

# Kina er 3–6 måneder bagud. Og 60–90 % billigere.

*De kinesiske frontier-modeller: hvem de er, hvor gode de faktisk er, hvad de koster — og hvad du skal svare, når spørgsmålet rammer dit ledelsesmøde.*

**AI OG STRATEGI · FRONTIER-MODELLER · AUGUST 2026**

Stefano Vincenti er AI-træner, konsulent og ekstern lektor ved ITU og DIS Copenhagen og co-founder af BotTellMe. Han hjælper nordiske ledelsesteams og fagfolk med at gå fra AI-nysgerrige til konkret, ansvarlig brug.

**Følg på LinkedIn**

**Tilmeld nyhedsbrevet**

[linkedin.com/in/stefanovincenti](https://linkedin.com/in/stefanovincenti) · [stefanovincenti.substack.com](https://stefanovincenti.substack.com) · [aitrainer.dk](https://aitrainer.dk)

Jeg underviser i de amerikanske frontier-modeller. På ITU og på DIS Copenhagen er det OpenAI, Anthropic, Copilot og Google, som primært står på pensum. Det er dem, de studerende møder, når de kommer ud. Det er dem, mine klienter køber. Og det er dem, jeg selv arbejder i hver dag.

Så kommer spørgsmålet, som jeg efterhånden får i næsten hver eneste klientsamtale: "Må vi bruge DeepSeek?" Det stilles sjældent af nysgerrighed. Det stilles, fordi nogen i organisationen allerede har regnet på prisen.

Jeg har brugt sommeren på at finde ud af, hvad det rigtige svar er. Her er det: hvem de kinesiske spillere er, hvor gode deres modeller faktisk er, hvad de koster, og hvad du konkret skal svare. Undervejs flyttede mit eget svar sig et sted hen, jeg ikke havde regnet med. Det kommer til sidst.

## ■ 1. UGE 29 VAR INGEN TILFÆLDIGHED

Den 16. juli udgav Moonshot AI deres Kimi K3, den største åbne model nogensinde med 2,8 billioner parametre. Den 19. juli svarede Alibaba med en preview af Qwen3.8-Max på World AI Conference i Shanghai. Den 24. juli lukkede DeepSeek de gamle API-navne, så alle med DeepSeek i produktion skulle migrere til V4.

Og mens modellerne landede, rykkede politikken. Den 21. juli rapporterede Financial Times og Reuters, at Beijing overvejer eksportkontrol på kinesiske modelvægte. Den 24. juli underskrev 77 virksomheder et åbent brev til forsvar for åbne vægte — Nvidia, Google, Meta, Microsoft, Mistral og OpenAI blandt dem. Anthropic skrev ikke under. Den 29. juli lukkede Moonshot en runde på 3,5 milliarder dollars til en værdi af 35 milliarder. Den 3. august lancerede Alibaba Qwen3.8-Max for fuld udblæsning, hvorefter aktien steg 9 procent.

Historien om kinesisk AI som billig efterligning døde et sted i 2025. Det her er et kapløb, hvor nummer to løber hurtigere og hurtigere — og det er næsten gratis.

## ■ 2. DEEPSEEK: HEDGEFONDEN DER KOSTEDE NVIDIA 590 MIA. DOLLARS

DeepSeek ligner ingen anden AI-virksomhed i verden. Selskabet blev grundlagt i 2023 i Hangzhou af Liang Wenfeng, manden bag kvant-hedgefonden High-Flyer, og finansieret af hans egne fondsoverskud. Omkring 150–170 ansatte, ingen KPI'er, unge forskere fra kinesiske eliteuniversiteter, og en kultur, der minder mere om et forskningsinstitut end en virksomhed.

Verden opdagede dem den 20. januar 2025, da R1-modellen udkom og DeepSeek-appen gik til tops i den amerikanske App Store. En uge senere faldt Nvidia 17 procent på én dag, cirka 590 milliarder dollars i markedsværdi — dengang det største endagstab for et enkelt selskab i amerikansk børshistorie. Pointen, der skræmte markedet: selve træningskørslen bag V3 kostede cirka 5,6 millioner dollars i GPU-tid. Det tal dækker kun den afsluttende kørsel, ikke al forskning og alle fejlslagne forsøg — men selv med forbeholdet står det i skærende kontrast til amerikanske frontier-budgetter på hundredvis af millioner.

