AgenticAssure Research

RESEARCH SERIES · ENTERPRISE AI BRIEFING · Nº 1

The Chinese LLM Opportunity

Cost, Capability & Trust

DeepSeek, Kimi K3, and Qwen are entering enterprise shortlists for one reason: the economics are hard to ignore. This paper answers the three questions every buying committee asks. Is it really cheaper? Is it really comparable? And can you trust it with your data?

July 2026

Compiled from vendor-published pricing, lab technical reports, and Gartner research

agenticassure.ai

Independent · Evidence-based · Defensible

Contents

ABOUT THIS REPORT

Prepared by AgenticAssure Research, July 2026. Pricing figures were fetched directly from each vendor's published pricing pages, not from third-party aggregators. Benchmark figures are self-reported by model developers unless otherwise noted, and are labelled as such throughout. An interactive edition with live cost calculators is available at **agenticassure.ai/reports**.

INTENDED READERS

CFOs and procurement leads evaluating AI spend; CTOs and platform teams weighing self-hosting; AI governance leaders assessing jurisdictional risk. This paper is not legal advice. Consult counsel on regulatory questions specific to your data and sector.

EXECUTIVE SUMMARY

Cheaper is proven. Comparable is close. Trust is the real decision.

IS IT REALLY CHEAPER?

Yes, 5–60×

On list price per token, DeepSeek and Qwen undercut US frontier tiers by an order of magnitude. The gap survives real workload maths.

IS IT REALLY COMPARABLE?

On many tasks

Lab-reported benchmarks put Kimi K3 and DeepSeek V4 near the frontier on coding and reasoning. Independent validation is thinner, so test on your own tasks.

CAN YOU TRUST IT?

Depends on the path

Hosted Chinese APIs carry documented jurisdictional exposure. Open weights, self-hosted in your own cloud, remove most of it.

1 **The cost gap is structural, not promotional.**

A representative workload of 5M input and 2M output tokens per month costs roughly $4 on DeepSeek V4-Pro against $75 on Claude Opus 5. Architectural efficiency (sparse activation, cheap KV cache) drives the difference, so it is unlikely to close quickly.

2 **Capability claims are lab-reported and should be re-verified quarterly.**

Kimi K3 and DeepSeek V4 publish near-frontier results on agentic and reasoning benchmarks. Gartner's independent read is that Chinese models are “closing the gap with speed to release.” Four major releases in 18 months means any static comparison ages fast.

3 **The trust question separates the hosted API from the open weights.**

Six governments restrict hosted DeepSeek. Every restriction on record targets the hosted app or API, not the open-weight model run independently. MIT and Apache licences let you run DeepSeek, Kimi, and Qwen entirely inside your own AWS, Azure, GCP, or on-prem estate.

4 **For most enterprises the sensible first move is a contained pilot.**

Run a Chinese open-weight model on non-sensitive workloads, benchmarked head-to-head against your incumbent. The decision framework in Section 05 maps four questions (data sensitivity, regulatory exposure, self-hosting capacity, cost pressure) to a starting position.

