Predibase

Train the best models. Serve at maximum speed.

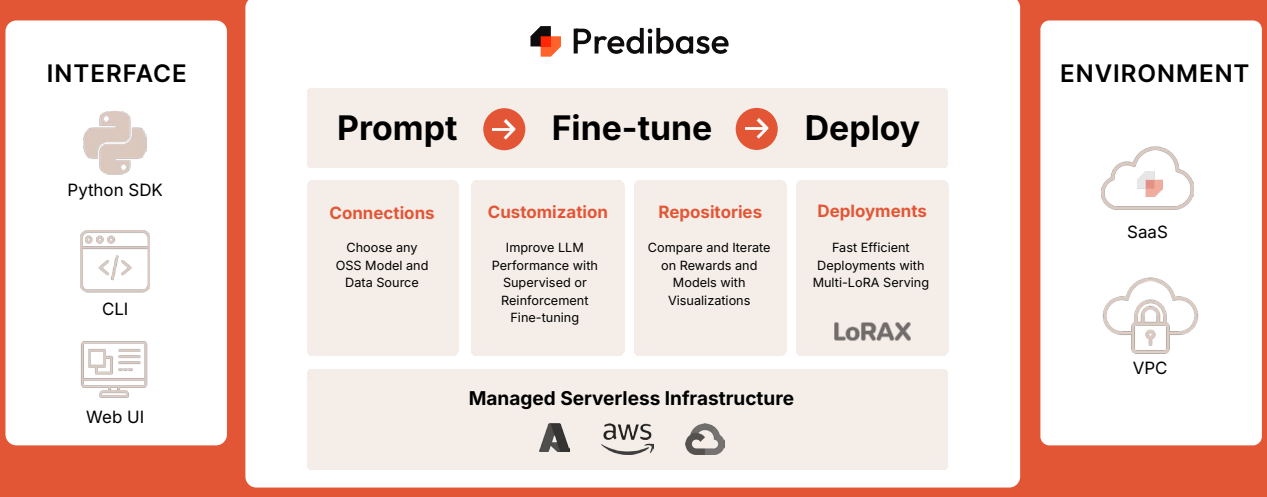
Managed Infra for Open-source AI

Enterprises are going all-in on open-source LLMs driven by the need for more control over their models, higher throughput and better accuracy.

GPT-4 Performance with Small, Lightning-Fast Models in your Cloud

Predibase makes customizing any open-source LLM effortless with the first managed platform for reinforcement fine-tuning. Our cost-effective serverless infrastructure scales to handle massive workloads, and with LoRAX — the leading open-source framework for multi-LoRA serving — you can deploy hundreds of fine-tuned models on a single GPU without sacrificing performance.

The Developer Platform for Open-source AI



Better Models with 1,000x Less Data

Transform any open-source LLM into a GPT-4–level reasoning powerhouse using fewer than 10 labeled examples.

Serve 100s of LLMs 10x Faster

Deploy dozens or hundreds of fine-tuned LLMs on a single GPU with blazing-fast inference, powered by LoRAX and Turbo LoRA.

Deploy in Our Cloud or Yours

Securely run mission-critical AI workloads in your own private environment — maintaining full data control and model ownership.

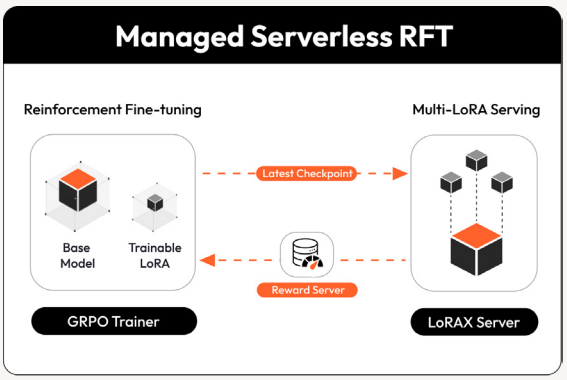
Predibase

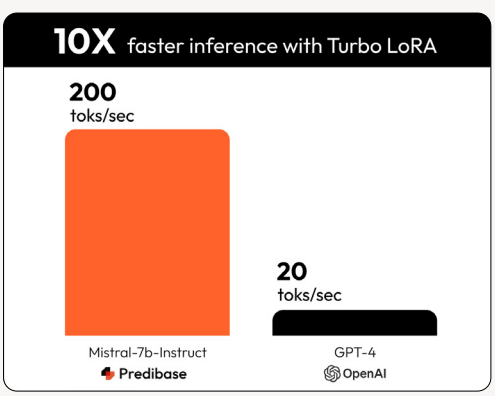
Adapt and Serve Models the Easy Way

Predibase offers a full suite of innovative capabilities to reliably and cost-effectively maximum the performance of any open-source LLM — all in just a few lines of code.

The First Platform for Reinforcement Fine-tuning

- Fine-tune powerful models with just a few examples — increase accuracy by 10-20%
- Models continuously learn from feedback using RFT for ever-improving accuracy
- Boost your LLMs with targeted rewards for deeper understanding and smarter decision-making





The Fastest, Most Efficient Multi-LoRA Serving

- Scale up or down automatically and only pay for the compute you need; reserve A100 or H100 GPUs for guaranteed capacity
- Dynamically serve 100s of fine-tuned LLMs on a single GPU with LoRAX, cutting infrastructure costs by up to 10x
- Unleash faster throughput with Turbo LoRA — no compromise on model accuracy

Built for Mission-Critical Workloads

- Flexible deployment in our cloud or your VPC (AWS, Azure, GCP)
- Peace of mind with multi-region failover, blue/green deployments, SLA guarantees, and robust logs/metrics
- Keep data private with our SOC 2 Type II–certified platform, no data sharing required

In production with the world's most innovative companies

sense · nubank · checkr · MarshMcLennan · Forethought · Qualcomm · convirza

TRY PREDIBASE FOR FREE:
pbase.ai/GetStarted