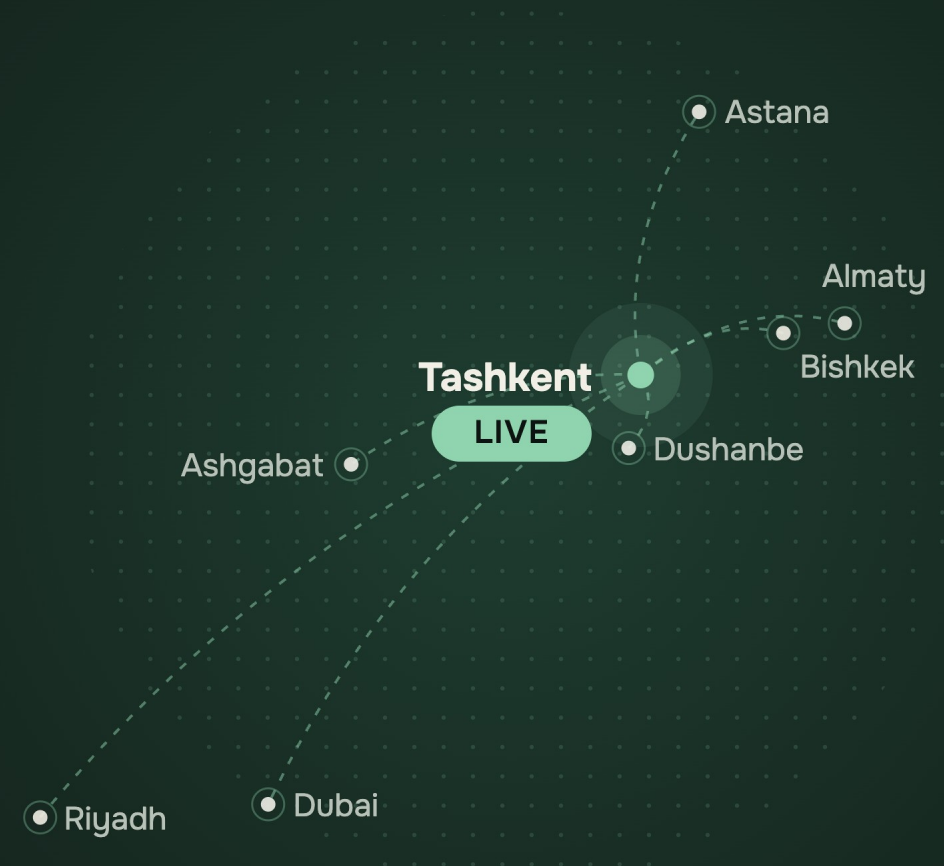


# The inference layer for Central Asia

Run, fine-tune and serve open models, built for Central Asia's companies and languages.



“

*Every country needs to own the production of their own intelligence.*

Jensen Huang, NVIDIA · World Governments Summit, February 2024

For Central Asia, that starts with open models, tuned for its languages and running in the region.

## WHY NOW

# Open models lead in usage.

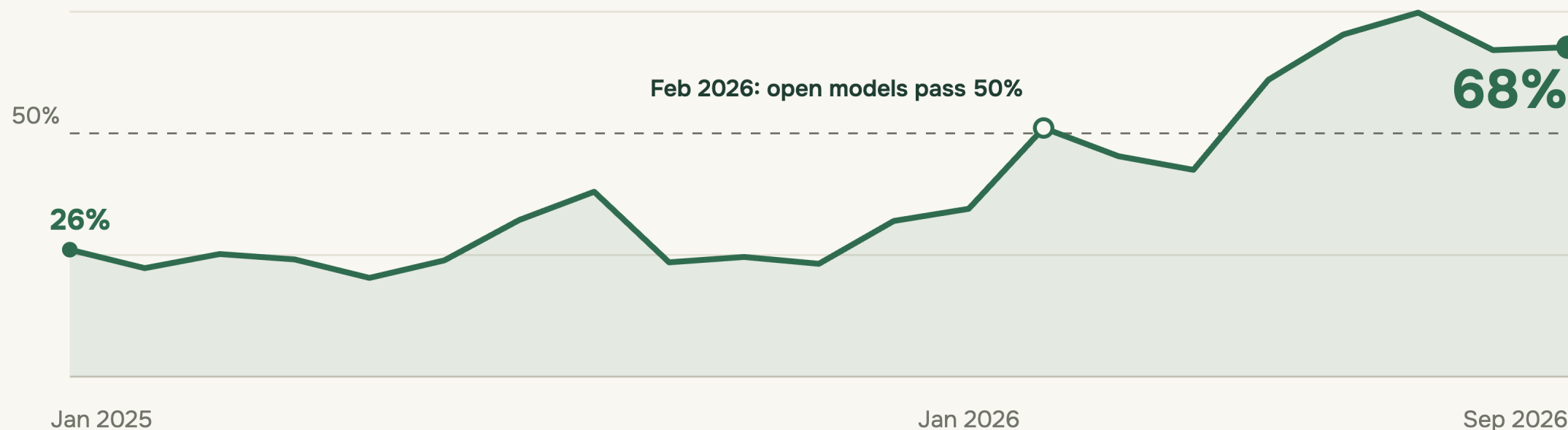
Among OpenRouter's top 50 models, open-weight ones handled 68% of tokens in September 2026, up from 26% in January 2025.

