yak.

RESEARCH NOTE · SEPTEMBER 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.	→ slides 06 · 16
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 17

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 5.1 · GPT-6 Astra

Worker

Smaller model, fast inference

Bounded, well-defined tasks

Less GPU, far cheaper to run

e.g. Claude Sonnet 5 · Gemini 3.8 Flash · GLM-5.3-Flash · DeepSeek V4.1 Flash

The top orchestrators are all closed. Open-weight ones trail well behind.

Model	Weights	AA-Briefcase	AA Index	Terminal-Bench 4	AA-LCR
Claude Opus 5.5 max	closed	1822	58	59.6%	—
Claude Fable 5.1 max	closed	1678	53	52.0%	85.3%
Qwen3.8 Max 0902	closed	1640	~45	~39%	~80%
GPT-6 Astra max	closed	1569	53	59.1%	80.7%
Kimi K3 max	open	~1490	~44	~13%	88.7%
GPT-5.6 Sol max	closed	1487	47	39.9%	84.0%
Qwen3.8-2.4T-A95B	open	1445	40	11%	80%
DeepSeek V4 Pro	open	1261	36	14%	80%

Artificial Analysis, Sept 2026 · AA-Briefcase is an Elo rating · — = not reported

On price, closed and open overlap.

Model	Weights	~$200 plan, $/BTok	API $ / BTok
Claude Opus 5.5	closed	~$20	~$634
GPT-5.6 Sol *promo*	closed	~$52	~$822
Kimi K3	open	~$69	~$579
DeepSeek V4 Pro *off-peak*	open	—	~$73
Qwen3.8 Max	closed	~$123	~$394
GPT-6 Astra	closed	~$133	~$2,054
Claude Fable 5.1	closed	~$164	~$1,349

BTok = billion processed tokens · 100:1 input:output, 95% cached, cache writes where billed · Sol at promotional price to at least 21 Nov 2026 (~$1,077 at launch price) · plan closest to $200 (Kimi: $199), a per-seat subscription used through the provider's own apps, not the API; capacity estimated, not a quota · DeepSeek: pay-as-you-go only

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

Model	Weights	~$200 plan, $/BTok	API $ / BTok	Decode tok/s
Gemini 3.8 Flash	closed	~$2.8	~$145	115–170
Claude Sonnet 5	closed	~$8.1	~$411	60–100
GLM-5.3-Flash	open	~$9.9	~$41	90–115
DeepSeek V4.1 Flash	open	—	~$16	100–240
GPT-5.6 Terra	closed	~$27	~$431	45–70
Kimi K3	open	~$69	~$579	17–20
Qwen3.8 Max	closed	~$123	~$394	55–80

DeepSeek at off-peak rates (~$32 peak) · Gemini at introductory price to 31 Dec 2026 (~$290 from 1 Jan 2027) · API latency, not full agent turns · varies with provider, cache and load · plans as on the previous slide; GLM plan = ZCode Max off-peak

GLM matches DeepSeek — available via a confidential API, and easier to run on premises.

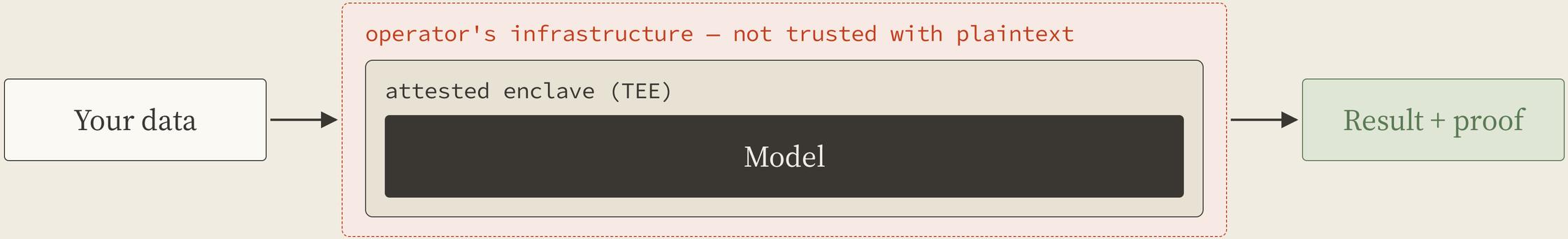
	GLM-5.3-Flash	DeepSeek V4.1 Flash
Bounded avg.†	40.4	40.4
API $/BTok	~$41	~$16 off-peak · ~$32 peak
Decode tok/s	90–115	100–240
First token	2–3 s	0.8–1.7 s
Parameters, total / active	320B / 18B	552B / 8B prefill, 16B decode
Memory, native checkpoint	~328 GB FP8	~510 GB FP8/FP4
Smallest private appliance ‡	2× DGX Spark, $9.4k	4× DGX Spark, $18.8k
Confidential API	Phala, $0.15/M input	$0.44/M input§
Used as a worker in my own work	*in daily use*	not tried

same mix · † mean of SciCode · HLE · GDP.pdf · CritPt · AA-LCR · ‡ GLM served at 4-bit (W4A16) on 2× DGX Spark; DeepSeek at its native FP8/FP4 with the Engram table on NVMe · § Phala's price for DeepSeek V4 Flash 0731, an older model of the same family; V4.1 Flash is not offered

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	$153.5	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: on premises wins from very modest workloads.

