HSE University · Faculty of Computer Science

AI Leaders

Business Lab for Executives

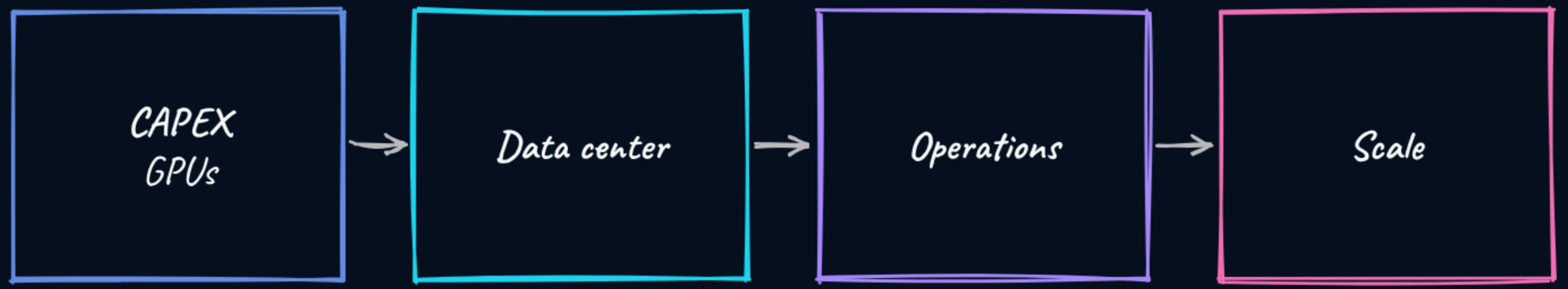
Evolution of AI Project Approaches



From resources to ready capabilities

On-Premise Hardware

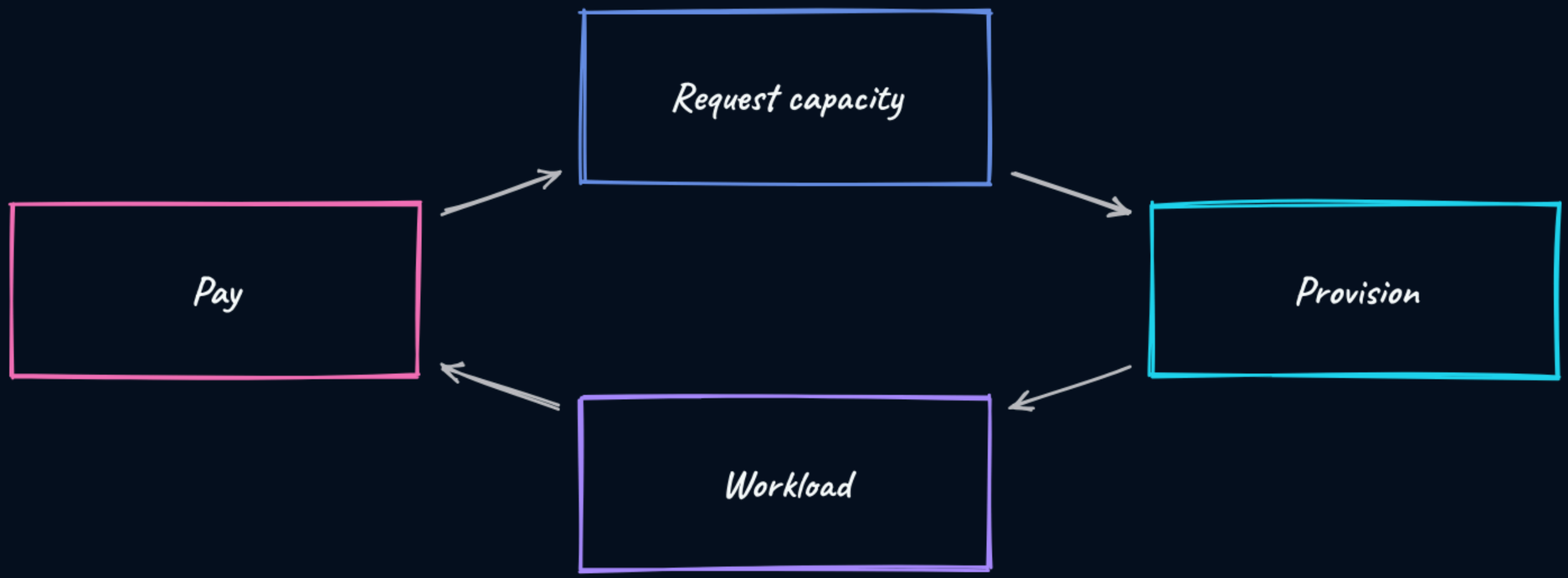