**3.1M**

open models on  
Hugging Face

**~4 mo**

behind the closed  
frontier, per Epoch AI



## THE PROBLEM

# Using AI is still hard in Central Asia.

Today a company has three options, and none of them works well.



### Foreign AI APIs

- Paid in USD by card, under foreign contracts
- Served from far away, adding latency to real-time apps
- Data leaves the country, a blocker in regulated sectors



### Local clouds

- Rent out virtual machines and GPUs
- No self-serve, per-token API for open models
- No managed fine-tuning, so teams do it themselves



### Build it yourself

- A local provider quotes a 35-day wait for an H200
- Import, capex and a datacenter to operate
- An AI team to make models fast, cheap and fluent in Uzbek

SO WE BUILT COREAI

# Fireworks and Baseten, built for Central Asia.

Pick an open model, fine-tune it on your data and call it through one API, on GPUs in the region, billed locally.

## PRODUCT

# One console for every model.

### Endpoints

Managed model endpoints with an OpenAI-compatible API

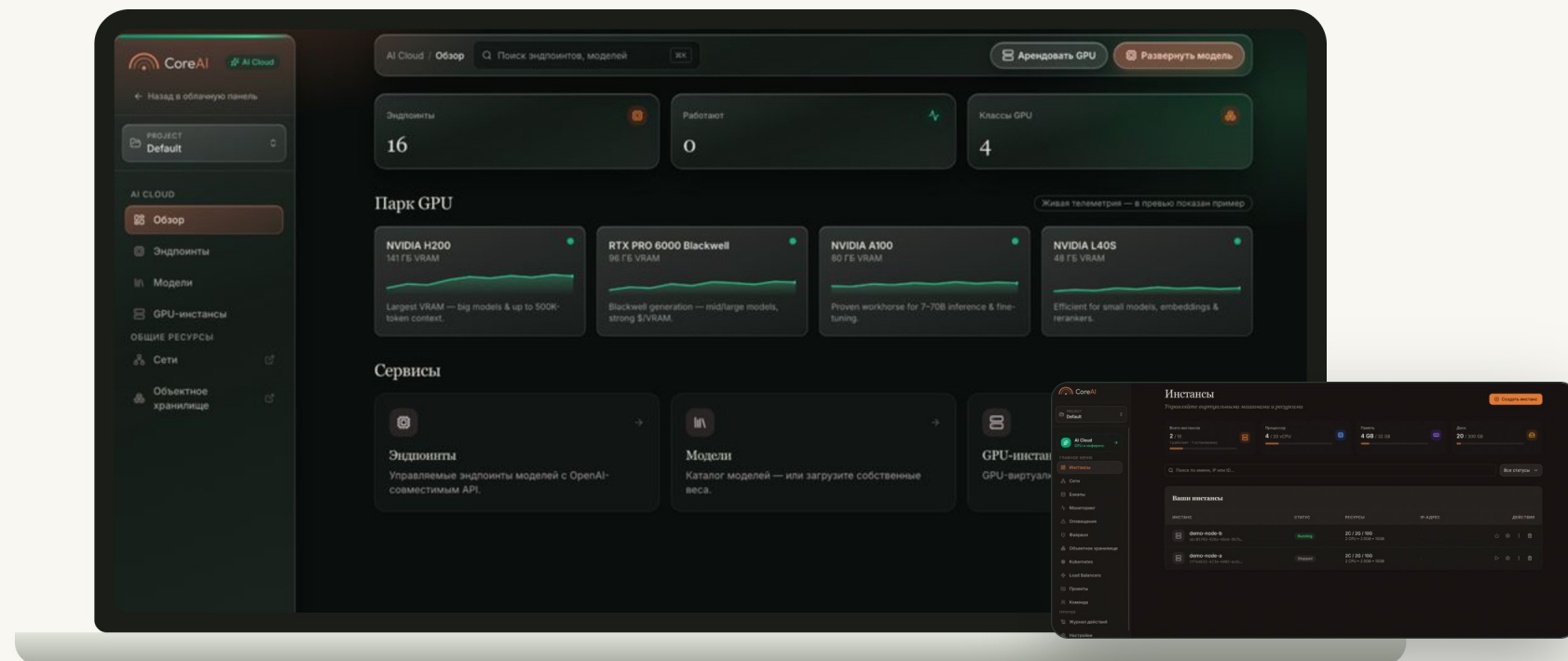
### Models

An open-model catalog, or bring your own weights

### GPU instances

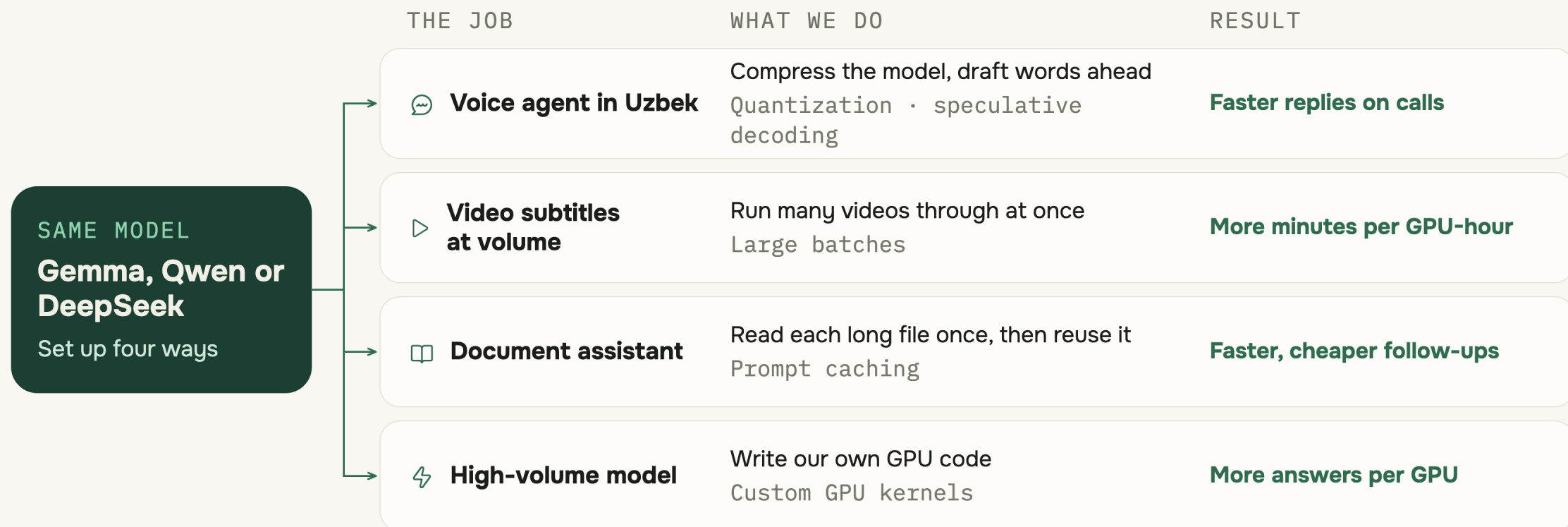
GPU machines on your own networks and SSH keys

The live CoreAI console



## INFERENCE ENGINEERING

# We make open models **fast and cheap** for each job.



We test every setup for speed, cost per answer and quality in Uzbek and Russian.

