

SERVE

BENCHMARK

GOVERN

ROUTE

Not every query needs a frontier model. Most don't.

One-click Terraform that stands up vLLM serving on your own GPUs, benchmarks it against your frontier baseline, governs every call through one gateway, and routes each query to the right tier.

INDUSTRIES

BFSI & fintech · Retail & D2C · Manufacturing · Media & OTT · Healthcare · PSU & government

WORKS WITH

Terraform · vLLM · LiteLLM · Bedrock · SageMaker · EC2 / ECS

BUILT FOR

CTOs & heads of platform · AI/ML platform engineering · Heads of data & infrastructure · CIOs managing AI spend

One-click

TERRAFORM: SERVE → BENCHMARK → GOVERN → ROUTE

Your GPUs, your data

ON-PREM, EC2, ECS OR AMAZON SAGEMAKER

Predictable TCO

PER-QUERY ROUTING ON COST × QUALITY, FRONTIER INCLUDED

Frontier or open-weight? The answer is both.

APPROACH A

Frontier-only

Top quality on everything — but every query, easy or hard, runs at frontier cost, and there's no owned open-weight option for high-volume workloads.

APPROACH B

Open-weight, self-built

Great for cost control, but you hand-build vLLM, benchmarking, gateways and routing — months of platform engineering before the first governed token ships.

HUMAINEETI

MODELSTACK

Stands up model serving, benchmarking, governance and intelligent routing (frontier included) in one click. More models, proven quality, governed, secure, predictable TCO.

Serve · benchmark · govern · route vLLM serving with your chosen models, quality and load testing, a governed model gateway, and per-query routing to the right model.

SERVE · BENCHMARK · GOVERN · ROUTE			
Serve Stands up vLLM serving with your chosen models	Benchmark Provides quality and load test capability	Govern Sets up a governed model gateway	Route Routes each query to the right model

At a glance

	FRONTIER-ONLY	OPEN-WEIGHT DIY	MODELSTACK
Frontier models for the hardest, highest-stakes queries	✓	—	✓
Open-weight models you own & can fine-tune	—	✓	✓
Runs on your GPUs — on-prem, EC2, ECS, SageMaker	—	✓	✓
One governed gateway across frontier & open-weight	—	—	✓
Built-in functional + load benchmarking, with report	—	—	✓
One-click Terraform: serve → benchmark → govern → route	—	—	✓
Per-query routing to the right tier (cost × quality)	—	—	✓

What you get

- › vLLM serving stood up with your chosen open-weight models
- › One governed gateway across frontier and open-weight models
- › Runs on your GPUs — on-prem, EC2, ECS or Amazon SageMaker
- › Functional and load benchmarking built in, with a report you can circulate
- › Per-query routing to the right tier on cost × quality
- › You keep control of the models, the data and the bill

INCLUDED FREE · DEMO + POC · YOUR WORKLOAD · YOUR GPUS OR ACCOUNT · NO COST

Benchmark open-weight against your frontier baseline — at no cost.

HOW IT RUNS

- We stand up the stack in your environment with one Terraform apply
- We benchmark open-weight candidates against your frontier baseline
- You see quality, latency and cost per query side by side

WHAT YOU PROVIDE

- A target environment — on-prem GPUs, EC2, ECS or SageMaker
- A representative sample of your real prompts and expected outputs
- Your current frontier model and spend, as the baseline

WHAT YOU GET BACK

- A functional and load benchmark report against your baseline
- A routing recommendation: which queries need frontier, which don't
- A projected TCO at your volume, with the assumptions shown

ONE FREE ENGAGEMENT · DEMO + POC

Find out what you are overpaying for.

humaineeti.ai/modelstack