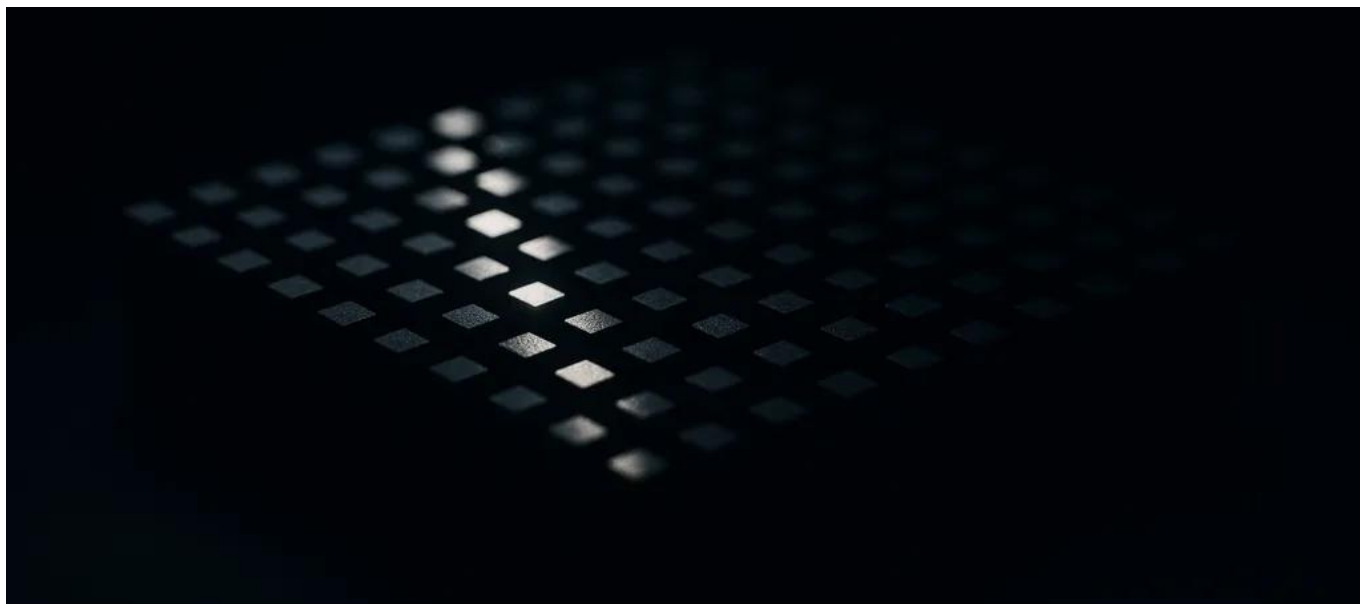




Reflection AI Ships Beam: 501B Open-Weight MoE, 23B Active,
Apache 2.0, Built at a \$25B Valuation



Reflection AI Ships Beam: 501B Open-Weight MoE, 23B Active, Apache 2.0, Built at a \$25B Valuation

Beam's pitch is not size. It is that 4.6% of the model fires per token, and Reflection AI claims that buys GLM-5.2-level reasoning at three to four times less inference compute.

Reflection AI announced Beam on October 5, 2026: a sparse mixture-of-experts model with 501 billion total parameters and 23 billion active per token, a 1 million token context window, trained on 23.8 trillion tokens. Weights and tooling were promised under Apache 2.0 during October 2026. Every launch benchmark is company-reported. The number that matters commercially is 23B active, not 501B total.

Reflection AI was founded in March 2024 by former Google DeepMind researchers Misha Laskin (CEO) and Ioannis Antonoglou (President/CTO), both tied to AlphaGo work. The company published [Introducing Beam](#) on October 5 and updated it on October 7. Beam is its first open-weight release, built for coding, reasoning and agentic workloads. Laskin



called it a “workhorse” in [Semafor’s coverage](#), which is an unusually modest word for a launch at this valuation.

Reflection has raised roughly \$4.7B in total. A ~\$2B round in 2025, including about \$800M from Nvidia, valued it near \$8B. Semafor describes the latest round as a \$25B pre-money valuation. So the first public artifact of all that capital is a model whose main selling point is that it is cheap to run.

