# Who captures the value of open AI?

A public-source briefing on model choice, operating control and inference economics

**Free edition · 19 September 2026 · Version 1.0**

Open weights expand the set of suppliers capable of serving a workload. Economic value then depends on who can deliver verified outcomes at lower total cost, who controls the customer relationship, and who retains bargaining power as alternatives improve. A downloadable model gives the buyer an option. Engineering, contracts and operating capacity determine whether that option is usable.

The practical unit of analysis is the **workload under a defined quality, latency and security requirement**. This briefing develops that unit into a cost model, a supplier-economics test and an exit test. Observations from public sources are cited; the economic mechanisms are analysis; all calculator defaults are explicitly hypothetical.

## 1. Read the denominator before declaring a winner

Vercel's latest displayed observation, September 18, reports **78.4% open-weight token share** on AI Gateway. Its definition concerns downloadable weights. On the same observation, DeepSeek V4.1 Flash accounts for **59.3% of tokens, 18.8% of requests and 5.1% of spend**. Those are three different measures of the same gateway's activity. [2]

Vercel's documentation says the figures beside models default to the most recent day, even when the chart displays a longer window. The published data are anonymized daily aggregates without absolute volumes or customer identities. A three-month chart therefore does not make its displayed end-point shares three-month averages. [3]

The economic inference is limited but useful: substantial token throughput can coexist with a much smaller share of expenditure. Workload length, prices, caching, tokenization and application mix could contribute; these aggregates do not identify their individual effects. They also leave unanswered whether traffic is broadly distributed across customers or concentrated in a few large workloads.

A forecast about the three best models requires an additional measurement system: a deadline, an openness definition, a frozen evaluation suite, comparable inference budgets, uncertainty estimates and rules for handling specialized models. Usage, frontier capability and profitability are separate hypotheses. A supplier can lead one measure and lag another without a contradiction.

For a buyer, the decisive question is narrower: which model completes this workload within its operating requirements, and what does each accepted result cost? A broad leaderboard helps generate candidates; a workload evaluation supplies the purchasing evidence.

## 2. Separate rights, capability and deployment

The Open Source Initiative's definition requires freedoms to use, study, modify and share, supported by parameters, code and sufficiently detailed training-data information. Downloading weights alone establishes less. A buyer should inspect the actual license and associated materials rather than infer permission from a marketing category. [4]

Companies can also participate in both distribution models. OpenAI released the gpt-oss models under Apache 2.0 in August 2025 while continuing its hosted model business. Google's Model Garden offers Google, open and third-party models. These are documented product choices, not evidence that either company has committed to a

permanent single-category strategy. [5] [6]

Capability is workload-specific. Stanford's 2026 AI Index describes markedly uneven performance across different tasks. For procurement, that means measuring the exact version, reasoning budget, tool environment, context length, quantization and retry policy you intend to operate. A historical gap on an aggregate benchmark does not specify the gap on your production tasks. [14]

Deployment adds another independent axis:

| Deployment | What the buyer gains | What the buyer must still verify |
|---|---|---|
| **Proprietary hosted API** | A managed service with provider-maintained models | Contract, retention, capacity, price changes and a workable fallback |
| **Open weights on a specialist platform** | A potential alternative host for the same weights | Portability of tuning, runtime, state and production behavior |
| **Open weights in a buyer-managed cloud environment** | Greater control of the deployment and data path | Cloud dependencies, administrator access, isolation and operating costs |
| **Open weights on locally operated infrastructure** | Direct control of the local serving environment | Staffing, hardware supply, software updates, security and recovery |

This table is an analytical checklist, not a universal ranking. Baseten's advertised managed, self-hosted and hybrid options illustrate why model access and operating location must be recorded separately. OpenAI's endpoint documentation similarly separates model access from data-control settings. [8] [10]

A substitution claim becomes credible when the buyer can change supplier while preserving acceptable behavior. Copying the weight files is only one step in that test.

## 3. Calculate the cost of a verified result

Use a common accounting horizon and define success before testing. Let **N** be submitted jobs, **F** fixed costs within that horizon, **v** the fully loaded variable cost per submitted job, and **p** the end-to-end fraction of jobs that pass the agreed validation and latency requirements. Include the complete retry policy in both costs and measured success.