Stage 1 2008-2012



Lead time includes procurement and site preparation

Cloud Compute

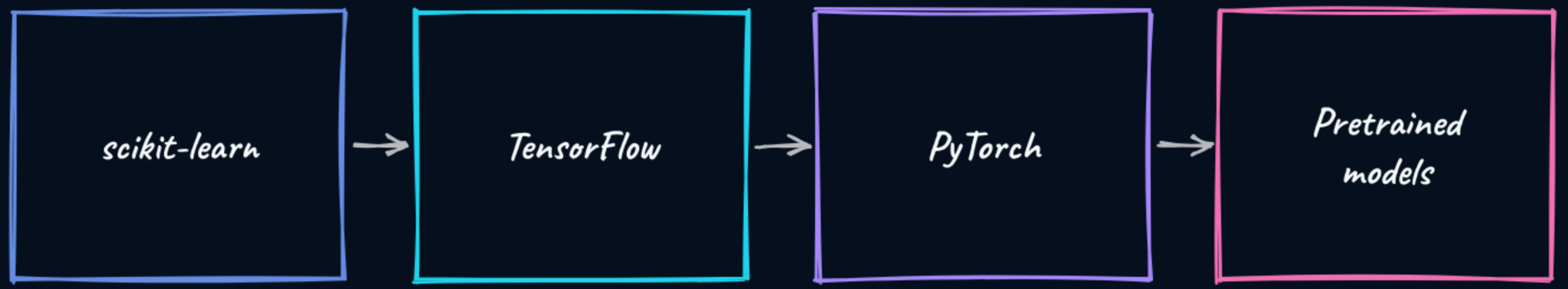
Stage 2 2013-2016



On-demand capacity; steady load changes the economics

ML Tooling Evolves