01

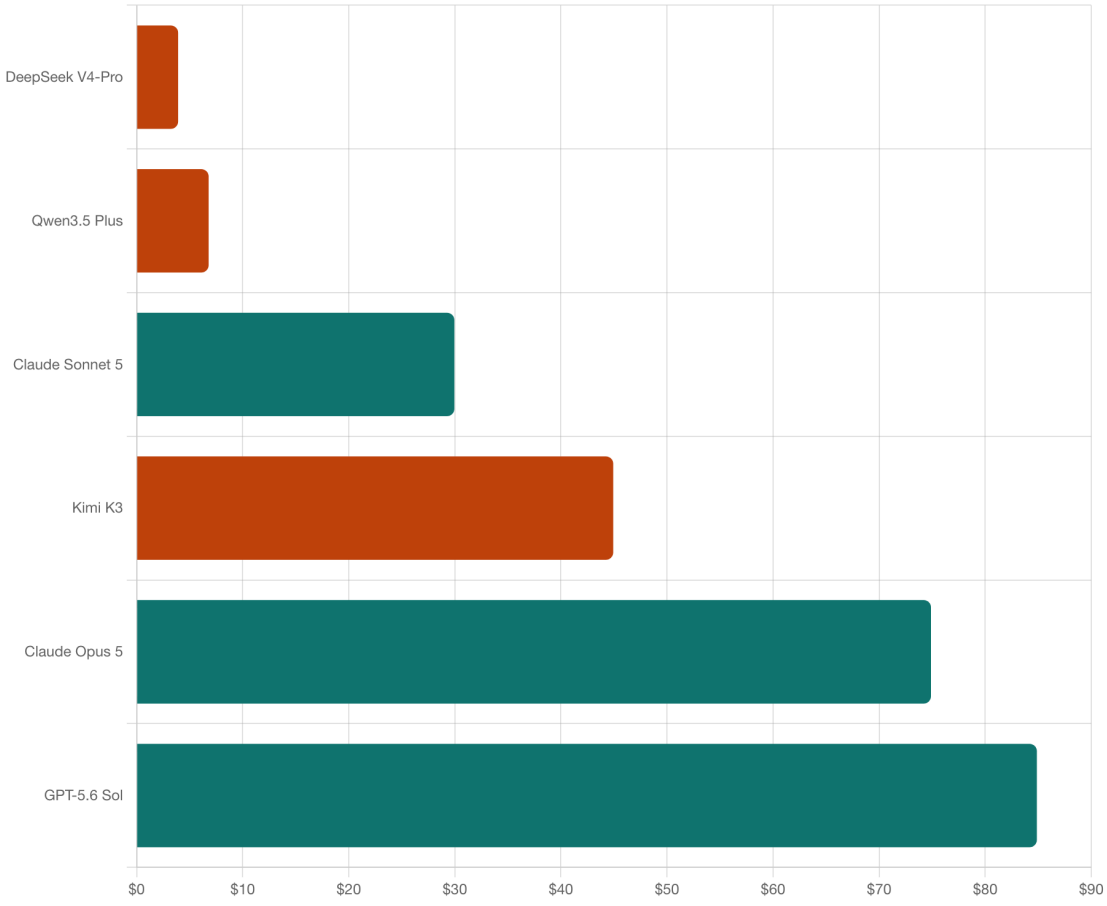
THE COST PICTURE

Is it really cheaper? Yes, and the gap is an order of magnitude.

Not marketing arithmetic: on vendor-published list pricing, the Chinese flagships price 5–60× below US frontier tiers for the same token volumes. The chart below prices a typical light-chat workload (5M input, 2M output tokens per month) across six models an enterprise would realistically shortlist.

EXHIBIT 1

A representative monthly workload: $4 on DeepSeek, $75 on Claude Opus



Illustrative monthly cost, 5M input + 2M output tokens, list pricing, July 2026. Chinese models shown in terracotta, US models in teal. Source: vendor-published pricing pages (DeepSeek, Moonshot AI, Alibaba Cloud, OpenAI, Anthropic), verified directly.

The full price list, thirteen models side by side

Sorted by monthly cost on the same representative workload. The pattern is consistent: every Chinese open-weight model lands below every US frontier tier, and the two efficiency tiers (DeepSeek V4-Flash, Qwen3.5 Flash) are effectively two orders of magnitude cheaper than the US flagships.