**Cost per accepted job = (F + N × v) / (N × p)**

The numerator includes unsuccessful attempts, tool calls, review, orchestration, storage and operating labor. For owned hardware, use a consistent allocation for depreciation or economic capital cost. For rented hardware, avoid adding the same hardware expense again under depreciation. Keep the horizon identical across alternatives and show which costs are incremental versus shared.

The following example uses invented inputs to expose the mechanism. It is neither a vendor quote nor a model benchmark.

| Hypothetical configuration | Fixed cost per horizon | Variable cost per submitted job | Verified success fraction |
|---|---|---|---|
| **A: managed API** | $0 | $0.10 | 95% |
| **B: buyer-operated open model** | $12,000 | $0.03 | 85% |

At **100,000 jobs**, A costs $10,000 and produces 95,000 accepted jobs, or **$0.1053 per accepted job**. B costs $15,000 and produces 85,000 accepted jobs, or **$0.1765**. At **1,000,000 jobs**, the corresponding costs are **$0.1053** and **$0.0494**, provided the assumed cost and success functions remain unchanged.

The mathematical crossover occurs at approximately **201,770 submitted jobs**. That is a conditional result of the assumptions, not a forecast about the utilization or staffing capacity of a real deployment. Capacity expansion, burstiness, queueing, new hardware and operational headcount can change both the fixed and variable terms.

There is also a hard acceptance constraint. With a required success rate of 90%, B is ineligible at either volume, regardless of its unit cost. Normalizing by accepted jobs does not erase failed demand, compensate users for errors, or make different service levels interchangeable. A fallback route needs its own measured costs, conditional success rates and latency.

The accompanying browser calculator exposes every input, shows the acceptance gate and exports its assumptions. It does not infer correlated retries, failure losses, capacity steps or quality from prices. Its role is to make the arithmetic inspectable before a real workload supplies the evidence.

## 4. Test where the supplier can retain a margin

An increase in qualified suppliers gives buyers additional negotiating options. That can pressure the owner of a model, the company hosting it, or both. The beneficiaries may include customers through lower prices, and no accounting identity assigns all the savings to the serving platform.

A serving business can charge for genuine work: improving throughput at a latency target, handling bursts, managing capacity, securing deployments and reducing the customer's operating burden. Fireworks publishes both token-based serverless billing and GPU-time deployment billing. Baseten offers infrastructure spanning its managed environment and customer-controlled deployments. Those offerings participate in related markets with different cost structures. [10] [11]

Large incumbent clouds also have access to open-model demand. Google's documented model catalog is enough to establish that downloadable weights do not reserve the hosting opportunity for a particular group of newer providers. Which supplier retains value depends on relative cost, deployment requirements, customer distribution, utilization and contract terms. [6]

A financial example shows why revenue growth needs a capital-accounting companion. For Q2 2026, Nebius reported **$582.3 million consolidated revenue**, **$236.2 million adjusted EBITDA**, and a **$190.4 million net loss from continuing operations**. Its operating cash flow was **$2,246.1 million**, while purchases of property, equipment and intangible assets were **$5,657.4 million**. Subtracting the latter from the former gives **negative $3,411.3 million**. [12]

These are unaudited group figures, including businesses beyond inference. The cash-flow subtraction is our calculation, not a company-reported free-cash-flow measure. Growth investment can support future earnings; the figures alone establish neither investment failure nor sustainable long-term returns. They establish the need to track capital deployment alongside growth. [12]

For an infrastructure thesis, request realized economics by hardware generation and workload cohort: contracted and achieved utilization, revenue per capacity unit, operating cost, refresh assumptions, customer concentration, cancellation rights and capital needed to serve renewals. For a software serving layer, ask how much of gross profit survives its underlying cloud bill, support requirements and competitive repricing.

Token growth also leaves spending growth unresolved. Total spend depends on the number of jobs, compute consumed per job and price per compute unit. Lower prices may expand demand, while efficiency lowers the resources needed for each completed task. The resulting revenue and margin depend on the sizes of those effects. A rising token counter supplies only part of the equation.

## 5. Protect proprietary knowledge at the actual data boundary

A request containing business context creates a data-handling exposure. The exposure depends on the receiver, the processing path, the permitted uses, retention and access controls. It does not automatically transfer ownership of the business's knowledge or establish that the provider trains on it.