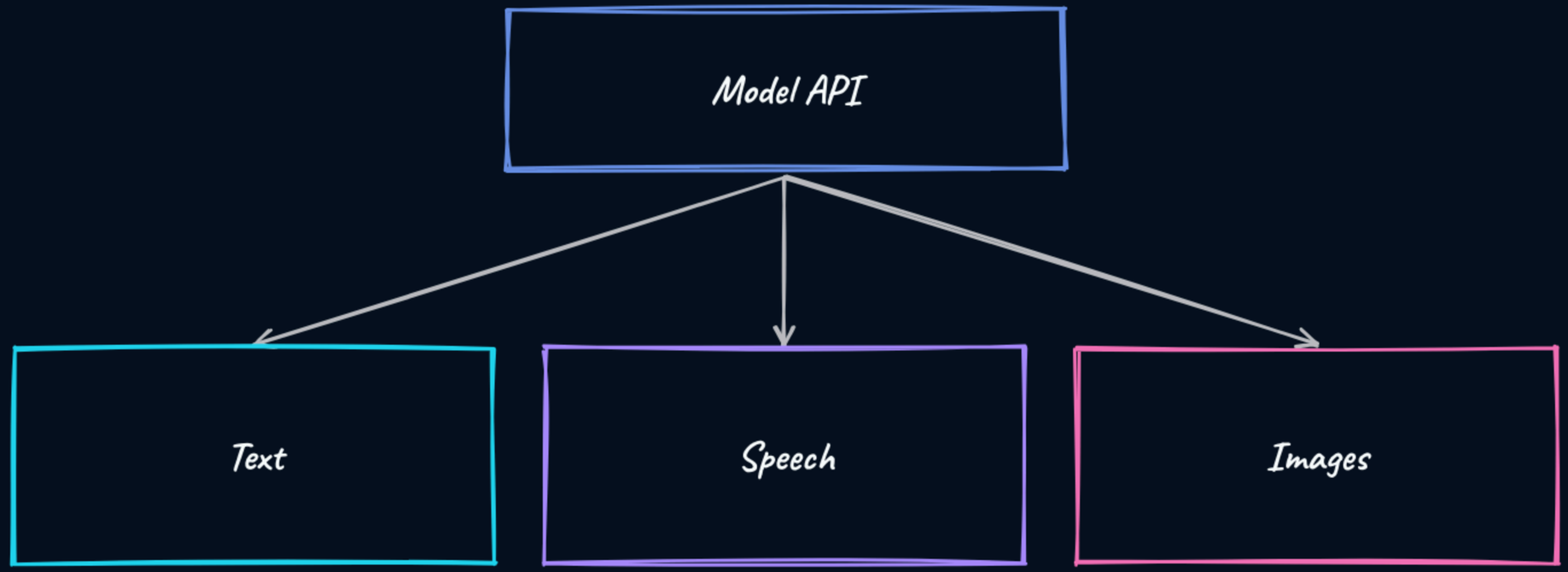
Stage 3 2016-2019



Tools lower entry costs; teams and pipelines remain

Foundation Models

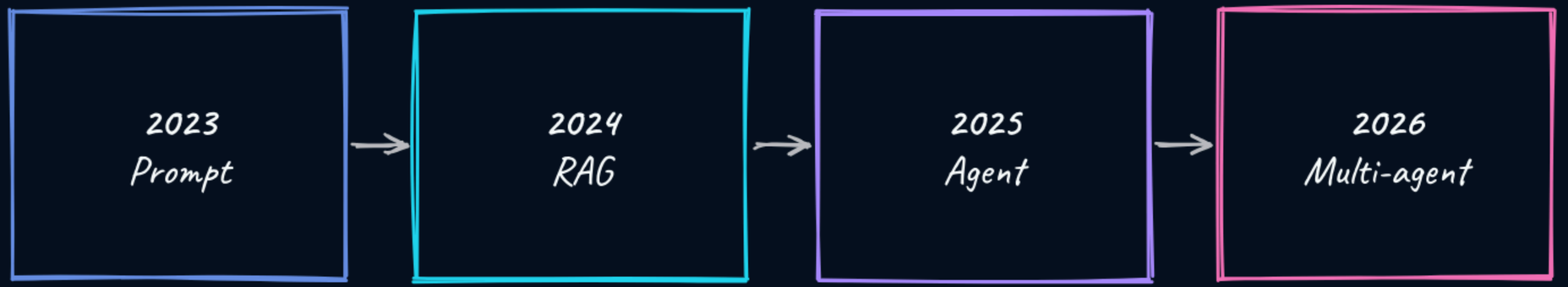
Stage 4 2020-2022



The product connects context, models and integrations