1.3B tokens a month, and a 2× DGX Spark ($9,398) beats the confidential API. In an agentic workflow that is about 97,000 pages of new documents a month.

	95% cached · *agentic workflow*		no caching · *unrelated queries*	
System	break-even	max load	break-even	max load
2× DGX Spark	1.3B/mo	17B/mo	1.4B/mo	3.3B/mo
2× RTX Pro 6000	5.2B/mo	98B/mo	5.6B/mo	13B/mo
DGX Station GB300	13B/mo	224B/mo	13B/mo	55B/mo

max load = capacity at full utilization · vs Phala $153.5/BTok, no cache discount · compute economics, not full TCO · *modeled*

Qwen3.8-Flash-Next: not an orchestrator, but a peer to GLM with a different profile.

Models evolve fast. On preliminary data this open model leads on AA-Briefcase, the benchmark closest to document and compliance work, yet it is far less proven in practice than Opus or GLM-5.3-Flash. It shows traits of both worker and orchestrator: more experimentation is due.

Artificial Analysis	Flash-Next	GLM Flash	Astra max
AA-Briefcase	1597	1451	1569
GDPval-AA	1612	1641	1542
AA Intelligence Index	40	42	53
Terminal-Bench 4	25%	33%	59.1%
AutomationBench	56%	60%	68%
Output tokens per task	108k	69k	27k

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*

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

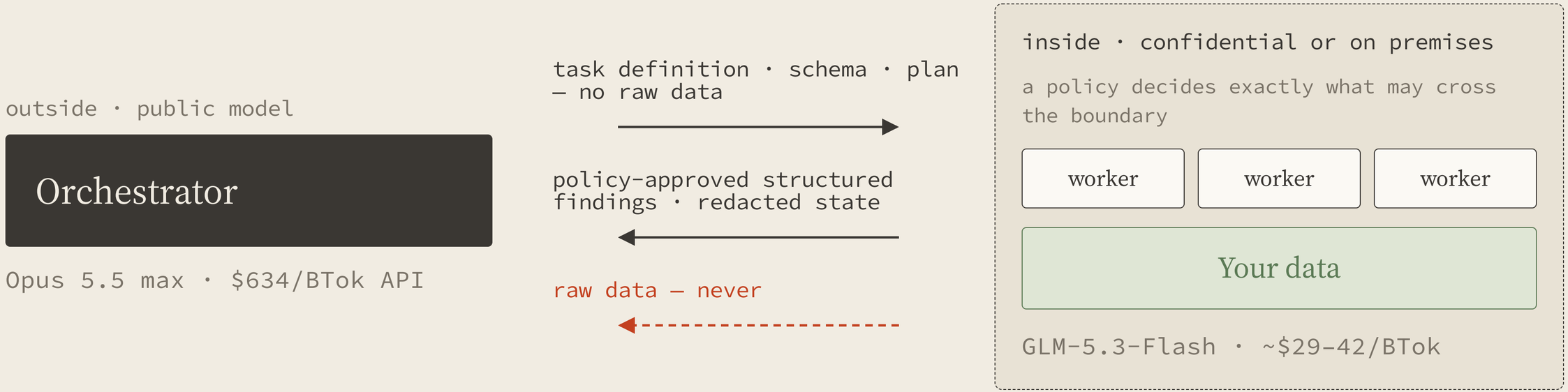
open-weight preview of the Qwen4 architecture · single-Spark cost from the single-Spark setup, NVIDIA's official NVFP4 recipe

No practical confidential orchestrator is much smarter than the workers.

Model	AA-Briefcase	AA Index	Terminal-Bench 4	Runs on
Claude Opus 5.5 max	1822	58	59.6%	hosted only
Qwen3.8-2.4T-A95B	1445	40	11%	8× B300 node, $300–690k
Kimi K3 max	~1490	~44	~13%	8× B300 node, $300–690k
Qwen3.8-Flash-Next	1597	40	25%	1× DGX Spark, $4.7k
workers, for comparison				
GLM-5.3-Flash	1451	42	33%	2× DGX Spark, $9.4k
DeepSeek V4.1 Flash	1424	39	27%	4× DGX Spark, $18.8k

Phala's confidential API serves none of these open models · the 8× B300 price is a purchase quote range; the Fixstars runs were not split into prefill and decode, so no $/BTok is modeled

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 September 2026 rate cards. Subscription capacities are estimates, never quotas.
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 and GPT-5.6 Sol prices are introductory or promotional.

yak.

Software built with machines, checked by people.

yak.ad · yak.finance