EXHIBIT 2

List pricing per million tokens, July 2026

MODEL	VENDOR	ORIGIN	WEIGHTS	$/M IN	$/M OUT	MONTHLY*	VS OPUS 5
DeepSeek V4-Flash	DeepSeek	CHINA	OPEN	0.14	0.28	$1.26	59.5× cheaper
Qwen3.5 Flash	Alibaba	CHINA	OPEN	0.10	0.40	$1.30	57.7× cheaper
DeepSeek V4-Pro	DeepSeek	CHINA	OPEN	0.435	0.87	$3.92	19.1× cheaper
Qwen3.5 Plus	Alibaba	CHINA	OPEN	0.40	2.40	$6.80	11.0× cheaper
Claude Haiku 4.5	Anthropic	US	API only	1.00	5.00	$15.00	5.0× cheaper
GPT-5.6 Luna	OpenAI	US	API only	1.00	6.00	$17.00	4.4× cheaper
Qwen3.7 Max	Alibaba	CHINA	API only	1.65	4.95	$18.15	4.1× cheaper
Claude Sonnet 5	Anthropic	US	API only	2.00	10.00	$30.00	2.5× cheaper
GPT-5.6 Terra	OpenAI	US	API only	2.50	15.00	$42.50	1.8× cheaper
Kimi K3	Moonshot AI	CHINA	OPEN	3.00	15.00	$45.00	1.7× cheaper
Claude Opus 5	Anthropic	US	API only	5.00	25.00	$75.00	baseline
GPT-5.6 Sol	OpenAI	US	API only	5.00	30.00	$85.00	1.1× dearer
Claude Fable 5	Anthropic	US	API only	10.00	50.00	$150.00	2.0× dearer

* Illustrative monthly cost at 5M input + 2M output tokens. Qwen and Kimi pricing shown is the International tier; regional, batch, and cached rates differ. Source: vendor-published pricing pages, July 2026, verified directly rather than via aggregators.

The honest caveat: list price is not total cost

Token price ignores integration effort, evaluation cycles, guardrails, and (for self-hosting) GPU capacity and MLOps headcount. The 10–60× sticker gap typically compresses to a 3–10× all-in advantage once those are counted. That is still large enough to change architecture decisions, which is why the adoption evidence on the next page exists.

The discount is already being banked in production

Three data points from teams that moved real workloads onto Chinese open-weight models. Each is directional rather than audited, but together they show the pattern: when the task tolerates it, the economics win.

100%

Lindy migrated its entire internal agent workload to DeepSeek V4, citing equivalent task success at a fraction of the cost.

4.5% → 46%

OpenRouter share of tokens served by Chinese open-weight models, growth driven by cost-sensitive high-volume workloads.

$4,000 → $200

Reported monthly spend after replatforming a production summarisation pipeline from a US frontier API to self-hosted Qwen.

> **What to take from this:** the cost story is no longer hypothetical. The open question for any given enterprise is not whether the discount exists, it is whether capability holds on your specific tasks, and whether the trust posture fits your data. Those are the next two sections.

A useful discipline when reading vendor-reported savings: ask what the workload was. Summarisation, extraction, classification, and internal agent loops tolerate model substitution well. Customer-facing reasoning, high-stakes drafting, and complex tool use are where frontier US models still justify their premium most often. The benchmark section that follows gives the capability evidence in both directions.

