

The Ox in the room

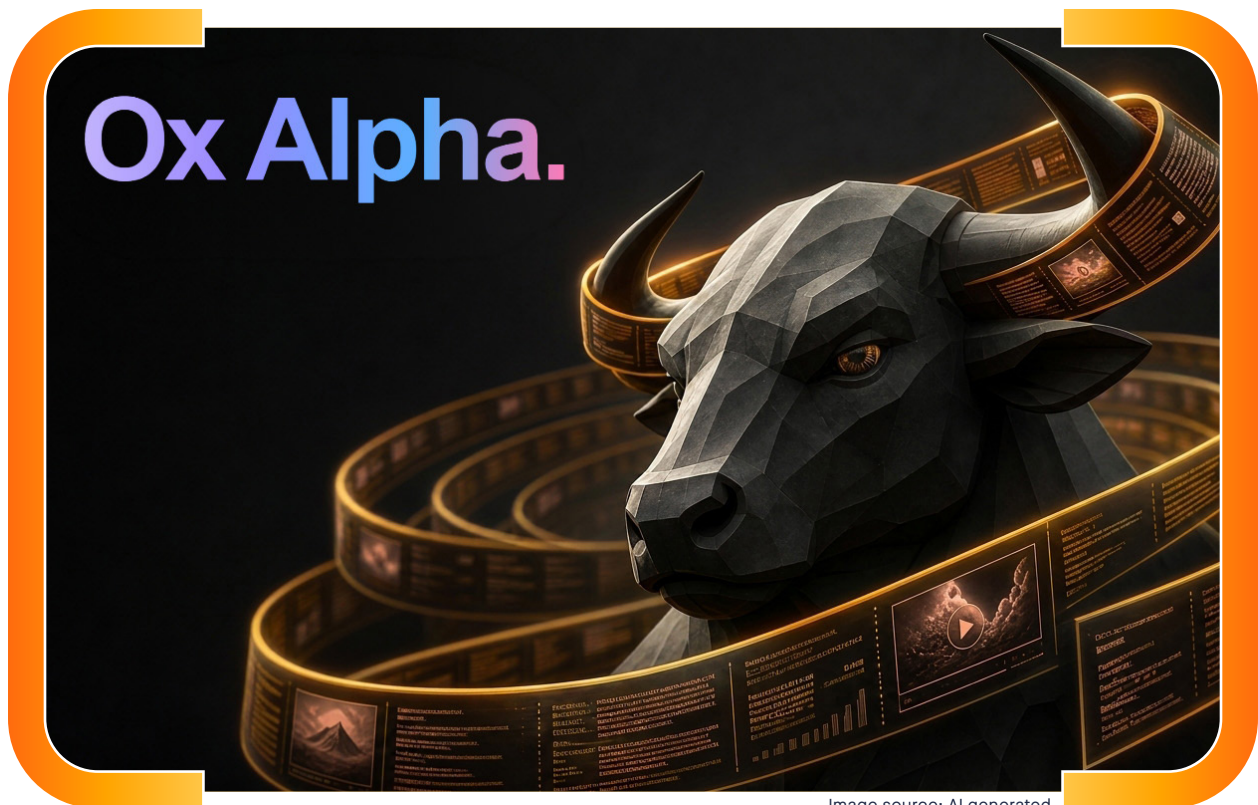


Image source: AI generated

Charles-Henry Monchau, CFA, CAIA, CMT
Chief Investment Officer
charles-henry.monchau@syzgroup.com

Ox Alpha, an anonymous model nobody has claimed, has just taken the 1st place on the world's busiest LLM marketplace. The story is not about China. It is about pricing power and about who ultimately earns a return on a trillion dollars of AI infrastructure.



Source: Bloomberg, Syz Group

A leaderboard worth reading

On 20 August, a model called Ox Alpha appeared on OpenRouter with no name attached to it. OpenRouter is the plumbing of the AI economy: a single interface through which developers can reach several hundred models and switch between them at will. Its weekly leaderboard does not rank models by benchmark scores. It ranks them based on the number of tokens developers actually consumed, capturing revealed preference rather than sell-side ratings, and actual order flow rather than research notes.