OpenAI states that business/API data are not used for training by default. Anthropic states the same default for its commercial inputs and outputs, with exceptions such as explicitly submitted feedback. These are published product policies; they are not independent audits of implementation, and consumer offerings have separate terms. [7] [9]

Retention is a different question. OpenAI's documentation separately describes abuse-monitoring logs, application state and eligibility or limitations for retention controls. The exact endpoint, features and settings matter. Buyers should preserve the applicable policy and contract version and test that the deployed configuration matches the intended boundary. [8]

The same scrutiny applies to a third party hosting open weights. Model portability does not make prompts private from the operator processing them. A locally controlled deployment can reduce some external exposures, while assigning more responsibility for patching, identity, secrets, logs and recovery to its operator. Evaluate the actual data flow rather than using the license as a security certificate.

An enterprise should separately control its documents, retrieval permissions, task definitions, evaluation examples, correction history and action-approval rules. Treat the model as one dependency behind that control structure. To verify an exit path, export those assets, revoke the old provider's credentials and run a representative workload through another approved route. Record the engineering time, residual dependencies and behavioral regressions.

Routing can help, provided the router is evaluated too. RouteLLM, submitted in 2024 and revised in 2025, supplies research evidence for learned routing between stronger and weaker models on its tested benchmarks. It does not promise the same savings on a new customer workload. Measure routing mistakes, privacy restrictions, evaluation cost and fallback behavior before assigning value to a multi-model design. [13]

The operational asset is the ability to make and verify a change. A diagram with several model logos demonstrates an intended architecture; a successful migration with preserved controls demonstrates an exercised option.

## 6. Run a decision process that can reject a preferred narrative

Start with a defined workload distribution, including difficult and high-cost cases. Freeze a held-out evaluation set, access permissions, retry budgets and a validation rubric. Choose sample sizes to support the desired uncertainty bounds. Measure candidates on matched jobs where possible, and report the uncertainty around both cost and success rather than publishing a single favorable average.

Separate shadow evaluation from production authority. A candidate can observe or propose before it receives permission to take external actions. Establish stop conditions for data leakage, unauthorized actions, latency violations and unacceptable error rates. A low price cannot waive a hard constraint.

| Decision | Evidence required | Condition that changes the decision |
|---|---|---|
| **Shift a workload to open weights** | Matched-task results meeting quality and latency floors, including review and retry costs | Savings disappear under real load, or the candidate misses a required constraint |
| **Pay a frontier-model premium** | Measured additional successful work or reduced downstream loss worth more than the premium | A qualified alternative supplies the same useful outcome at lower full cost |
| **Commit to a hosting platform** | Capacity, reliability, security controls and economics under an actual contract | Repricing, congestion or retained dependencies invalidate the assumed advantage |
| **Claim meaningful portability** | A completed export and migration with known time, cost and regressions | Critical state, permissions or behavior remain tied to the original provider |

This is a proposed decision protocol. It is not a claim that any named provider has passed these tests. After every material model, price or infrastructure change, rerun the affected slices. Preserve failures and superseded configurations so that a later result can be reconciled to an earlier claim.

For an investment narrative, preregister what would count as confirmation. More open-model traffic would support adoption. Improving retained unit margins and returns on deployed capital would support value capture by a supplier. Successful switches with lower total cost would support buyer bargaining power. The three findings require different evidence.

## 7. Keep the commercial ledger separate from the evidence ledger

Subscription growth is useful feedback about demand for a paid bundle. Price also selects for willingness and ability to pay. Inferring analytical accuracy from those observations requires an additional causal argument: acquisition channels, audience composition, access benefits and churn can change without a corresponding change in factual quality.

A research product can publish dated claims, named sources, explicit assumptions, forecasts with terminal conditions, alternative explanations and a correction history. Readers can then evaluate what the work establishes independently of its price. Free material deserves the same scrutiny, and a free substitute proves its value through the work it enables.

The defensible open-AI thesis is therefore conditional: downloadable weights can increase buyer choice when licenses permit the intended use, models meet the workload requirements, and a buyer can actually operate or move the deployment. Suppliers earn durable returns only to the extent that their remaining advantages survive that choice. An operating plan should preserve the evidence and exit paths needed to test both sides.