GenAI Applications

Stage 5 2023-now

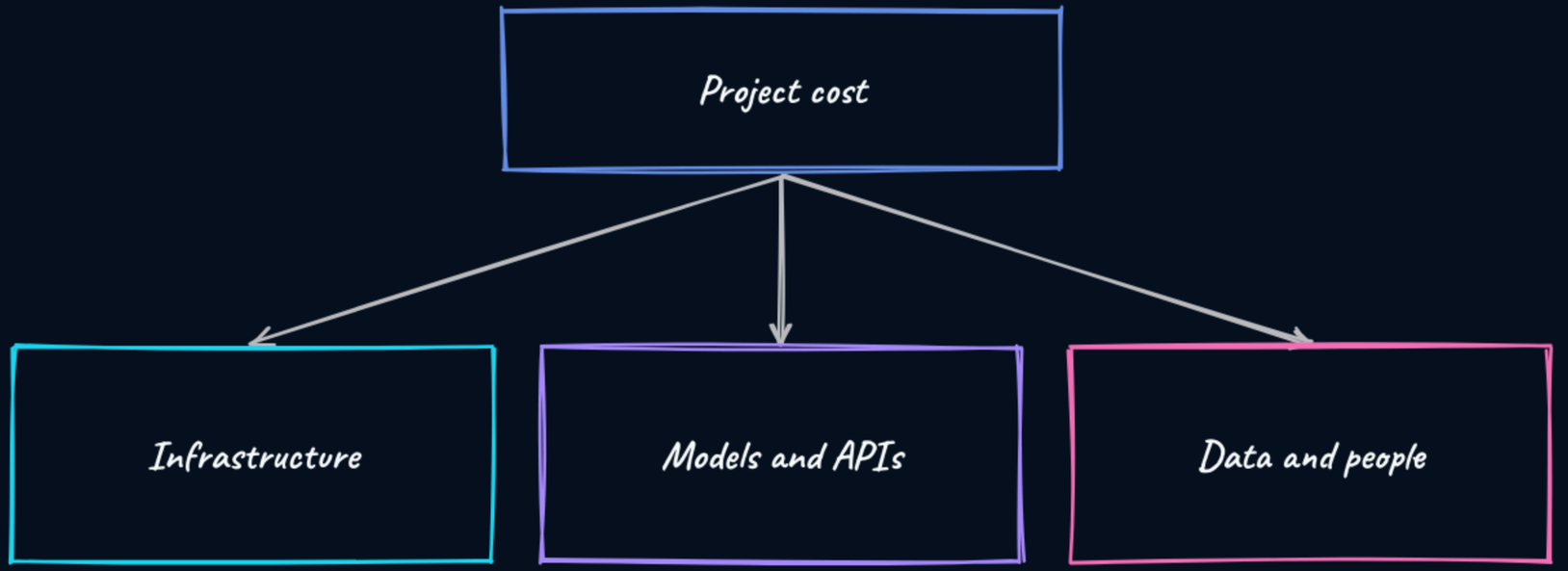


Roles Shrink to Tiny Teams

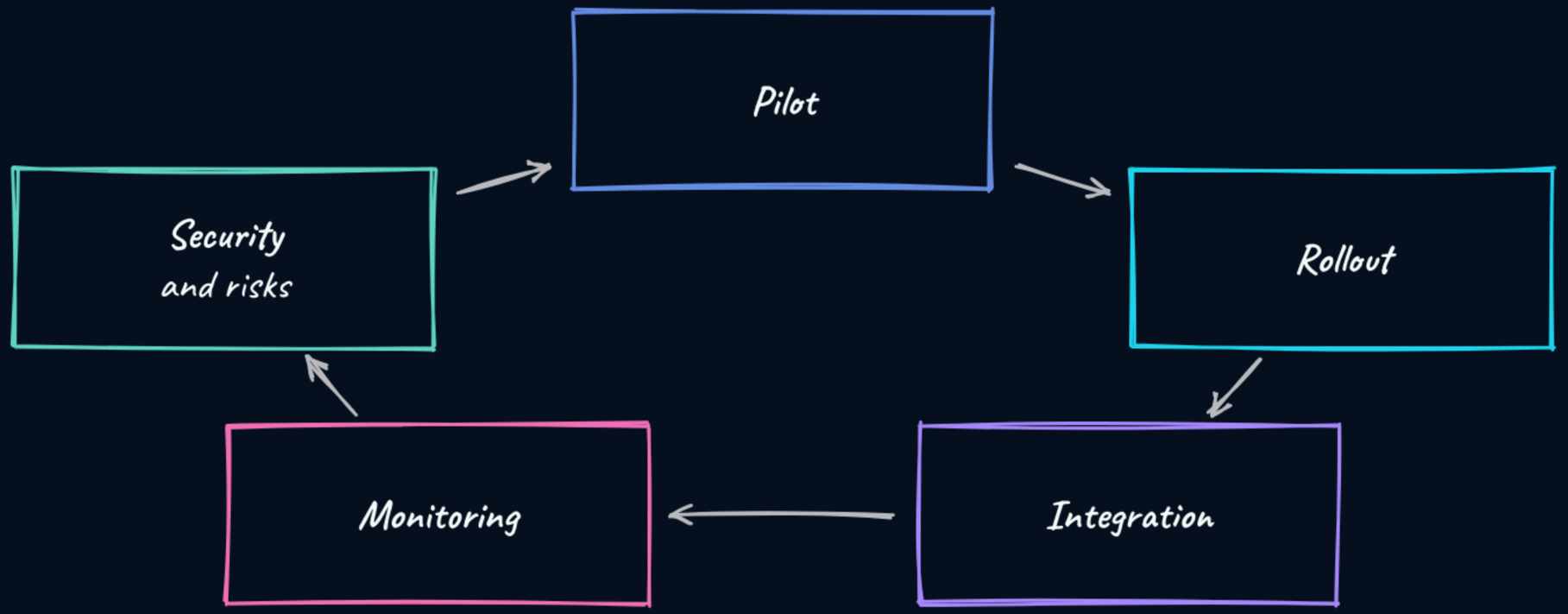
Before	Pilot today
7–10 roles	2–3 people
Data · ML	AI + MLOps
Infrastructure	Domain + product
Integration	

