitment

e.g. Phala

Confidential GPU VM or node

rented · hourly or reserved

e.g. Phala, Google Cloud, Microsoft Azure

alternative trust model: become the operator

Own hardware

on premises · own cache

upfront cost

Attestation: the enclave proves what hardware and software it runs before any key or data is released.

What confidential compute does — and does not — solve.

Does

protects data in use from the host and cloud operator

attestation verifies hardware and software state before secrets are released

data can be released only to a verified environment

no-retention and no-leakage become verifiable properties of the attested code, not contract promises

Does not

hallucinations

prompt injection

flawed application code

excessive network egress

compromised employee endpoints

authorization mistakes

governance and vendor risk

Confidential compute is a technical control, not a compliance certificate. Third-party and ICT-risk obligations (e.g. DORA in the EU) still apply.

At sufficient utilization, owned hardware can undercut the confidential API on compute cost.

GLM-5.3-Flash on	Upfront	$/BTok, fresh	$/BTok, 95% cached	Privacy
Z.ai API	$0	$153.5	$40.6	cloud
Phala confidential API	$0	$153.5	$40.6	TEE, attested
2× DGX Spark	$9.4k	~$216	~$42	on premises
2× RTX Pro 6000	~$38k	~$212	~$29	on premises
DGX Station GB300	$95k	~$122	~$30	on premises

compute economics, not full TCO — excludes redundancy, support, operations and spares · 100:1 input:output · 4-year life, 30% utilization, $0.15/kWh · cached rows modeled, ±25–40%

Break-even: a pair of Sparks pays off at a few billion tokens.

5.2B tokens a month, and a 2× DGX Spark ($9,398) beats the confidential API in an agentic workflow — about 390,000 pages of new documents a month. With nothing to reuse, from 1.4B.

	95% cached · *agentic workflow*		no caching · *unrelated queries*	
System	break-even	max load	break-even	max load
2× DGX Spark	5.2B/mo	17B/mo	1.4B/mo	3.3B/mo
2× RTX Pro 6000	20B/mo	98B/mo	5.6B/mo	13B/mo
DGX Station GB300	50B/mo	224B/mo	13B/mo	55B/mo

max load = capacity at full utilization · vs Phala at the same cache share ($40.6/BTok at 95% cached, $153.5 with none) · compute economics, not full TCO · *modeled*

Qwen3.8-Flash-Next: DeepSeek's document score on a single Spark.

Models evolve fast. This open model scores like DeepSeek V4.1 Flash on AA-Briefcase, the benchmark closest to document and compliance work, and fits one DGX Spark. It is far less proven in practice than GLM-5.3-Flash, and its overall index is GLM's: more experimentation is due.

Artificial Analysis	Flash-Next	GLM Flash	DeepSeek
AA-Briefcase	1568	1448	1570
GDPval-AA	1637	1649	1746
AA Intelligence Index	28	28	32
Output tokens, index run	156M	124M	174M
Output tok/s	74	58	175

Qwen3.8-Flash-Next on	Upfront	$/BTok
1× DGX Spark *(n-gram on NVMe)*	$4.7k	~$11
2× DGX Spark *(table in memory)*	$9.4k	~$13
Hosted API (Alibaba)	$0	~$30
Phala confidential API	—	*not offered*

Measured on an RTX 4060 laptop, 8 GB	output tok/s
2-bit (Strata), experts in system RAM	22.3

HYPOTHESIS · *Flash-Next on document judgment, GLM or DeepSeek on execution — private peers checking each other, no public orchestrator. Untested.*

open-weight preview of the Qwen4 architecture · Artificial Analysis, best published effort · single-Spark cost from NVIDIA's official NVFP4 recipe

Flint: a coding worker small enough for the analyst's laptop.

On the device, the laptop itself is the trust boundary. Scores are the vendor's; laptop speeds were measured.

Vendor-reported	full weights	4-bit (NVFP4)
Terminal-Bench 2.1	54.9%	53.7%
SWE-bench Verified	74.2%	74.1%
LiveCodeBench v6	80.3%	80.8%
IFEval	95.1%	94.6%

Measured on an RTX 4060 laptop, 8 GB	output tok/s
Flint, 2-bit (FreeToken)	53.8
Flint, 4-bit (NVFP4)	33.1

Footprint	
Parameters	39.8B total / 2.5B active
Weights	79.6 GB full · 24.20 GB 4-bit file
Speed (vendor)	125+ tok/s on an M4 Pro laptop · 150+ tok/s on an RTX Spark laptop (from $2,599)
License	Apache 2.0, open weights

UNVERIFIED · *Tuned for coding, not documents, and not yet on Artificial Analysis. Worth a benchmark on your own work, not a recommendation today.*