# Provenance and limitations

**Scope.** This is independent analysis prepared from public primary sources on September 19, 2026. Social Capital's current public preview identified the open-versus-closed AI topic and advertised a 99-page subscriber report. The paid deck was not accessed or reproduced, and this briefing does not purport to summarize or assess its unseen contents. [1]

**Evidence boundary.** Gateway figures describe one service on a specified day. Vendor documentation establishes stated offerings and policies, not independent verification of their implementation. The Nebius figures are unaudited, consolidated and historical. The routing paper supports a method under its evaluated conditions. Derived mechanisms are analysis; the numerical cost example and every calculator default are illustrative assumptions.

**Operational boundary.** No model evaluation, infrastructure benchmark, penetration test or customer-data audit was performed for this briefing. The calculator normalizes assumed cost by assumed verified success; it does not measure either. A real purchasing decision also needs workload-specific error costs, capacity, contracts and security review. No named company's private margins or valuation have been estimated.

# Sources

All sources were checked on September 19, 2026. Dates below distinguish publication dates from access dates. Numbers in the text point to the source entry, whose scope records the limit of its use.


**[1] Social Capital / Chamath Palihapitiya. Deep Dive: The Open vs. Closed AI Race (public preview).** 2026-09-18. Identifies the topic and advertised subscriber report. The paid PDF was not accessed; this briefing makes no claim about its full contents.

**[2] Vercel. AI Gateway model leaderboards.** Latest displayed observation: 2026-09-18. AI Gateway traffic only. Open weights 78.4%; DeepSeek V4.1 Flash token share 59.3%, request share 18.8%, spend share 5.1%. Shares do not measure global market share or model capability. Vercel licenses its leaderboard data CC BY 4.0.

**[3] Vercel. AI Gateway Leaderboards: methodology and exports.** Accessed 2026-09-19. States that displayed model/lab shares default to the most recent day, and that data are anonymized daily aggregates without absolute volumes or customer identifiers.

**[4] Open Source Initiative. The Open Source AI Definition, version 1.0.** Version 1.0; accessed 2026-09-19. Use, study, modify and share freedoms; parameters, code and sufficiently detailed training-data information. This is OSI's definition, not a claim that every vendor uses the term consistently.

**[5] OpenAI. Introducing gpt-oss.** 2025-08-05. Historical example of a provider offering both hosted proprietary models and Apache-2.0-licensed open weights. No current model-performance ranking is inferred.

**[6] Google Cloud. Model Garden.** Accessed 2026-09-19. Documents access to Google, open and third-party models. Establishes availability, not independently measured quality or provider margins.

**[7] OpenAI. Enterprise privacy at OpenAI.** Updated 2026-01-08. Business/API data are not used to train models by default. This is a stated policy, not an audit of implementation or a universal no-retention promise.

**[8] OpenAI. Data controls in the OpenAI platform.** Accessed 2026-09-19. Distinguishes training use, abuse-monitoring retention, application state and eligibility/limitations of retention controls. Exact endpoint and feature selection matter.

**[9] Anthropic. Is my data used for model training?.** 2026-08-18. No training on commercial inputs/outputs by default, with exceptions including explicitly supplied feedback. Consumer plans are described separately.

**[10] Baseten. Inference platform and deployment options.** Accessed 2026-09-19. Managed, single-tenant, self-hosted/VPC and hybrid options. Vendor statements establish offered configurations; they do not independently verify reliability or cost advantages.

**[11] Fireworks AI. Pricing.** Accessed 2026-09-19. Per-token serverless and GPU-time on-demand billing. No live vendor rates are used as defaults in the illustrative calculator.

**[12] Nebius Group. Second-quarter 2026 financial results.** 2026-08-12. Quarter ended June 30, 2026; pages 1 and 8. Group figures include other businesses and are not standalone inference unit economics. EBITDA is non-GAAP. The cash-flow subtraction in this briefing is our calculation, not a company-reported FCF measure.

**[13] Ong et al.. RouteLLM: Learning to Route LLMs with Preference Data.** Submitted 2024-06-26; revised 2025-02-23. Evidence that learned routing can trade off model quality and cost on evaluated benchmarks. Does not establish savings on a new enterprise workload or September 2026 model pairs.

**[14] Stanford HAI. 2026 AI Index Report.** 2026; retrospective performance evidence. Describes uneven performance across task families. Used for the limitation of single-number capability comparisons, not as a live September ranking.