Combining roles does not remove accountability

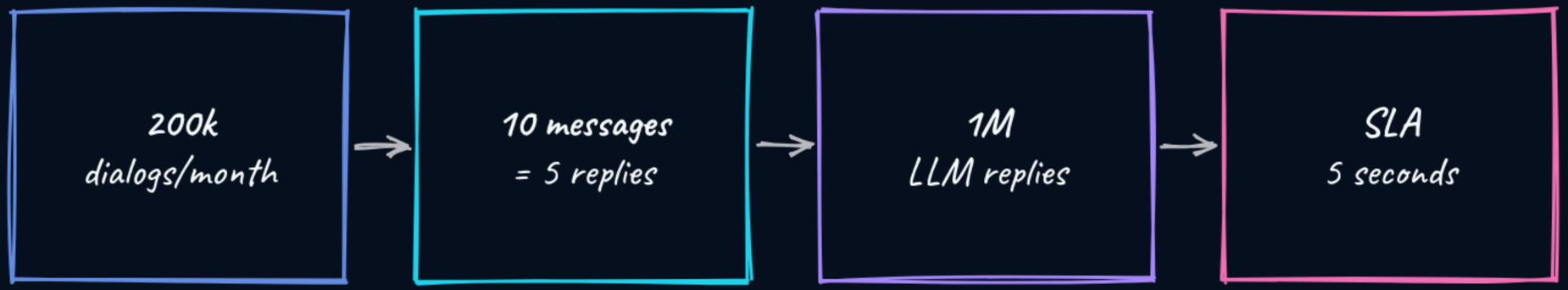
Cost Estimation and Hidden Line Items



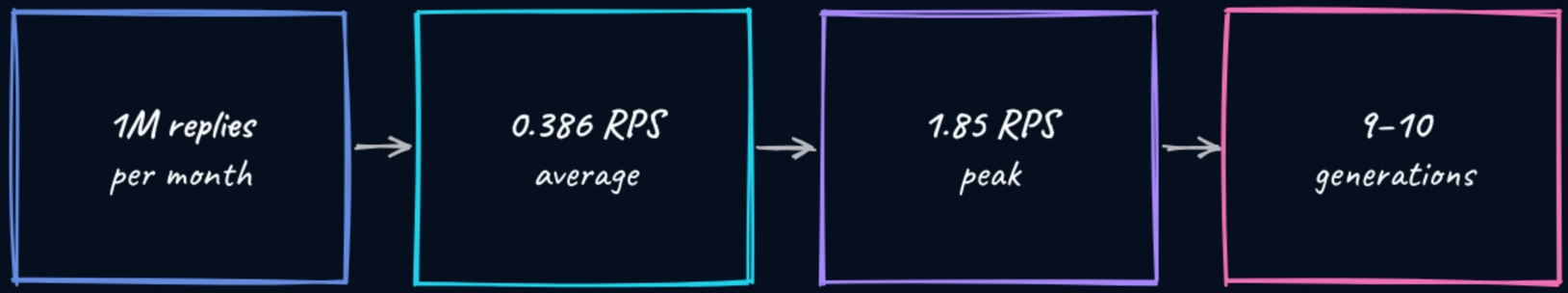
Technical Side



Support Bot TCO



From Parameters to Load



Concurrency ≈ peak RPS × 5 seconds

OpenAI gpt-5.2: Cost

Input	Output	Total
1.4B tokens $1.75/M $2,450	180M tokens $14/M $2,520	$4,970/month $0.0249/dialog

gpt-5.2: token economics

OpenAI API: CAPEX, OPEX, team, speed

Finance	Team	Time
CAPEX ≈ 0 LLM ≈ $5k/month Platform $0.5–5k	Launch: 5–7 FTE Run: 1.5–3 FTE	MVP: 4–6 weeks Prod: 8–12 weeks

Qwen Self-Host: 5 sec SLA

Qwen-14B · INT4

~90 tokens/s
180 tokens: 2.0s
~2 requests/GPU
6–8 GPUs

Qwen-32B · INT4

~45 tokens/s
180 tokens: 4.0s
~1 request/GPU
≈12 GPUs

Peak ≈10 replies, 5 s SLA; with batching, fewer GPUs are needed

Self-Host: CAPEX, OPEX, team

CAPEX	Team	Time
14B: $120–220k 32B: $200–350k+	Launch: 8–12 FTE Run: 3–5 FTE	12–20+ weeks Plus delivery

Depreciation of $180k ≈ $5k/month; OPEX: colocation, maintenance, GPU ops

Qwen on Rented GPUs

Case · Step 6 GPU Rental

Finance

CAPEX ≈ 0
OPEX: GPUs, network, storage
Costlier than own hardware at high load

Implementation

Faster than buying servers
Needs MLOps/SRE
Risk: GPU price and availability

Lower CAPEX, higher OPEX

GPU Rental OPEX



Rate: cloud.ru price list, February 2026; network and reserve extra

Three Deployment Strategies

	API	Rented GPUs	Own GPUs
CAPEX	Low	Low	High
OPEX	~$5k LLM/month	1.17–2.33M RUB/month	Infrastructure + team
Launch	Fast	Medium/fast	Slower
Control	Lower	High	Maximum

For this workload: API-first

Case recommendation



Strict on-prem requirements change the choice

Materials

Sources References

API, GPU, self-host

- OpenAI API docs/pricing
- ↗cloud.ru GPU price list, February 2026
- Hugging Face, Qwen: models and speed
- FinOps unit economics

Key Conclusions

