Predibase

Smaller, Cheaper, Faster And Fine-Tuned

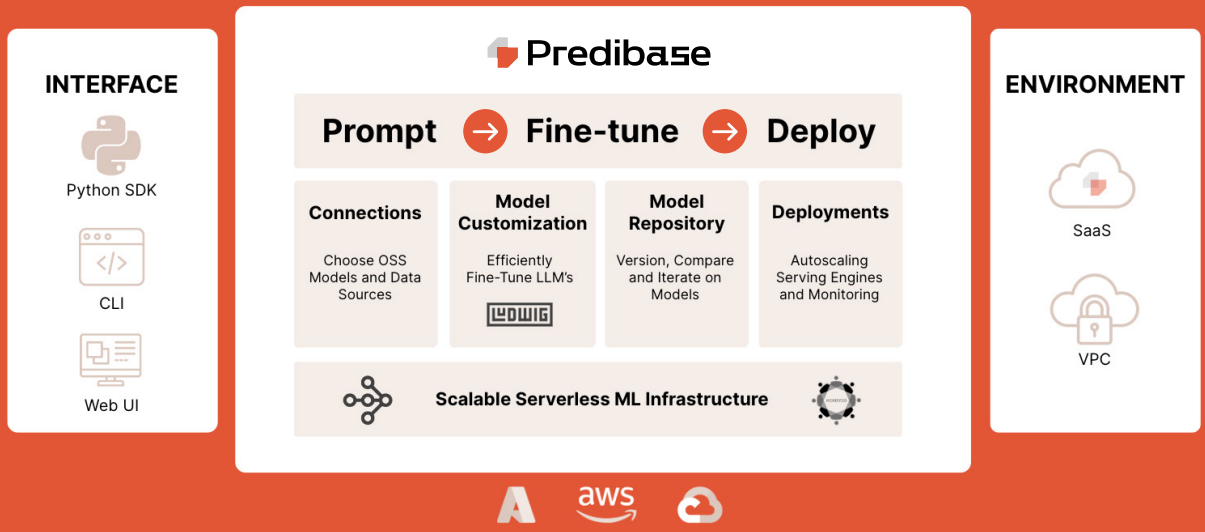
Open Source AI Infra

Over 75% of organizations won't use commercial LLMs in production due to concerns over privacy, cost and security, but productionizing open-source LLMs comes with its own set of infrastructure challenges.

Predibase: The fastest most efficient way to productionize open-source AI

Teams want to customize LLMs without giving commercial vendors access to their proprietary data. Predibase addresses these challenges with the first open-source AI infrastructure platform designed to help developers fine-tune and serve smaller, faster task-specific LLMs without comprising on performance. Built on best-in-class managed infrastructure, Predibase is the fastest way to fine-tune and deploy open-source LLMs in your cloud.

The Developer Platform for Open-Source AI



Private LLMs That You Own

Deploy and customize the latest open-source LLMs — like Llama-2-70b — in your cloud

Smaller, Faster and Trained on Your Data

Compress and efficiently fine-tune LLMs for your task with best practice optimizations

World-class Managed Infra

Serverless autoscaling infra — from T4s to A100s — that can dynamically serve 100s of models on a single GPU

Predibase

Easy and efficient model fine-tuning and serving

Predibase offers a full suite of innovative capabilities to reliably and cost-effectively productionize even the largest open-source LLMs — all in just a few lines of code.

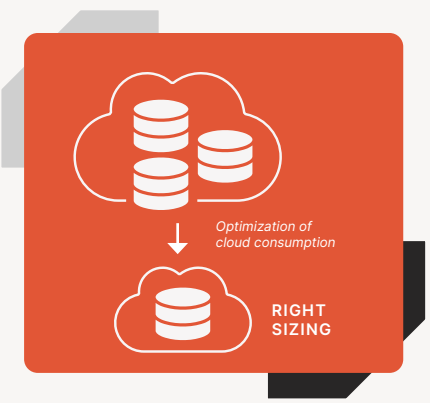
Automatic Memory-Efficient Fine-Tuning

- Reliably fine-tune any open-source LLM on even the cheapest most readily available GPUs
- Simply specify your base model, dataset, and prompt template in a few lines of code to get started
- Out-of-the-box optimizations like LoRA along with right-sized compute ensures training succeeds–no more OOMs

```
# Specify a Huggingface LLM to fine-tune
llm = pc.LLM("hf://meta-llama/Llama-2-13b-hf")

# Kick off the fine-tune job
job = llm.finetune(
    prompt_template=prompt_template,
    target="output",
    dataset="s3_bucket/code_alpaca",
    repo="finetune-code-alpaca"
)

# Stream training logs and metrics
job.get()
```



Serverless Right-Sized Training Infrastructure

- Built-in orchestration logic determines the most cost-effective hardware for each training job including support for A100s
- Built-in fault tolerance and metric and artifact tracking
- One-click deployment capabilities

Cost-Effective Serving for Fine-Tuned Models

- Only pay for compute you use by automatically scaling your deployment up and down with traffic
- Dynamically serve hundreds of fine-tuned LLMs from a single GPU and cut costs by 100x
- Load and query each fine-tuned LLM in seconds

```
# Specify a base LLM and fine-tuned LLM
base = pc.LLM("pb://deployments/llama-2-13b")
model = pc.get_model("finetune-code-alpaca")

# Prompt the fine-tuned LLM instantly using
# dynamic adapter loading
finetuned_deployment = base.with_adapter(model)
finetuned_deployment.prompt(
    "Write an algorithm in Java to reverse "
    "the words in a string."
)
```

Built by AI leaders from Uber, Apple and Google and developed and deployed with the world's most innovative companies.

payscale Uber Koo Adobe KOBLE WWF PARADIGM sekure

TRY PREDIBASE FOR FREE:
pbase.ai/GetStarted