02

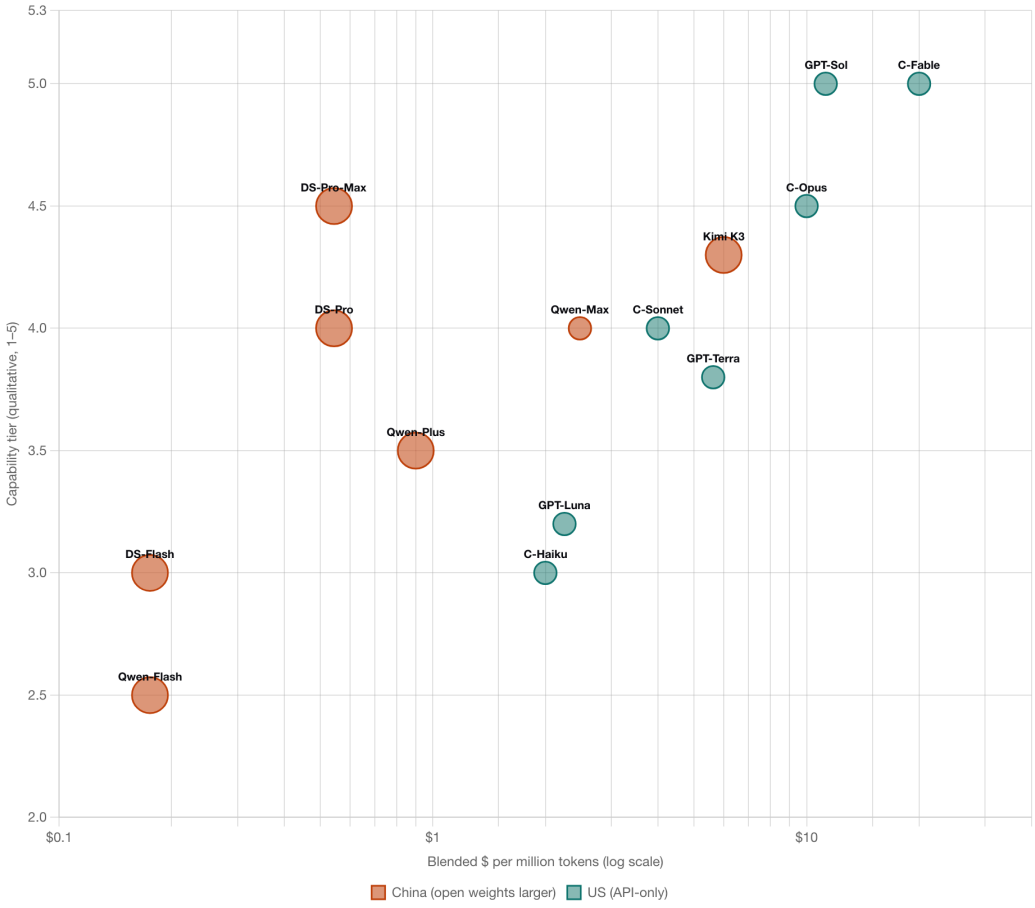
CAPABILITY

Cheap means nothing if the model can't do the work

The capability question is where scepticism is healthiest. The chart below places each model by capability tier (based on lab-reported benchmark clusters) against blended price. The pattern to notice: the Chinese open-weight models cluster in the high-capability, low-cost quadrant that was empty eighteen months ago.

EXHIBIT 3

Chinese open-weight models now occupy the quadrant US vendors priced as impossible



Capability tier vs. blended price per million tokens (log scale). Tiering reflects lab-reported benchmark clusters, July 2026; treat vertical position as directional. Chinese models in terracotta, US models in teal.

What the labs themselves claim, on the record

All figures below are self-reported in each lab's technical report and should be read as a vendor's best case. They are quoted because they are specific and falsifiable, and because the pattern across labs is consistent.

EXHIBIT 4

Lab-reported positioning of the three flagships, July 2026

MODEL	ARCHITECTURE	LAB-REPORTED CLAIMS
Kimi K3 Moonshot AI	2.8T parameters, 104B active · 1M-token context · full weights released (modified MIT)	Positions ahead of GLM-5.2 and Claude Opus 4.8, trailing Claude Fable 5 and GPT-5.6 Sol. GPQA-AA v2 Elo 1682 (vs. 1747 for Fable 5, 1726 for Sol). SWE-Marathon 42.0. Claims 2.5× better scaling efficiency than K2; native vision.
DeepSeek V4 Pro-Max DeepSeek	1.6T parameters, 49B active · 1M-token context · open weights (MIT)	Benchmarks against Claude Opus 4.6-Max, GPT-5.4-aiHigh, and Gemini 3.1-Pro-High. Million-token context at 27% of V3.2's FLOPs and 10% of the KV cache cost, the efficiency claim that underwrites the pricing in Section 01.
Qwen3-235B-A22B Alibaba	235B parameters, 22B active · 119 languages, 36T training tokens · Apache 2.0	AIME'24 85.7 / AIME'25 81.5; LiveCodeBench 70.7; CodeForces Elo 2056; BFCL v3 70.8. Unified thinking and non-thinking modes with a controllable reasoning budget; family spans 0.6B to 235B for edge-to-datacentre deployment.