Within days, Ox Alpha ranked number one, at 17.5 trillion tokens for the week, way ahead of DeepSeek. It offers a one-million-token context window, accepts text, images and video, is free during its preview, and its provider states it does not train on user prompts. The open-source agent framework OpenCode reported that the provider had capacity for 100 trillion tokens a day. Patrick Collison, co-founder and CEO of Stripe, the global payments technology company, called it “very impressive.” Some of the more spectacular early figures rested on tiny samples, and broader evaluations are more nuanced. But the essential fact stands: an unidentified player can now release a model good enough that the community argues in earnest over whether it comes from Google, Microsoft or a Chinese laboratory such as Z.ai. Two years ago that was unthinkable.

Gemini 3.7 Flash is 75% off for a limited time. See all discounted models [here](#)

LLM Leaderboard

Compare the most popular models on OpenRouter

Rank	Model	Provider	Tokens	% Change
1	Ox Alpha	by stealth	17.5T	new
2	DeepSeek V4 Flash 07...	by deepseek	11.6T	↑3%
3	MiMo-V2.5	by xiaomi	9.69T	↑94%
4	Hy3	by tencent	7.76T	↓20%
5	DeepSeek V4 Flash 04...	by deepseek	5.53T	↑16%
6	Nemotron 3 Ultra (free)	by nvidia	5.12T	↑114%

OpenRouter weekly token consumption. Ox Alpha, provider unknown, enters directly at number two.

Look past the mystery to the rest of the table. Six of the top ten models are Chinese: DeepSeek, Xiaomi, Tencent, Z.ai. The highest-ranked model from a US frontier laboratory sits in sixth place. On the metric that measures what developers actually run in production, the American lead has already gone.

The frontier, in plain words

The “technological frontier” is simply the best performance anyone can obtain from a machine at a given moment. What matters for investors is not where that frontier sits, but how many players stand on it and how quickly the tenth-best catches the first.

Three forces are compressing that distance. Research diffuses through published papers, open-weight models, the distillation of larger models, and the movement of a relatively small number of researchers. Efficiency also compounds: mixture-of-experts architectures, quantisation and speculative decoding have reduced the compute required for each unit of capability by an order of magnitude. And each new capability, once demonstrated to be possible, becomes far cheaper to reproduce than it was to discover. US semiconductor restrictions remain a real constraint on China, but they are being partly offset by algorithmic efficiency, inference optimisation and disciplined use of the compute available. The frontier is not advancing more slowly. The queue behind it is getting much shorter.

What this does to the price of a token

The token is the unit of account of this industry, and it is behaving like a commodity. Academic work tracking OpenRouter and Epoch AI data documents a roughly 600-fold decline in token prices since 2020. GPT-4-class capability cost around USD 30 per million input tokens in 2023; equivalent quality is now available in the region of USD 0.10 to 0.40. The spread between the cheapest and the most expensive model on the market is now measured in thousands of times.

The market has split in two. A frontier tier still commands premium pricing for genuinely hard, high-stakes, long-horizon work. Below it sits a commodity tier deflating relentlessly, with no visible floor, where suppliers compete almost exclusively on price and at least one major participant appears willing to price below cost.

The mechanism that transmits this deflation into revenue is routing. Once a company places a router in front of its applications, the model becomes a fungible input, purchased on specification, latency, context length and cost per successful outcome, rather than on brand. AT&T recently reported 56% savings on coding costs from routing alone. This is the crucial point for markets: commoditisation does not require the challengers to match the leader at the very top. It only requires them to be good enough for the 90% to 95% of workload that is not at the top. The frontier laboratories may well keep the crown and lose the volume. Volume is what pays for the data centres.

Where the value added is moving

If the model itself becomes an input rather than a product, the moat has to sit somewhere else. It is migrating towards distribution and default placement, which decide who owns the end user; towards proprietary, non-public data; towards ownership of the workflow and the system of record, where switching costs actually live; towards agents, orchestration and evaluation layers; and towards genuinely scarce physical assets such as power, grid connections, cooling, high-bandwidth memory and advanced packaging.

The analogy is cloud computing. Raw compute was commoditised and its price fell continuously for fifteen years, and yet enormous fortunes were made by those who owned the distribution and the workloads, not by whoever sold the cheapest cycle. Our own industry offers the same lesson in miniature. No private bank will derive a durable advantage from having access to a frontier model, because every competitor will have one. The advantage lies in what it is connected to: client data, regulatory perimeter, human judgement, distribution.