## DEPLOYMENT PASSPORTS

# Customers deploy a tested recipe in one step.

## DEPLOYMENT PASSPORT

Example

## Voice agent · Uzbek

|           |                                                    |
|-----------|----------------------------------------------------|
| Model     | Gemma 4 31B + LoRA adapter                         |
| Hardware  | 1 × NVIDIA H200                                    |
| Settings  | FP8, speculative decoding,<br>prefix caching       |
| Tested on | TTFT, tokens/s, \$ per 1M tokens,<br>Uzbek quality |
| History   | Version 3; any version restorable                  |

## WITHOUT COREAI

Hire AI engineers → Pick a model and GPU → Tune settings →  
Benchmark → Deploy → Monitor → Repeat

## WITH COREAI

Pick a passport → Production endpoint

Every deployment we tune becomes a passport the next customer can start from.

## ADVANTAGES

# Why companies choose CoreAI.



## Per-token API

Self-serve, plus managed fine-tuning.  
Local clouds offer neither.



## 18–30% cheaper

Than local GPU providers.



## Under 5 ms

In-country latency, for real-time  
workloads like voice agents.



## Billed in som

Invoices under Uzbek law. No USD cards,  
no foreign contracts.



## Easier to use

It's why Acclaim AI left a state-run GPU  
for CoreAI.



## GPUs in Uzbekistan

Beeline's tender made it a requirement.

## THE CATEGORY

# The inference layer is worth billions.

**Fireworks AI** Series D · July 2026

Over \$1B in annualized revenue; \$1.5B raised in the round

**Baseten** June 2026

\$1.5B raised in two tranches

**Together AI** Series C · July 2026

\$800M raised in the round

**CoreAI brings this layer to Central Asia.**

## ACQUIRED

**CentML → NVIDIA**

Inference optimization software. Deal reported above \$400M, 2025.

*“More than 95% of those tokens come from models specialized on customers' proprietary data.”*

Fireworks AI, July 2026

## TRACTION

# Already running production AI.



Dedicated GPUs at BeeLab, Beeline  
Uzbekistan's subsidiary. Won by tender.

## Acclaim

Voice agents in production: speech-to-text, text-to-speech and LLM inference.



Video subtitles. Fine-tuning and  
distillation of their models next.

**3.5×** Acclaim AI moved its voice agents off a state-run GPU onto CoreAI, then grew 3.5× in six months.

**\$58,000**

revenue to date

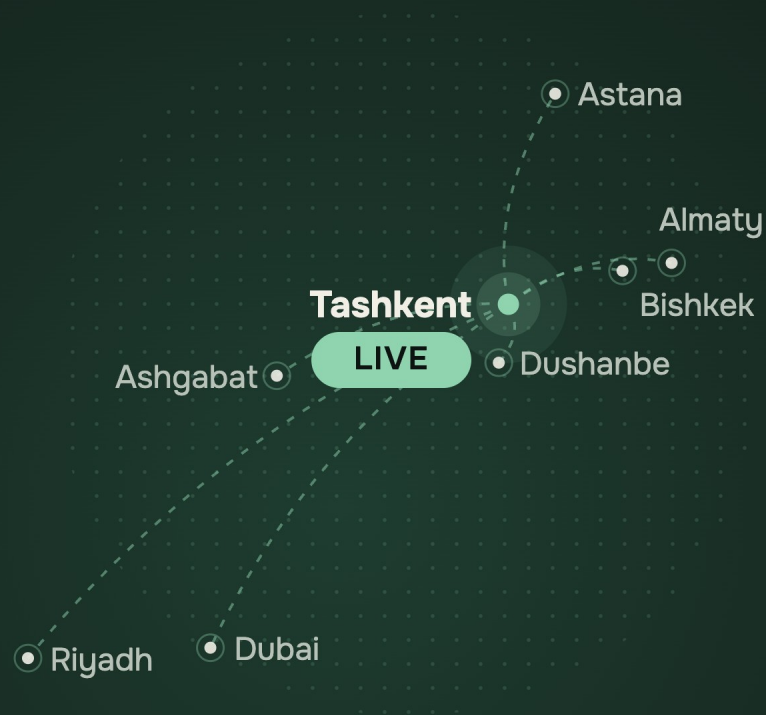
**\$0**

spent on acquisition; every customer came to us

Next: a private-cloud deployment for a large financial institution, a confidential-computing pilot, and VoiceLab testing managed inference.

## THE REGION

# Built in Tashkent, for Central Asia and MENA.



Live in Tashkent today. Provider Edition is built for partner datacenters across the region.