What the 23B active number actually buys you

Sparse MoE models route each token through a subset of experts. Beam has 501B total parameters and activates 23B per token. Memory footprint tracks total parameters (you still have to hold the weights somewhere), while arithmetic per token tracks active parameters. That split is the whole argument.

Reflection claims Beam matches GLM-5.2 on advanced reasoning while using roughly three to four times less inference compute, per [TechCrunch](#) and Bloomberg. [SiliconANGLE](#) reports the company also claims Beam approaches Qwen 3.8-Max performance despite Qwen exceeding 2 trillion parameters.

4.6%

of parameters active per
token (23B of 501B)
Reflection AI blog, Oct 5 2026

23.8T

training tokens, curated web
plus licensed data
Reflection AI blog, Oct 5 2026

1M

token context window
TechCrunch, Oct 5 2026

For anyone sizing self-hosted inference, the sparsity ratio is where the money is. A dense model chasing comparable reasoning quality has to push every parameter through every token. Beam pushes 4.6% of them. The cost you pay is VRAM: you host all 501B parameters regardless of how few fire, so the hardware floor is set by memory capacity while the throughput ceiling is set by the much smaller active slice. Good trade for teams that already own the boxes, bad for teams renting per-GB.

The architectural shape is familiar. I covered the same pattern a month earlier when [Tencent open-sourced Hy4 Preview at 770B total, 49B active, 1M context, Apache 2.0](#). Beam is smaller in total and roughly half as active per token. Whether it is better depends



entirely on benchmarks nobody independent has run yet.

Every benchmark at launch came from Reflection itself

The reported scores are strong: SWE-Bench Verified 80.9, Terminal-Bench v2.1 80.1, AIME 2026 97.8, per Startup Fortune's writeup. On SWE-Bench Pro v2-Hard, Reflection reports Beam at 77.2 against Inkling at 56.9, a twenty-point gap. [Fortune](#) covered the launch the same day.

! The catch

All launch benchmarks are company-reported. TechCrunch and SiliconANGLE both note that no standardized third-party table existed at launch comparing Beam against DeepSeek or Kimi under identical prompting and tool-use conditions. The 3-4x compute claim against GLM-5.2 is also Reflection's own measurement, with no published methodology for how inference compute was counted.

This is how every model launch works now, and the numbers may well hold. But a twenty-point gap on a hard agentic benchmark and a 97.8 on AIME are the kind of figures that move procurement decisions, and procurement decisions should not run on a vendor's own scorecard. Agentic benchmarks are especially sensitive to scaffolding: the same weights with a different harness, retry policy, or tool-call format can swing results.

The other missing piece at launch: Beam was available only through an early access program, and reports do not specify eligibility, quotas or pricing. Apache 2.0 weights were promised during October 2026. At the time of the announcement they were not out.

Does Beam change the self-hosting math for European teams?

Conditionally yes, and the condition is the Apache 2.0 release landing as promised.

The constraint many European engineering teams operate under is jurisdiction, not model quality. Shipping production data to a US API triggers one compliance conversation; running Chinese-origin weights triggers a different one, usually with procurement or a customer's security questionnaire rather than with the law. Open weights under a permissive licence



remove the network hop entirely. Beam is a US-built open-weight model at this capability tier positioned explicitly against GLM-5.2 and Qwen 3.8-Max.

Apache 2.0 matters more than people assume. It grants patent rights, permits commercial use and modification, and does not impose use restrictions that a legal team has to re-read every quarter. Many “open” model releases are not this clean, a distinction I have written about before in [open source AI: myth vs reality](#). Weights under a permissive licence are still not open source in the OSI sense, since the training data and pipeline are not published, but for a CTO signing off on a deployment the licence text is what the lawyers actually read.

Reflection says it will ship weights, model card, documentation and the stack to run, evaluate and fine-tune Beam, all under Apache 2.0. If the fine-tuning stack is genuinely usable, that is worth more than the weights alone. Most teams do not need frontier general capability. They need a competent coding and tool-use model adapted to their own codebase and their own internal APIs, running somewhere they control.