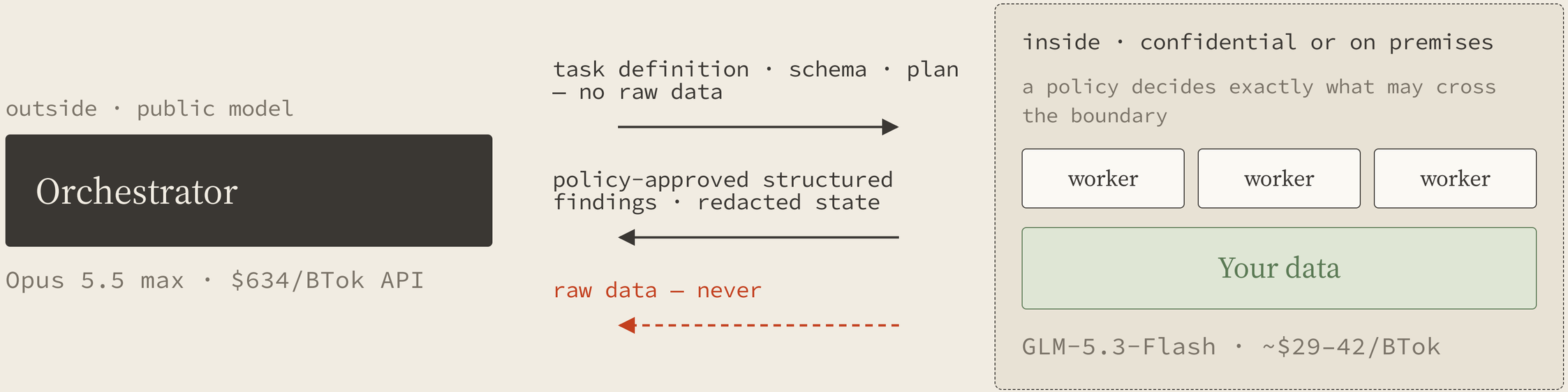
Recursive, released in October · vendor scores and speeds from the model card; 4-bit loses at most about a point · laptop: own runs, experts in system RAM, one request at a time, median over a mixed task set (Flint 60.1 on a fixed prompt)

A confidential orchestrator now exists: Kimi K3.1 Pro on Phala, a step behind Opus.

Model	AA-Briefcase	GDPval-AA	AA Index	Runs on
Claude Opus 5.5 max	1797	1866	59	hosted only
Kimi K3.1 Pro	1672	1856	47	Phala confidential API, or 8× B300 node, $300–690k
Qwen3.8-2.4T-A95B	1429	1613	—	8× B300 node, $300–690k
Qwen3.8-Flash-Next	1568	1637	28	1× DGX Spark, $4.7k
workers, for comparison				
GLM-5.3-Flash max	1448	1649	28	2× DGX Spark, $9.4k
DeepSeek V4.1 Flash max	1570	1746	32	4× DGX Spark, $18.8k

Kimi K3.1 Pro on Phala: ~$386/BTok on this deck's mix · the 8× B300 price is a purchase quote range; no owned-node $/BTok is modeled (no prefill/decode split published) · — = not evaluated

For tasks a confidential worker cannot solve alone: a public orchestrator coordinates, private workers touch the data.



Use AI on the work your data policy keeps *off-limits*.

INSIDE YOUR BOUNDARY

Confidential compute or your own hardware — verified by attestation, not by contract alone.

SIZED TO THE TASK

The smallest setup that passes your own benchmark, not a frontier model by default.

MEASURED

Productivity against a baseline, at the lowest cost your quality and security bar allow.

How we work.

1	Map the workflow	Which work, which data, and what must never leave.
2	Benchmark the minimum architecture	Your representative work decides between one worker, several, or a frontier orchestrator. Leaderboards only screen.
3	Choose the trust boundary and deployment	Managed confidential API, confidential VM or node, or on premises.
4	Integrate, attest and measure	Attested deployment, policy on what crosses the boundary, and productivity measured against a baseline.

Start with one workflow that is valuable but currently blocked by data policy.

On the numbers.

estimated	Prices from November 2026 rate cards. Subscription capacities are estimates, never quotas; newer models are scaled from their predecessor's estimate by list price.
one mix	Headline figures assume 100:1 input to output with 95% of input cached; the break-even slide also shows no caching.
modeled, not measured	Owned-hardware costs are a capacity model, ±25–40% until benchmarked on real traffic.
time-limited	Gemini 3.8 Flash is at its introductory price; Gemini 4 Carbon is shown at its December price.

yak.

Software built with machines, checked by people.

yak.ad · yak.finance