## \$255B

AI inference market by 2030 ·  
MarketsandMarkets

## ~\$1.8B

AI cloud and model platforms, Central  
Asia and MENA, 2028 · IDC

### THE REGION IS BUILDING GPU CAPACITY

**Kazakhstan.** Alem.Cloud national supercomputer (512× H200), Kazakhtelecom GPU cloud, a \$10B Data Center Valley under construction.

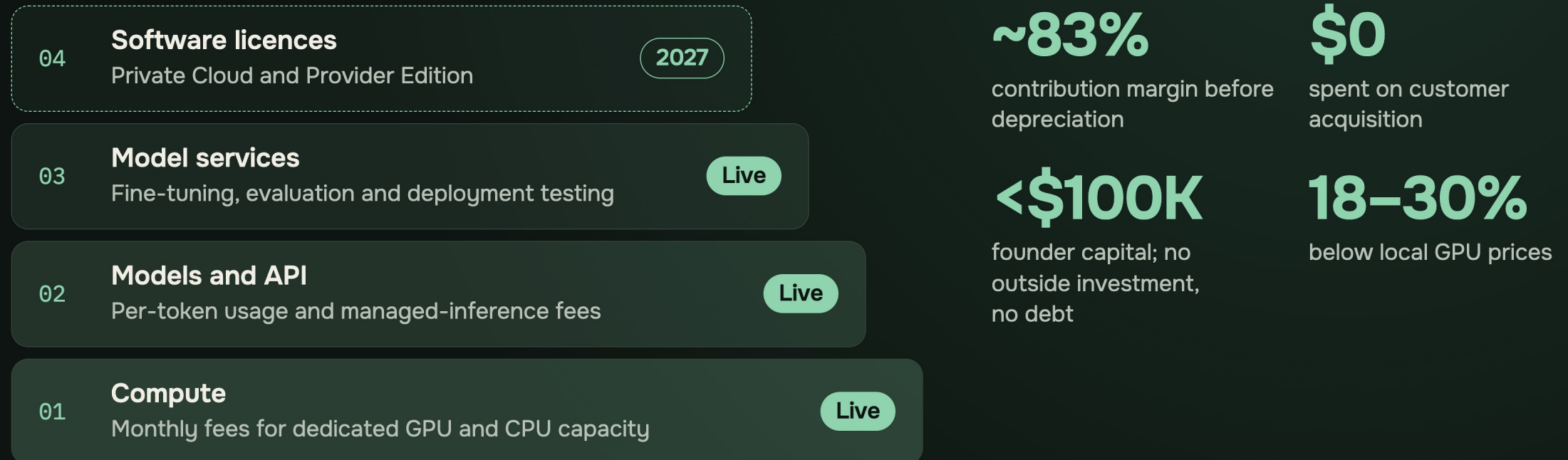
**Uzbekistan.** DataVolt's 12 MW AI datacenter in Tashkent, national NVIDIA clusters, an Uzbek LLM programme.

**Gulf.** Sovereign-AI programmes and datacenter operators looking for platform partners.

**Every new datacenter in the region is a potential CoreAI partner.**

## BUSINESS MODEL

# We earn at every layer and have never raised outside capital.



## TEAM

## Two founders, friends since high school.



### Sanjarbek Rakhmonov

Co-founder and CEO

Leads AI engineering: model evaluation, fine-tuning and inference. Runs enterprise sales, including the Beeline tender.

**KAIST** · full scholarship

**FAU Erlangen-Nürnberg** · AI, graduate

**AI engineer** · Seoul startup, 2 years

**AI engineer** · Alstin Capital, Munich



### Sherzod Safarov

Co-founder and CTO

Wrote the CoreAI control plane and built the Tashkent infrastructure. Over seven years in backend and platform engineering.