I juni 2026 tog DeepSeek imod ekstern kapital for første gang: 7,4 milliarder dollars til en værdiansættelse over 50 milliarder. Læg mærke til konstruktionen: de fleste investorer købte sig ind i et kommanditselskab kontrolleret af Liang selv, uden stemmeret og med fem års binding. Den eneste investor med stemmeret og uden lock-up er Kinas statslige AI-industrifond.

DeepSeek V4 findes i to udgaver: V4 Pro med 1,6 billioner parametre (49 milliarder aktive) og V4 Flash med 284 milliarder (13 milliarder aktive), begge med 1 million tokens kontekst og begge under en ægte MIT-licens. Prisen er det vilde: Flash koster 0,28 dollars pr. million output-tokens. OpenAI's GPT-5.6 Sol tager 30 dollars for det samme.

Svaghederne er lige så konkrete: first-party-hosting i Kina under kinesisk lov, dokumenteret censur på politisk følsomme emner, svagere multimodalitet, og en R2-model der aldrig udkom, angiveligt fordi træning på Huaweis chips fejlede. Og bemærk: DeepSeeks egen API træner på dine data. Det gør de vestlige hosts, der serverer samme model, ikke.

### ■ 3. MOONSHOT: MODELLEN DER BLEV SÅ POPULÆR, AT DE LUKKEDE FOR TILMELDING

Moonshot AI blev grundlagt i Beijing i 2023 af Yang Zhilin, en forsker med en ph.d. fra Carnegie Mellon. Alibaba og Tencent er blandt investorerne. Værdiansættelsen gik fra 2,5 milliarder dollars i februar 2024 til 20 milliarder i maj 2026, og den 29. juli lukkede de 3,5 milliarder dollars til en værdi af 35 milliarder.

Kimi K3, udgivet 16. juli, er deres svendestykke: 2,8 billioner parametre, 104 milliarder aktive, 1 million tokens kontekst, nativ multimodalitet. Samme dag den udkom, gik den ind som nummer 1 i Arena.ai's Frontend Code Arena med 1.679 point, foran Claude Fable 5 (1.631) og GPT-5.6 Sol (1.618). Et spring fra plads 18 til plads 1. Tre dage senere måtte Moonshot lukke for nye abonnenter med deres egen formulering: "Kimi K3 has received far more love than we expected, and our GPUs are feeling it."

Den 27. juli kom vægtene, og K3 blev den hurtigst voksende udgivelse i Hugging Faces historie med næsten en halv million downloads den første måned. Fra det øjeblik kan enhver hoste modellen selv. Men læs licensen, før du planlægger noget: K3 ligger under en selvskrevet "Kimi K3 License", ikke MIT eller Apache. Åbne vægte er ikke det samme som frie vilkår.

To ærlige forbehold. Moonshot skriver selv, at K3 stadig ligger bag de bedste amerikanske lukkede modeller på samlet kapabilitet. Og forspringet på kodning holdt ikke: K3 ligger i dag nummer 2 i Frontend Code Arena med 1.676 point, overhalet af Claude Opus 5 Max med 1.705. Tre uger på førstepladsen.

### ■ 4. ALIBABA: QWEN ER FUNDAMENTET UNDER DET MESTE OPEN SOURCE-AI

Mens DeepSeek og Moonshot tager overskrifterne, har Alibaba bygget noget mere strukturelt. Qwen-familien rundede en milliard downloads på Hugging Face i januar 2026, med omkring 200.000 afledte modeller. Det er verdens mest brugte åbne modelfamilie, og den overhalede Llama undervejs. Da Apple Intelligence den 15. juli fik myndighedsgodkendelse til Kina, var Qwen den navngivne partner i stakken.

