together.ai

Generation Digital

The Together AI Platform

Accelerate training, fine-tuning and inference on performance-optimized GPU clusters

Reliable at production scale

Built for scale, with customers going to trillions of tokens in a matter of hours without any depletion in experience.

Industry-leading unit economics

Continuously optimizing across inference and training to keep improving performance, delivering better total cost of ownership.

Frontier AI systems research

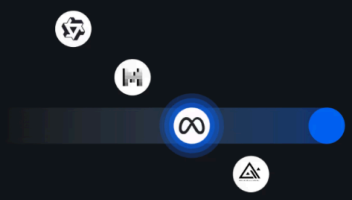
Proven infra and research teams ensure the latest models, hardware, and techniques are made available on day 1.

Industry leading AI research and open-source contributions

- FlashAttention
- Mixture of Agents
- Dragonfly
- Red Pajama Datasets
- DeepCoder
- Open Deep Research
- Flash Decoding
- Open Data Scientist Agent

Maximize your generative AI investments and take control over your models

01



Run production-grade inference

Run popular models like Llama or Qwen, or your own custom model, on our highly scalable Together Inference Engine. Deploy on our serverless or dedicated endpoints, inside your VPC or on Together GPU Clusters to match your workload needs.

02

Fine-tune and experiment easily

Easily fine-tune and deploy new models for testing. Manage and orchestrate all your models in a single place. Quickly iterate and test different configurations for optimal performance.

03

Optimize performance & cost

Achieve the lowest price and latency, with the best accuracy, for your use case. Automatically fine-tune models, and use adaptive speculators and model distillation to drive better performance and costs for your models.

Flexible deployment options

Together Cloud

- Get started quickly with fully managed serverless endpoints with pay-per-token pricing
- Dedicated GPU endpoints with autoscaling for consistent performance

Your Cloud

- Dedicated serverless deployments in your cloud provider
- VPC deployment available for additional security
- Use your existing cloud spend

Together GPU Clusters

- For large-scale inference workloads or foundation model training
- NVIDIA H100 and H200s clusters with Infiniband and NVLink
- Available with Together Training and Inference Engines for up to 25% faster training and 75% faster inference than PyTorch

Enterprise-grade security and data privacy

We take security and compliance seriously, with strict data privacy controls to keep your information protected. Your data and models remain fully under your ownership, safeguarded by robust security measures.

AICPA Service Organization Control Reports
AICPA SOC 2
Formerly SAS 70 Reports

HIPAA Compliant