Source: DeepSeek-AI (arXiv:2606.19348), Qwen Team (arXiv:2505.09388), Kimi Team technical report. All figures self-reported by the model developers.

How to read lab-reported numbers

Treat them as claims of parity, not proof of it. The correct enterprise response is a two-week evaluation on your own task set: same prompts, same harness, blind scoring. The consistent experience of teams that have run these bake-offs is that Chinese flagships win or draw on structured tasks (code, extraction, summarisation) more often than the market's priors expect, and lag most on long-horizon agentic work under ambiguity.

The independent check: what Gartner tells its clients

Vendor claims deserve an outside witness. Gartner's Emerging Tech Impact Radar (February 2025) addresses Chinese foundation models directly, and three findings matter for a buying committee.

> "Chinese LLMs are **closing the gap with speed to release**, putting pressure on global competitors' positioning in price and performance."

> "DeepSeek's models achieve performance **comparable to top-tier global models** in mathematical reasoning, coding, and multilingual tasks, at significantly lower training and inference cost."

> Gartner names the pattern a **"proof of feasibility"**: efficient architectures from Chinese labs have demonstrated that frontier-adjacent capability does not require frontier-scale budgets, and it advises clients to treat them as a legitimate lever in AI cost strategy.

Two qualifications keep the picture honest. First, Gartner's assessment predates the two most recent releases (DeepSeek V4, Kimi K3), so it validates the trajectory rather than today's exact standings. Second, independent benchmark coverage of Chinese models remains thinner than for US flagships; the evaluation burden falls more heavily on the buyer.

The pragmatic conclusion: capability parity is plausible enough, on enough task families, that cost alone no longer settles the shortlist. What settles it is the subject the rest of this paper turns to now: whose jurisdiction your tokens flow through, and whether you can remove that question entirely.

Source: Gartner, Emerging Tech Impact Radar: Generative AI, 14 February 2025, ID G00809486. Quotations abridged.

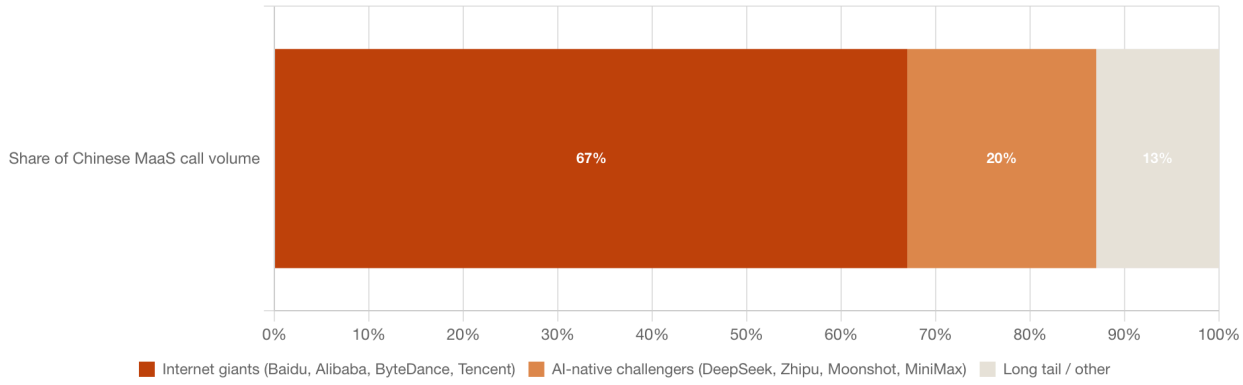
03

THE WIDER LANDSCAPE

Three labs got the headlines. Thirteen models matter.