Qwen3.8-Max blev previewet 19. juli og lanceret for fuld udblæsning den 3. august: 2,4 billioner parametre, multimodal, 1 million tokens kontekst. Alibabas eget claim er, at den kun overgås af Claude Fable 5. En uafhængig måling placerede den i Arena.ai's Frontend Code Arena den 2. august på plads 4 med 1.668 point, kun 8 point efter Kimi K3. Men der er stadig ingen offentliggjort licens og intet fuldt model card, så self-hosting kan du ikke planlægge på endnu.

Alibabas reelle styrke er porteføljen: modeller i alle størrelser, en cloud-forretning der tjener penge på inferensen, Apple-aftalen som forbrugerdistribution, og et investeringsprogram på 380 milliarder yuan

(cirka 53 milliarder dollars) over tre år. Svagheden er talent og tillid: Qwens tekniske chef Junyang Lin forlod abruptt selskabet den 3. marts, dagen efter en større modeludgivelse.

## ■ 5. BAG DE TRE STORE STÅR EN HEL LIGA — OG DEN ER BØRSNOTERET NU

Zhipu AI blev den 8. januar verdens første børsnoterede LLM-selskab. Markedsværdien toppede over 1.000 milliarder Hongkong-dollars (cirka 128 milliarder dollars) i slutningen af juni efter GLM-5.2. GLM-5.2 var i juni den bedste åbne model på Artificial Analysis' intelligensindeks og fører på PostTrainBench for agentisk kodning. Både GLM-5 og GLM-5.2 er efter selskabets eget udsagn trænet udelukkende på Huawei's Ascend-chips, uden en enkelt Nvidia-GPU. Det er en påstand fra udgiveren, ikke en uafhængig verifikation — men hvis den holder, betyder den, at amerikansk chipeksportkontrol har en bagdør.

MiniMax fulgte på børsen dagen efter Zhipu og fordoblede kursen på debutdagen. ByteDance kører Seed-modellerne med aggressive priser, Tencents Hunyuan er gået open source under Apache-licens, Baidus Ernie er trænet på deres egne Kunlun-chips, og StepFun kører effektivitetssporet med små, hurtige åbne modeller. For to år siden talte man om "de seks AI-tigre". I dag er billedet konsolideret, og de fire tech-giganter er gået fra sponsorer til konkurrenter. Våbnet, de deler, er det samme: åbne vægte.

## ■ 6. FORSPRINGET ER 39 POINT PÅ ELO — OG DET ER BLEVET STØRRE

Her skal man passe på med tallene, fordi to helt forskellige målinger konstant bliver blandet sammen. Stanfords AI Index 2026 opgør forskellen mellem den bedste amerikanske og den bedste kinesiske model på LMArenas Elo-skala. I marts 2026 førte Claude Opus 4.6 med 1.503 point over den bedste kinesiske model, Dola-Seed-2.0 Preview, med 1.464. En forskel på 39 point, altså 2,7 procent.

### Hvor stort er gabet — målt rigtigt?

*Elo-skala (LMArena) og samlet intelligensindeks, ikke enkelt-benchmarks*

# 39 point

US-forspring på Elo-skalaen (2,7 %), marts 2026 — større end året før

*Stanford AI Index 2026*

# 3–6 mdr

konstant gab bag den amerikanske frontier, hverken voksende eller lukkende

*OpenRouter, 18 mdr. data*

# nr. 4

Kimi K3 på det samlede intelligensindeks — bag Opus 5, Fable 5 og GPT-5.6 Sol

*Artificial Analysis*

### USA beholder forspringet — men det er lejet, ikke ejet

Fortællingen om, at Kina overhaler indenom, holder ikke på Elo-målingen. I feb. 2025 var gabet 5 point; et år senere 39.

*Model 1. De store "17–31 procentpoint"-gab stammer fra 2023-benchmarks (MMLU m.fl.) — anden skala, ikke sammenlignelig.*

Et sted skal du være særligt skeptisk: kodning. Der findes endnu ingen uafhængig SWE-bench Verified-score for Kimi K3. Alle de procenttal, du ser cirkulere, er Moonshots egne kørsler eller rene gæt. Moonshots eget tal er 67,5 på DeepSWE mod Claude Fable 5's 70,0 og GPT-5.6 Sol's 73,0. Tæt på, men bagefter, og målt af udgiveren selv.

