

Serverless open-model inference lands builders on per-token APIs with no provisioning; Batch and cached-input discounts cut cost on async and repeated prompts; Dedicated Inference sells H100 GPU-hours (list \$5.49, promo \$3.99 through 09/30/26) and B200 at \$8.99; GPU Clusters rent H100 from \$3.99 on-demand down to \$3.19 reserved (91–180 days) with preemptible at \$1.99; Fine-Tuning, Sandbox, and Storage attach usage; Enterprise/PTU sales close steady production and residency buyers.

KEY PARTNERS

- **NVIDIA** — investor + Cloud Partner narrative (FACT blog/news).
- **Open-weight labs hosted** — DeepSeek, Nemotron, MiniMax, Kimi, GLM, Llama, Qwen (FACT blog + pricing catalog).
- **Series C investors** — Aramco Ventures, Vista, GC, Emergence, Salesforce Ventures, Lux, others (FACT blog).
- 500 MW+ compute capacity commitments capitalized by investors (FACT blog).
- **Named customers** — Cognition, Decagon, Eleven Labs, Cursor, Suno (FACT blog claim).
- OpenAI-compatible API ecosystem for tool/SDK partners (FACT docs).

KEY ACTIVITIES

- Operate serverless multi-model inference APIs.
- Provision Dedicated endpoints + GPU clusters (on-demand/reserved/preemptible).
- Ship Fine-Tuning (SFT/DPO), Sandbox, Code Interpreter, Storage.
- Advance kernels/compilers and model serving (company research narrative).
- Publish pricing + docs; run Batch and cache optimizations.
- Sell PTUs / enterprise capacity for steady production.

KEY RESOURCES

- **GPU inventory** — H100, H200, B200, GB200/B300 paths (FACT pricing tables).
- 500 MW+ committed capacity pipeline (FACT Series C blog).
- Serverless + Dedicated + Cluster product surfaces (FACT pricing).
- OpenAI-compatible inference APIs + model catalog (FACT docs).
- **Research/serving IP narrative (FlashAttention-4, Megakernel** — company blog).
- Named production customer relationships (FACT blog claim).

VALUE PROPOSITIONS

- **Serverless** — 100+ open models, no provisioning, pay per use (FACT docs).
- **OpenAI-compatible APIs** — swap providers with less rewrite (FACT docs).
- Batch up to 50% off; cached input discounts on select models (FACT).
- Dedicated single-tenant GPUs for latency/custom models (FACT pricing).
- GPU Clusters on-demand → reserved → preemptible (FACT pricing).
- Fine-Tuning + Sandbox + Storage on the same platform (FACT pricing).
- **Company claim** — 6x–20x lower cost vs closed frontier at scale (blog).

CUSTOMER RELATIONSHIPS

- Self-serve signup + public pricing cards (FACT site).
- **Docs-led expansion** — serverless → dedicated when busy (FACT DMI docs).
- Batch/cache education to keep cost competitive (FACT docs).
- PTU estimator + Contact sales for reserved capacity (FACT pricing).
- Startup Accelerator program narrative (company blog index).
- Reference customers named in Series C communications (FACT claim).

CHANNELS

- **Primary** — together.ai + /pricing public cards.
- docs.together.ai quickstarts and model catalogs.
- OpenAI-compatible API / SDK discovery.
- **Company blog** — research, model library, funding narrative.
- NVIDIA Cloud Partner + marketplace surfaces (company news).
- Contact sales for PTUs, reserved clusters, enterprise.

CUSTOMER SEGMENTS

- App/agent builders on serverless open models (FACT docs ICP).
- **AI-native production companies (named blog logos** — FACT claim).
- Teams needing dedicated/custom model endpoints (FACT pricing).
- ML orgs training or self-serving on GPU clusters (FACT pricing).
- Residency/compliance buyers pinning dedicated regions (FACT docs).
- Products attaching Fine-Tune + Code Sandbox (FACT pricing).

COST STRUCTURE

- GPU / power / colo COGS behind serverless + dedicated (INFERENCE).
- Cluster inventory and reserved-capacity utilization risk (INFERENCE).
- Serving-stack and kernel R&D (company research narrative).
- Fine-tune + sandbox compute COGS (INFERENCE from pricing).
- MW-scale capacity commitments with investors (FACT blog claim).
- **GTM** — PLG content + enterprise sales (INFERENCE).
- No public Together AI P&L line invented.

REVENUE STREAMS

- **Serverless** — PAYG tokens/images/audio/video; no minimums (FACT).
- **Examples** — Llama 3.3 70B \$1.04/\$1.04; gpt-oss-120B \$0.15/\$0.60 per 1M (FACT pricing).
- Batch up to 50% off; cached input discounts (FACT docs).
- **Dedicated** — H100 \$5.49/hr (promo \$3.99 thru 09/30/26); B200 \$8.99 (FACT).
- **Clusters** — H100 \$3.99 OD → \$3.19 reserved (91–180d); preemptible \$1.99 (FACT).
- **Fine-Tuning** — \$/1M tokens + minimums; Sandbox/Storage attach (FACT).
- **PTU / Enterprise** — contact sales for reserved throughput (FACT).

Sources: together.ai/blog/announcing-our-series-c · together.ai/pricing · docs.together.ai/docs/serverless · docs.together.ai/docs/dedicated-endpoints
Not affiliated with Together AI.