EXHIBIT 5

Internet giants carry roughly two-thirds of Chinese model-as-a-service call volume



Directional, not audited; compiled from industry landscape reporting, mid-2026. The top 10 providers combined account for roughly 85–90% of volume.

EXHIBIT 6

The market organises into three tiers with different incentives

TIER	WHO	WHAT DEFINES THEM
1 · Big Tech	Alibaba (Qwen), ByteDance (Doubao), Baidu (ERNIE), Tencent (Hunyuan)	Distribution plus balance-sheet scale; models subsidise a bigger consumer and enterprise business
2 · Independents	DeepSeek, Zhipu (GLM), Moonshot (Kimi), MiniMax (the "Four Dragons," combined valuation above $140B), plus Baichuan, StepFun, 01.AI	Model quality *is* the business; this is where the frontier-pushing releases come from
3 · Hardware-adjacent	Xiaomi (MiMo), Huawei (Pangu/Ascend), iFlytek (Spark), Kuaishou (KwaiKAT)	Models exist to sell chips, phones, or a platform; worth watching, not yet frontier-competitive

DeepSeek sits administratively among the independents but behaves more like a research lab than a product company; it is the outlier that makes the "Four Dragons" framing useful shorthand rather than a precise taxonomy.

Release velocity is the real headline

Jan 2025
DeepSeek R1: the reasoning-model wake-up call

May 2025
Qwen3: unified thinking modes, 119 languages

Apr 2026
DeepSeek V4: 1M-token context at 10% of the KV cache cost

Jul 2026
Kimi K3: 2.8T parameters, full weights released

Four major releases from three labs in eighteen months. Whatever this paper says about capability gaps today is a snapshot, not a conclusion. Budget for re-evaluating quarterly.

BEYOND THE BIG THREE

Five more names a buying committee will encounter

MODEL · LAB	POSITION	KNOWN FOR
GLM-5 · Zhipu AI	Startup turned public; HKEX IPO Jan 2026	Leads open-weight agentic and coding benchmarks; approaches Claude Opus 4.5
MiniMax · MiniMax	HKEX listing 2026	Largest context window among Chinese models (256K) at the lowest price point
ERNIE · Baidu	2.4T-parameter omnimodal flagship	Trained on Baidu's own Kunlun silicon, a hedge against GPU export controls
Doubao · ByteDance	Consumer-facing flagship	China's most-used consumer AI app at roughly 155M weekly active users
Hunyuan · Tencent	Enterprise + WeChat ecosystem	Deep distribution advantage via Tencent's messaging and payments footprint

Mistral (France) and other sovereign-AI efforts form a real and growing non-China, non-US category, scoped out of this paper and flagged as a natural follow-up.

A different kind of entry: Manus, the execution layer

Chinese-founded, Singapore-based. Not counted among the thirteen models because it is not a foundation model: it is an agent product that orchestrates other labs' models (Claude, Qwen) to plan and execute multi-step work. It earns a mention because for over six months it has shipped genuinely strong deliverables (researched slide decks, working front ends, formatted reports) straight from a prompt. For buyers evaluating “Chinese AI” broadly rather than one model API, it belongs in the conversation, and it became the sharpest live example of the jurisdictional dynamics covered next.

04

PRIVACY, TRUST & SECURITY

The question that actually approves or kills the deal

Cost and capability get the headlines; this section decides the outcome. The honest framing: hosted Chinese APIs carry real, documented jurisdictional exposure, and open weights are the practical way around most of it.