**Salesforce** · software

**Uztelecom** · cloud ops

**ProData** · datacenter

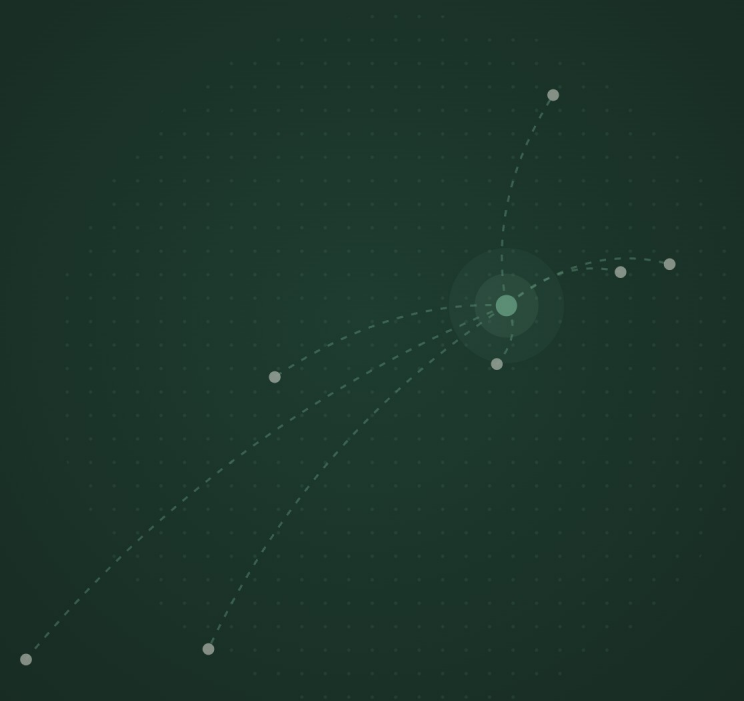
First hires: two platform engineers in the next six months, paid from current revenue.

# Central Asia will run its own AI.

## CoreAI is the layer it runs on.

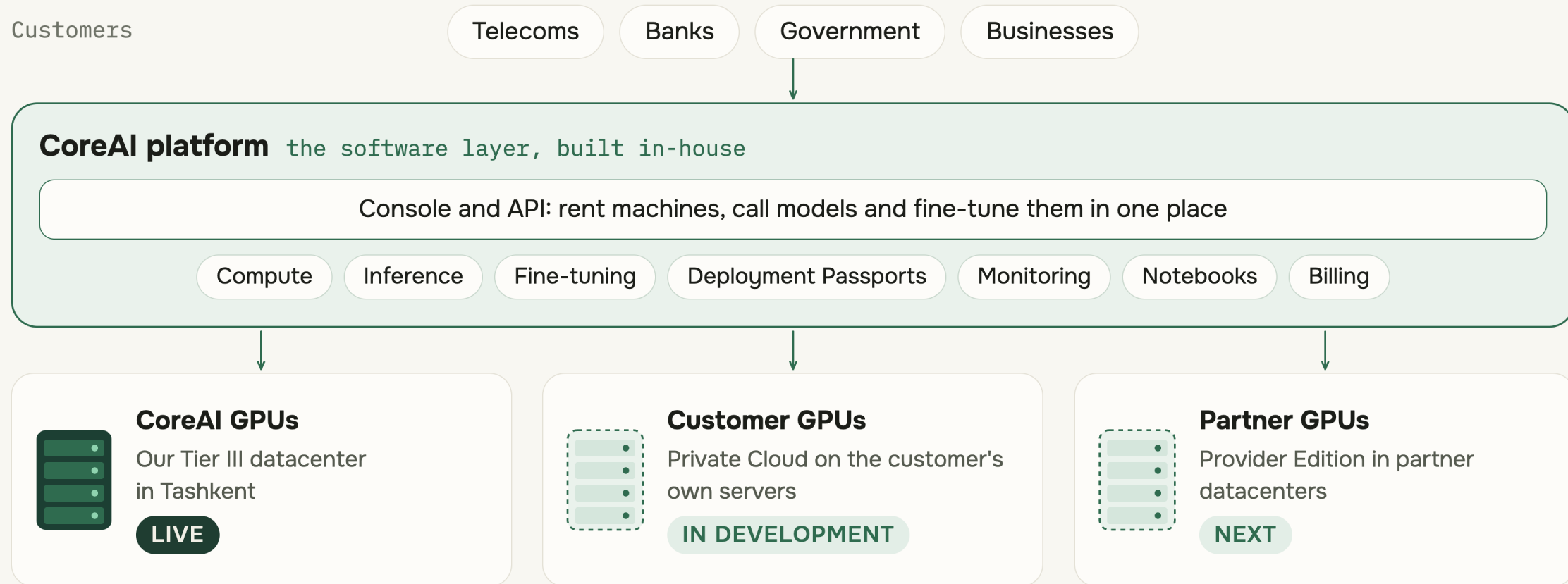
---

[coreai.uz](https://coreai.uz)   [info@coreai.uz](mailto:info@coreai.uz)   [linkedin.com/in/rakhmonov-sanjar](https://linkedin.com/in/rakhmonov-sanjar)