## ■ 7. REGNESTYKKET DINE UDVIKLERE ALLEREDE HAR LAVET

En typisk automatiseringsworkload på 50 millioner output-tokens om måneden koster meget forskelligt afhængigt af model:

| Model                 | Pris pr. 1 mio. output-tokens | 50 mio. tokens/md |
|-----------------------|-------------------------------|-------------------|
| OpenAI GPT-5.6 Sol    | \$30                          | ~\$1.500          |
| Claude Fable 5        | \$50                          | ~\$2.500          |
| Claude Opus 5         | \$25                          | ~\$1.250          |
| Google Gemini 3.1 Pro | \$12                          | ~\$600            |
| DeepSeek V4 Flash     | \$0,28                        | ~\$14             |

Tabel 1. Vær fair: der findes billige amerikanske muligheder (Gemini). Men afstanden fra \$600 til \$14 er stadig en faktor 40.

Markedet har reageret. På OpenRouter, den største uafhængige model-router, overhalede kinesiske modeller de amerikanske i tokenforbrug i begyndelsen af juni 2026. I dag er over 60 procent af den routede trafik kinesisk, alle fem mest brugte modeller er kinesiske, og Metas Llama ligger under 1 procent. Amerikanske modellers andel er ifølge en Bloomberg-opgørelse faldet fra omkring 70 procent i juni 2025 til omkring 30 procent et år senere.

### TO FORBEHOLD, FØR DU REGNER VIDERE

OpenRouter måler udviklertrafik, ikke omsætning — på omsætning fører de amerikanske labs stadig klart. Og de billigste priser er first-party fra Kina, hvor dine data ryger til kinesiske servere. Vestlige hosts (Fireworks, Together, Azure) tager typisk det dobbelte af den kinesiske first-party-pris for samme model, uden at træne på dine data. Det dobbelte af næsten ingenting er stadig næsten ingenting.

## ■ 8. BEGGE SUPERMAGTER OVERVEJER AT LUKKE GRÆNSEN

Washington er i krig med sig selv om svaret. Den 24. juli underskrev 77 virksomheder brevet "Open Weights and American AI Leadership" — Nvidia, Google, Meta, Microsoft, IBM, Hugging Face, Mistral, Mozilla, Linux Foundation og OpenAI. Anthropic skrev ikke under og stod dermed alene. Den 27. juli svarede de med et indlæg, der åbner med, at de aldrig har argumenteret for et forbud mod åbne vægte, og i stedet foreslår chipeksportkontrol, en indsats mod destillation og obligatorisk sikkerhedstest af alle tilstrækkeligt kapable modeller.

Og så det spejlvendte, som europæiske virksomheder bør læse to gange. Den 12. juni pålagde det amerikanske handelsministerium Anthropic at kræve eksportlicens for enhver udenlandsk statsborgers adgang til Claude Fable 5 og Mythos 5, inde i som uden for USA. Anthropic slukkede begge modeller for alle kunder for at overholde ordren. Kontrollen blev ophævet den 30. juni, og Fable 5 var globalt tilgængelig igen den 1. juli. Atten dage uden adgang.

#### LEKTIE

En amerikansk lukket model kan forsvinde fra den ene dag til den anden, uden varsel, og af grunde der ikke har noget med dig at gøre. Ironisk nok gjorde episoden de kinesiske åbne modeller mere attraktive — for en downloadet vægt-fil kan ingen regering slukke. Beijing overvejer nu selv eksportkontrol på deres egne modelvægte. Vinduet, hvor frontier-vægte flyder frit over grænserne, står måske ikke åbent for evigt.

### ■ 9. HVAD DU SKAL SVARE, NÅR SPØRGSMÅLET RAMMER LEDELSESMØDET

Efter en sommer med det spørgsmål er mit svar kogt ned til tre regler.