EXHIBIT 7

Open weights are the structural difference, not the flag on the datacentre

VENDOR	HOSTED DATA LOCATION	LEGAL JURISDICTION	OPEN WEIGHTS	SELF-HOSTABLE OUTSIDE CHINA	LICENCE
DeepSeek	PRC servers (hosted API)	China: 2017 National Intelligence Law	**Yes**	**Yes**	MIT
Moonshot (Kimi K3)	PRC servers (hosted API)	China: 2017 National Intelligence Law	**Yes, full weights**	**Yes**	Modified MIT
Alibaba (Qwen)	PRC or Singapore endpoint	China: 2017 National Intelligence Law	**Yes**	**Yes**	Apache 2.0
OpenAI	US / allied regions	United States: CLOUD Act, FISA 702	**No**	**Hosted only**	Proprietary
Anthropic	US / allied regions	United States: CLOUD Act, FISA 702	**No**	**Hosted only**	Proprietary

Both jurisdictions can compel provider cooperation with government data requests; enterprises weigh both. The practical difference is optionality.

> The distinction that matters is not “China bad, US good.” It is **optionality**: DeepSeek, Kimi, and Qwen ship as open weights under permissive licences, so you can download the model and run it entirely inside your own AWS, Azure, GCP, or on-prem environment. No data ever touches a China-hosted endpoint. OpenAI and Anthropic's frontier models offer no equivalent path: it is their hosted API or nothing.

Where governments have already drawn a line

Government and regulated-sector restrictions on hosted DeepSeek specifically, as of mid-2026. Every restriction on record targets the hosted app or API, not the underlying open-weight model run independently, a distinction most coverage glosses over.

Italy
Blocked from app stores (Jan 2025) over unresolved GDPR data-practice questions

Australia
Banned on all government devices and systems

Taiwan
Prohibited across public sector, state-owned enterprises, schools, critical infrastructure

South Korea
Banned at Ministry of Trade, Industry & Energy and Korea Hydro & Nuclear Power

United States
Blocked by NASA, US Navy, Pentagon, and Dept. of Commerce on government-furnished devices

Canada & India
Formal restrictions on deployment and commercial use under review or in force

The exposure runs in both directions

The list above is Western governments restricting Chinese-hosted AI. The reverse case matters just as much for anyone assessing jurisdictional risk, and the clearest recent example is not a model at all.

Manus and the blocked Meta acquisition

In late 2025, Meta agreed to acquire Manus, the Chinese-founded, Singapore-based agent platform, for over $2B. On 27 April 2026, China's state planner (the body that reviews outbound investment and technology transfer) intervened and ordered both sides to unwind the deal. Manus continued shipping through the process, including a March 2026 desktop app.

Read the most direct way: Beijing does not block a $2B outbound sale of a product it considers replaceable. The intervention signals how much strategic value China places on keeping frontier-adjacent AI onshore, the same instinct behind the open-weight releases from DeepSeek, Kimi, and Qwen. The jurisdictional story is not one-directional restriction; it is two governments, each moving to keep the AI they see as strategic inside their own borders.

05

DECISION FRAMEWORK

So which one, for you?

Not a recommendation, but a way to reason about your own situation. Four questions set the starting position: data sensitivity, regulatory exposure, self-hosting capacity, and cost pressure.

EXHIBIT 8

Four questions map to five starting positions

YOUR SITUATION	SUGGESTED STARTING POINT
High data sensitivity or regulatory exposure, no self-hosting capacity	Stay on US-hosted frontier models for now. The hosted Chinese APIs put your prompts under PRC jurisdiction, and without an infrastructure team the open-weight escape hatch is not yet practical for you.
High sensitivity or exposure, and you can self-host	Self-host an open-weight Chinese model (DeepSeek V4 or Qwen3.5) inside your own cloud tenancy. You capture most of the cost advantage while keeping every token inside your own jurisdiction and control plane.
Low sensitivity, cost is a real constraint	Use the hosted DeepSeek or Qwen API for the workloads that tolerate it; route the remainder to your incumbent. Kimi K3 is the candidate where coding and agentic quality matter most.
Low sensitivity, cost is flexible	Run a two-week head-to-head: your incumbent frontier model against Kimi K3 and DeepSeek V4 on your actual task set, blind-scored. Let the evidence, not the priors, set the shortlist.
Everything in between	Pilot on non-sensitive workloads first. Prove capability and integration cost on your own tasks before the trust conversation reaches procurement and legal.