The assumption embedded in the capex

The four largest hyperscalers are guiding to roughly USD 725 billion of capital expenditure in 2026, up around 77% on last year, with consensus already pencilling in more than a trillion for 2027. Underneath that number lies an implicit assumption that deserves to be stated plainly: that inference will be sold at a price containing an economic rent for a small number of model providers, and that this rent will persist long enough to depreciate the assets.

If capability-adjusted prices continue to fall by a factor of four to ten per year while data centre assets are depreciated

over five or six years, that assumption requires volume to grow faster than price falls. So far, it comfortably has. The Jevons effect is real: greater efficiency is driving greater usage, with agentic workloads consuming tokens voraciously. The risk is not that demand for AI disappears. It is that unit revenue falls faster than utilisation rises, and that the marginal token ends up being served by whoever is willing to accept the lowest return on capital. Today, a Chinese laboratory or an anonymous provider giving it away for free.

And free is not charity. Someone is paying for that inference. They are buying evaluation data, developer habit, mindshare and the option to reprice later. That is precisely how a market behaves on its way to commoditisation.

Investment conclusion

The conclusion is not to sell AI. Rather, it is to be much more discerning about which parts of the AI value chain offer attractive returns.

Separate rents from revenues. For every AI-related holding, ask a simple question: does falling AI cost make this business cheaper to operate, or easier to replace? Own the first group. Be sceptical of valuations that assume a durable competitive advantage when technological leadership is increasingly measured in months rather than years.

The application layer is an underestimated beneficiary of AI. Falling inference costs reduce the expense of embedding intelligence into existing workflows, creating a powerful margin tailwind for software companies. The popular view that AI will destroy software may therefore have it backwards. As intelligence becomes abundant and inexpensive, the companies closest to the user may capture more, not less, value.

Watch the right indicators, not the benchmarks. Track token volumes, frontier-model pricing, enterprise savings from model routing, and hyperscaler depreciation trends. These data points reveal where value is accruing and where pricing power is eroding. The most important signal will be a leading Western AI lab lowering prices to protect market share. That would mark a turning point, forcing the market to reprice expectations for profitability across the AI ecosystem.

We may well find out next week that Ox Alpha is actually an unreleased Microsoft model rather than a Chinese breakthrough. But it would barely change the analysis. What matters is not who developed the model, but how difficult it was to tell the difference. The real signal is that the gap between leading models may be narrowing to the point where ownership matters less than many investors assume.

Welcome to Syzerland®

For further information

Banque Syz SA

Quai des Bergues 1
CH-1201 Geneva
T. +41 58 799 10 00
syzgroup.com

Charles-Henry Monchau, CFA, CAIA, CMT

Chief Investment Officer
charles-henry.monchau@syzgroup.com

This marketing document has been issued by Bank Syz Ltd. It is not intended for distribution to, publication, provision or use by individuals or legal entities that are citizens of or reside in a state, country or jurisdiction in which applicable laws and regulations prohibit its distribution, publication, provision or use. It is not directed to any person or entity to whom it would be illegal to send such marketing material.

This document is intended for informational purposes only and should not be construed as an offer, solicitation or recommendation for the subscription, purchase, sale or safekeeping of any security or financial instrument or for the engagement in any other transaction, as the provision of any investment advice or service, or as a contractual document. Nothing in this document constitutes an investment, legal, tax or accounting advice or a representation that any investment or strategy is suitable or appropriate for an investor's particular and individual circumstances, nor does it constitute a personalized investment advice for any investor.

This document reflects the information, opinions and comments of Bank Syz Ltd. as of the date of its publication, which are subject to change without notice. The opinions and comments of the authors in this document reflect their current views and may not coincide with those of other Syz Group entities or third parties, which may have reached different conclusions. The market valuations, terms and calculations contained herein are estimates only. The information provided comes from sources deemed reliable, but Bank Syz Ltd. does not guarantee its completeness, accuracy, reliability and actuality. Past performance gives no indication of nor guarantees current or future results. Bank Syz Ltd. accepts no liability for any loss arising from the use of this document.