**1. Kinesiske first-party-API'er med persondata: nej.** At sende EU-persondata direkte til deepseek.com eller kimi.com har ingen dokumenteret GDPR-overførselsmekanisme. Italiens Garante blokerede DeepSeek den 30. januar 2025, syv tyske delstatstilsyn åbnede sag i februar 2025, og Folketingets Præsidium forbød DeepSeek på alle Folketingets enheder i vinteren 2025. Det spor er lukket for enhver seriøs europæisk virksomhed.

**2. Samme modeller via vestlige hosts eller self-hosting: en reel mulighed — men læs licensen.** Kører du vægtene hos en europæisk eller amerikansk udbyder med no-train-garanti, forlader dine data aldrig din kontrol. Men "åben" dækker over meget forskellige juridiske virkeligheder: DeepSeek V4 er ægte MIT, Zhipu GLM-5.2 er MIT, Kimi K3 har en selvskrevet licens, og Qwen3.8-Max har slet ingen offentliggjort licens endnu.

**3. Governance før pilotprojekt — og nu er der en deadline bag.** EU AI Act's håndhævelse over for general purpose-modeller trådte i kraft den 2. august 2026. Kommissionen kan nu kræve dokumentation, evaluere modeller, kræve tilbagetrækning og udstede bøder på op til 3 procent af global omsætning eller 15 mio. euro. Finjusterer eller videredistribuerer I en åben model, kan I selv blive "provider" med de forpligtelser, det medfører.

Tre spørgsmål, jeg stiller mine klienter lige nu: Ved I, hvilke modeller jeres leverandører og værktøjer faktisk kører jer på i dag? Har I regnet på, hvad jeres største AI-workload ville koste på en åben model via en vestlig host? Og har jeres AI-politik et svar på kinesiske modeller — eller opdager I det først, den dag en udvikler allerede har dem i produktion?

### ■ 10. OG SÅ DET, JEG SELV ER LANDET PÅ HEN OVER SOMMEREN

Jeg underviser stadig i de amerikanske modeller på ITU og DIS, fordi det er dem, markedet køber, og fordi det er dem, de studerende møder, når de kommer ud. Men noget flyttede sig. Når jeg sidder med en

europæisk klient og regner på data, jura og leverandørrisiko, er det ikke længere et amerikansk navn, jeg skriver først på tavlen.

De atten dage, hvor Fable 5 var slukket, viste, hvad leverandørafhængighed koster i praksis. De kinesiske first-party-API'er er lukket land, når der er persondata i spil. Og fra august har AI Act sat en pris på compliance, som ledelsen kan se i et regneark. Tilbage står to muligheder, der peger samme vej: åbne vægte hos en europæisk host, eller de europæiske modeller selv. For en EU-virksomhed er det efterhånden tæt på en no-brainer. Det havde jeg ikke sagt for et halvt år siden — og det er ikke, fordi de europæiske modeller er blevet bedst. Det er, fordi resten af regnestykket har ændret sig. Det er den artikel, jeg skriver næste gang.

## ■ KILDER

Denne analyse bygger på et bredt kildegrundlag; de vigtigste er (status pr. august 2026):