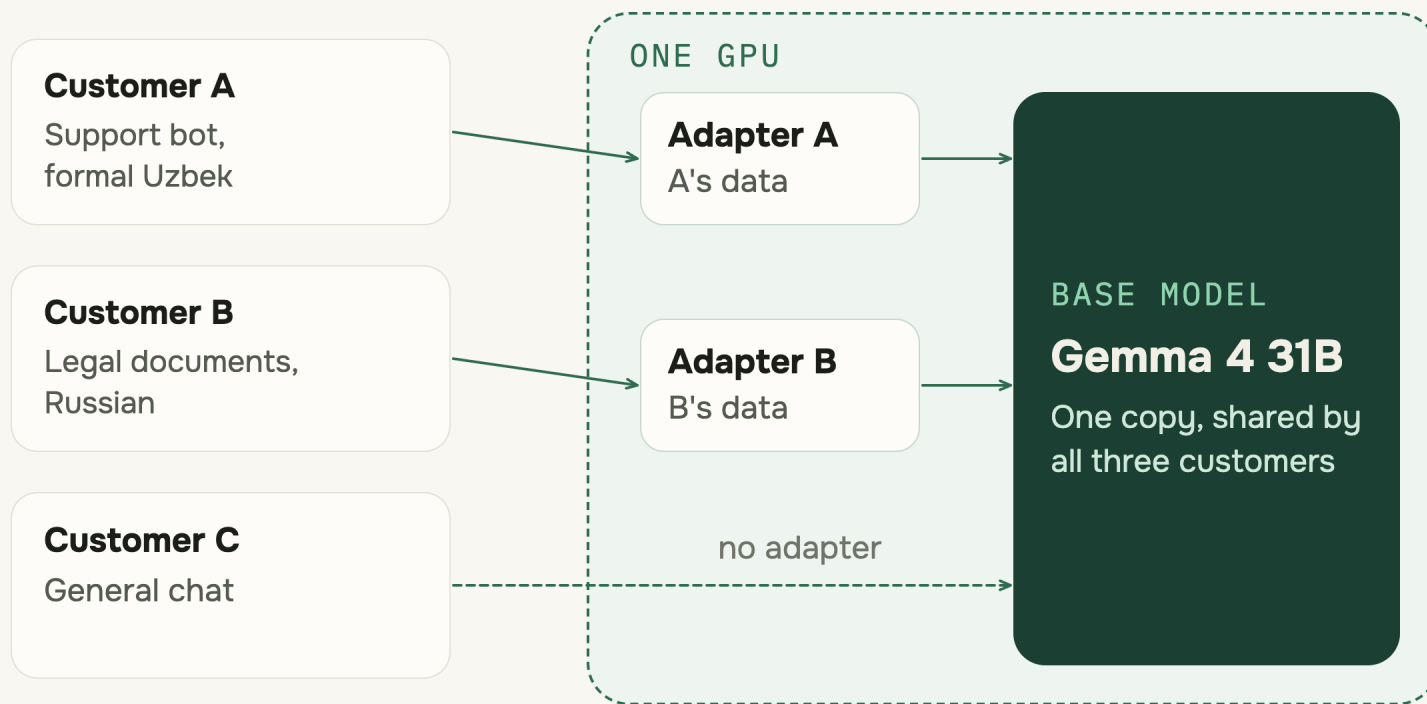
## APPENDIX · HOW IT WORKS

# The same software runs on anyone's GPUs.



## APPENDIX · SCALING PAST OUR GPUS

# One GPU serves many customers.



## We add customers faster than GPUs

**Shared models.** A GPU rented to one customer idles between requests; a shared one stays busy.

**Adapters.** Each customer gets its own fine-tuned model without a GPU of its own.

**Custom kernels.** More answers out of every GPU we run.

**Any hardware.** The same API runs on customer and partner GPUs, and we're already in talks with companies that have idle capacity.

## APPENDIX · COMPETITION

# CoreAI is both **AI-native** and **local**.