The \$25B is not priced on Beam

Reflection has given away its only shipping product. That is the part worth sitting with.

MY READ

The \$25B is priced on the assumption that whoever supplies the default American open-weight model captures the enterprise layer above it: support, hosted inference, fine-tuning services, and the next model that is not open. Open weights are distribution, and distribution is the thing Chinese labs built faster than anyone expected. Reflection is buying its way into that position with roughly \$4.7B of raised capital, and the Nvidia stake of ~\$800M tells you who benefits if the default American open model happens to be an inference-hungry MoE that enterprises self-host on GPUs. None of that is stated by any party. It is my reading of the incentives.

The sparsity angle cuts the same way. A model that is cheap to run is a model more organisations will run themselves, which means more GPUs sold to more buyers rather than fewer GPUs sold to a handful of hyperscalers. I do not think that is a conspiracy. I think it is an alignment of interests that explains why the compute-efficiency framing leads the



announcement instead of a raw capability claim.

Wait for the weights, build the eval harness now

The detail to verify first is whether the Apache 2.0 release in October 2026 includes the fine-tuning and evaluation stack Reflection promised, or weights only. That decides whether Beam is a deployable platform or a benchmark announcement with a download link.

Concretely, for a team already evaluating self-hosted open-weight models:

- Wait for the weights. Early access with unspecified eligibility, quotas and pricing is not something to plan a quarter around.
- Size the memory, not the FLOPs. 501B parameters have to live in VRAM even though 23B fire. Run that number against your actual hardware before the efficiency claim means anything to your budget.
- Build your own eval harness now, against your own tasks, before any weights land. Company-reported SWE-Bench numbers tell you about SWE-Bench. Your agentic workload is not that.
- Read the model card when it ships, specifically for what “proprietary licensed datasets” covers. Reflection discloses 23.8T tokens of curated web data plus licensed data and no more. For regulated deployments, the provenance gap is the thing your auditor will ask about.

The honest uncertainty: I cannot tell from the available reporting whether Beam’s efficiency advantage survives contact with real agentic workloads, where long tool-call chains and a 1M-token context change the arithmetic considerably. A 3-4x compute advantage measured on reasoning benchmarks may compress or widen under long agent traces. Nobody has published that number.

Where this goes by mid-2027

I expect three things, stated as prediction rather than fact.

First, independent benchmarks will land within weeks of the weights dropping and will



Reflection AI Ships Beam: 501B Open-Weight MoE, 23B Active, Apache 2.0, Built at a \$25B Valuation

come in below the company-reported figures, more so on the agentic benchmarks than on AI ME. That is the usual pattern, and it would not make Beam a bad model.

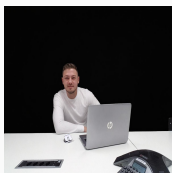
Second, the 3–4x inference compute claim will prove more durable than any individual benchmark score. Sparsity ratios are structural. Capability rankings reshuffle every quarter.

Third, I expect at least one more US lab to ship a permissively licensed frontier-tier open-weight model within six months, framed on compute efficiency rather than parameter count, because Reflection has now established that as the competitive axis against GLM–5.2 and Qwen 3.8–Max. Qwen crossing 2 trillion parameters while Beam claims comparable territory at 501B is the comparison every rival marketing team just read.

What I am least sure about is whether any of this holds commercially. A company valued at \$25B that open-sources its flagship has to monetise something else, and nothing in the launch coverage explains what that is.

WHAT TO DO WITH THIS

Beam is the most interesting open-weight release for European self-hosting since Hy4, and it is also entirely unverified: Apache 2.0 weights were promised during October 2026 and every benchmark came from Reflection. Plan an evaluation, do not plan a migration. If you are weighing self-hosted open weights against API dependency and want the jurisdictional and cost analysis done properly, [book a call](#) →



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams



Reflection AI Ships Beam: 501B Open-Weight MoE, 23B Active,
Apache 2.0, Built at a \$25B Valuation

turn AI from a chatbot into a reliable operational layer.

Book a meeting →