- DeepSeek: [api-docs.deepseek.com](https://api-docs.deepseek.com) · [huggingface.co/deepseek-ai](https://huggingface.co/deepseek-ai) (specs, MIT-licens) · [forbes.com](https://forbes.com) 17/6 2026 · [arxiv.org/pdf/2412.19437](https://arxiv.org/pdf/2412.19437) (V3-træning) · [forbes.com](https://forbes.com) & [reuters.com](https://reuters.com) 27/1 2025 (Nvidia-tab)
- Moonshot/Kimi: [huggingface.co/moonshotai/Kimi-K3](https://huggingface.co/moonshotai/Kimi-K3) · [x.com/arena](https://x.com/arena) 16/7 2026 · [arena.ai/leaderboard/code/webdev](https://arena.ai/leaderboard/code/webdev) · [bloomberg.com](https://bloomberg.com) 7/5 & 29/7 2026 · [technode.com](https://technode.com) 3/8 2026
- Alibaba/Qwen: [x.com/alibaba\\_cloud](https://x.com/alibaba_cloud) 21/1 2026 · [marktechpost.com](https://marktechpost.com) 19/7 & 3/8 2026 · [techcrunch.com](https://techcrunch.com) 15/7 2026 (Apple i Kina) · [technode.com](https://technode.com) 4/3 2026 (Junyang Lin)
- Øvrige labs: [yicaiglobal.com](https://yicaiglobal.com) & [scmp.com](https://scmp.com) (Zhipu) · [deeplearning.ai/the-batch](https://deeplearning.ai/the-batch) 26/6 2026 · [huggingface.co/blog/mlabonne/glm-5](https://huggingface.co/blog/mlabonne/glm-5) (Ascend) · [cnbc.com](https://cnbc.com) 9/1 2026 (MiniMax)
- Kapløb og priser: [hai.stanford.edu](https://hai.stanford.edu) (AI Index 2026 & 2025) · [artificialanalysis.ai](https://artificialanalysis.ai) · [openrouter.ai/blog/insights](https://openrouter.ai/blog/insights) 30/6 2026 · [dataconomy.com](https://dataconomy.com) 29/7 2026 · [platform.claude.com/docs](https://platform.claude.com/docs) · [developers.openai.com](https://developers.openai.com) · [ai.google.dev](https://ai.google.dev)
- Politik: [axios.com](https://axios.com) 17/7–22/7 2026 · [ppc.land](https://ppc.land) (77-underskriverbrevet) · [anthropic.com/news/position-open-weights-models](https://anthropic.com/news/position-open-weights-models) 27/7 · [fable-mythos-access](https://fable-mythos-access) 12/6 & [redeploying-fable-5](https://redeploying-fable-5) 30/6 · [cnbc.com](https://cnbc.com) 8/7 & [qz.com](https://qz.com) 31/7 (Kongressen) · [reuters.com](https://reuters.com) & [ft.com](https://ft.com) 21/7
- Kunde-eksempler: [scmp.com](https://scmp.com) 22/10 2025 (Airbnb) · [bloomberg.com](https://bloomberg.com) 20/5 2026 · [cursor.com/blog/composer-2-5](https://cursor.com/blog/composer-2-5)
- EU/compliance: [artificialintelligenceact.eu](https://artificialintelligenceact.eu) (kap. V, 2/8 2026) · [garanteprivacy.it](https://garanteprivacy.it) 30/1 2025 · [datenschutz.hessen.de](https://datenschutz.hessen.de) 20/2 2025 · [ft.dk](https://ft.dk) (Folketingets DeepSeek-forbud)

© 2026 Stefano Vincenti · [aitrainer.dk](https://aitrainer.dk) · AI gengivelse med kildeangivelse.

## Stefano Vincenti

AI-træner, konsulent og ekstern lektor · 20+ år i software og digital transformation · Co-founder af BotTellMe · Ekstern lektor ved ITU og DIS Copenhagen · Partner hos TryZone.

### LinkedIn

[linkedin.com/in/stefanovincenti](https://linkedin.com/in/stefanovincenti)

### Nyhedsbrev

[stefanovincenti.substack.com](https://stefanovincenti.substack.com)

### Mere

[aitrainer.dk](https://aitrainer.dk) · [tryzone.dk](https://tryzone.dk) ·  
[bottellme.com](https://bottellme.com)

*Dette er generel praktisk vejledning, ikke juridisk rådgivning eller compliance-rådgivning, og jeg er ikke advokat eller compliance-specialist. Brug det ikke som beslutningsgrundlag på egen hånd. Tag altid jeres egen compliance- eller juridiske afdeling og jeres DPO med på råd, og bekræft datoer og status mod primære kilder (EUR-Lex, Europa-Kommissionen, Datatilsynet). Alle priser er oplyst i USD ekskl. moms og kan være ændret. Status pr. august 2026.*

**Stefano Vincenti · AI, Built Human**