An interactive version of this framework, which scores your answers and returns a tailored starting point, is available in the online edition at agenticassure.ai/reports.

> **The one-line version:** the cost gap is real, the capability gap is task-dependent and shrinking, and the trust decision is yours to engineer. Open weights turn a jurisdiction problem into an infrastructure project, and infrastructure projects are things enterprises know how to run.

06

SOURCES & METHODOLOGY

Appendix

Pricing and benchmark figures compiled July 2026. Pricing changes frequently; treat every cost figure as directional and re-verify against the vendor pages below before budgeting a production deployment. All six pricing sources were fetched directly from the vendor's own documentation in the course of preparing this paper, not from third-party aggregators.

PRIMARY TECHNICAL REPORTS

DeepSeek-AI, *DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence*, arXiv:2606.19348

Qwen Team, *Qwen3 Technical Report*, arXiv:2505.09388

Kimi Team, *Kimi K3: Open Frontier Intelligence*, Technical Report

Gartner, *Emerging Tech Impact Radar: Generative AI*, 14 Feb 2025, ID G00809486

PRICING (VENDOR-PUBLISHED, VERIFIED DIRECTLY)

DeepSeek · Official API pricing docs · api-docs.deepseek.com

Moonshot AI · Kimi K3 official pricing · platform.kimi.ai

Alibaba Cloud · Model Studio pricing (Qwen) · alibabacloud.com

Anthropic · Claude platform pricing · platform.claude.com

OpenAI · GPT-5.6 announcement and pricing · openai.com

OpenAI Help Center · GPT-5.6 Sol / Terra / Luna preview · help.openai.com

PRIVACY, JURISDICTION & RESTRICTIONS

Security Magazine · Dangers of DeepSeek's privacy policy

Theori · DeepSeek security, privacy & governance

Demandsage · DeepSeek banned countries 2026

BankInfoSecurity · Asian governments ban DeepSeek

MindStudio · Self-hosting open-weight Chinese models

Zenn · Is self-hosting Chinese open-weight LLMs economically rational?

ADOPTION EVIDENCE & WIDER LANDSCAPE

BayTech Consulting · Enterprise decision framework for DeepSeek V4 (Lindy case study)

eCorp IT · Chinese open models cutting enterprise AI bills 60–90% (2026)

Digital Applied · Chinese AI models Q2 2026: 10-provider landscape report

TechRadar · Meta's Manus acquisition and the blocked deal

Manus.im · Product and capabilities

Presenc.ai · Chinese open-source LLM companies leaderboard 2026

Methodology note

Benchmark figures are self-reported by model developers unless attributed to Gartner. The illustrative monthly workload (5M input, 2M output tokens) approximates a light internal-chat deployment; your blend of input to output tokens materially changes the rankings for output-heavy workloads. Market-share figures are directional estimates compiled from industry landscape reporting, not audited disclosures. Not legal advice; consult counsel on jurisdictional and regulatory questions specific to your data.

ABOUT AGENTICASSURE

Govern every AI. Prove it to every auditor.

AgenticAssure is an AI governance and assurance platform: continuous evidence collection, control mapping, and audit-ready reporting for enterprises deploying LLMs and agentic systems in regulated environments.

This paper is part of the AgenticAssure Research Series: independent, evidence-based briefings written for the executives who have to sign the deployment decision, not just admire the technology.

If your organisation is weighing Chinese open-weight models, or needs the governance evidence to deploy any model with confidence, we would be glad to talk.

BOOK A BRIEFING
tidycal.com/agenticassure

ENQUIRIES
contact@agenticassure.ai

WEB
agenticassure.ai