ФКН

ИИ-лидеры: бизнес-лаборатория для руководителей

СТАРТ: 26 ФЕВРАЛЯ 2026

Образовательная программа

охватывает **все основные этапы внедрения искусственного интеллекта** в компании от формирования стратегии и выбора приоритетных кейсов до запуска пилотов и масштабирования успешных решений

ДАТЫ КОНСУЛЬТАЦИЙ
С АВТОРАМИ ПРОГРАММЫ:
13 ФЕВРАЛЯ, 10:00
17 ФЕВРАЛЯ: 17:00

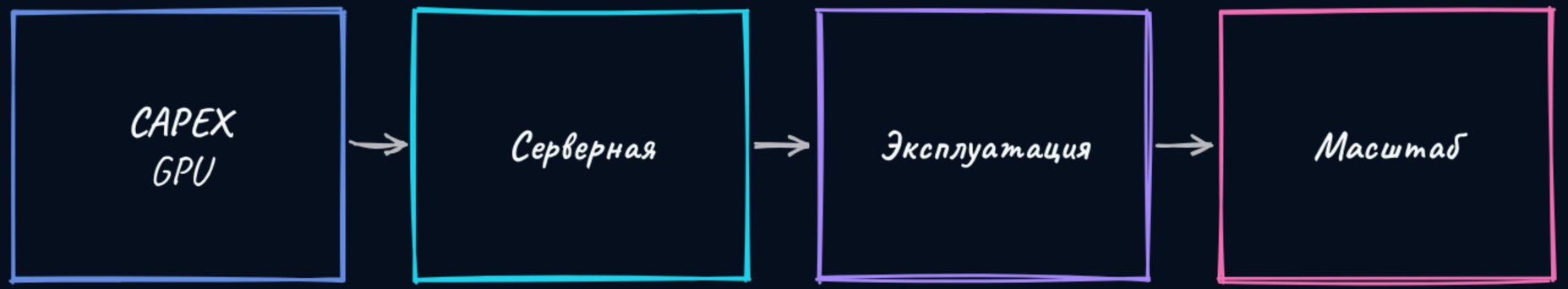
Эволюция подходов к ИИ-проекту



От ресурсов — к готовым возможностям

Собственное железо

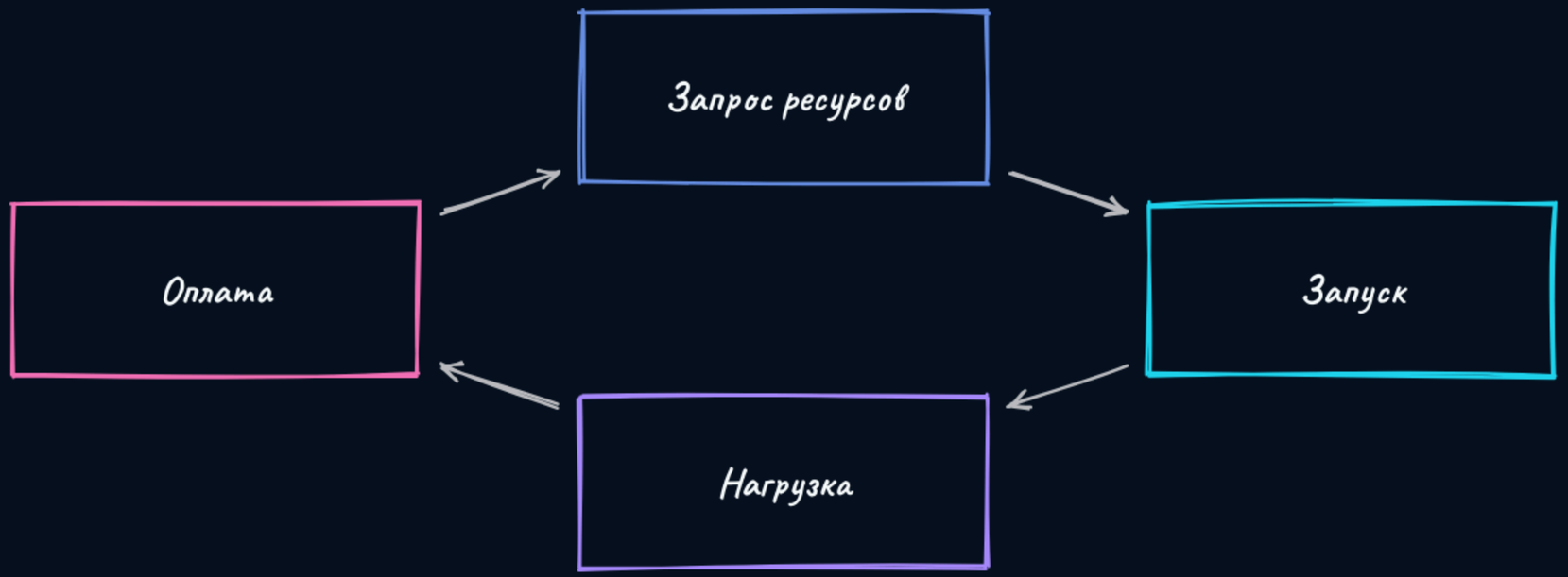
Этап 1 2008–2012



Срок запуска включает закупку и подготовку площадки

Облака:
compute on demand

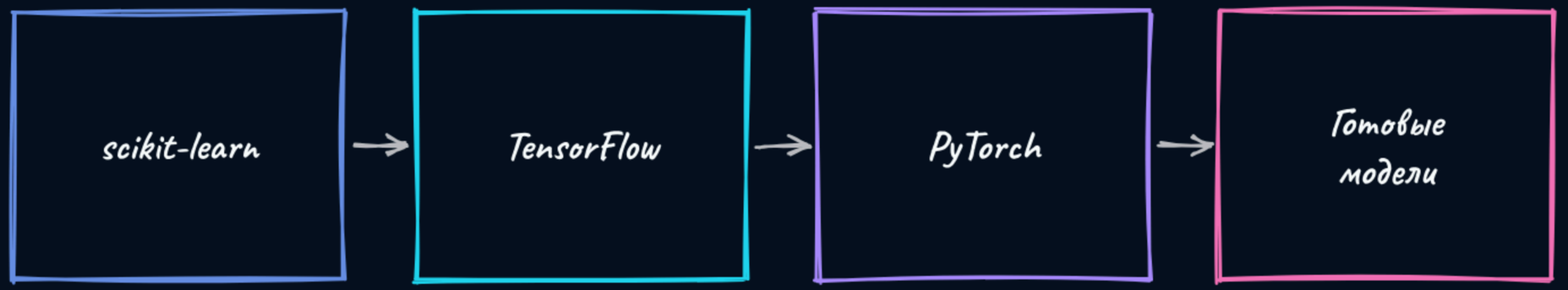
Этап 2 2013–2016



Ресурсы по требованию; постоянная нагрузка меняет экономику

ML-инструменты взрослеют

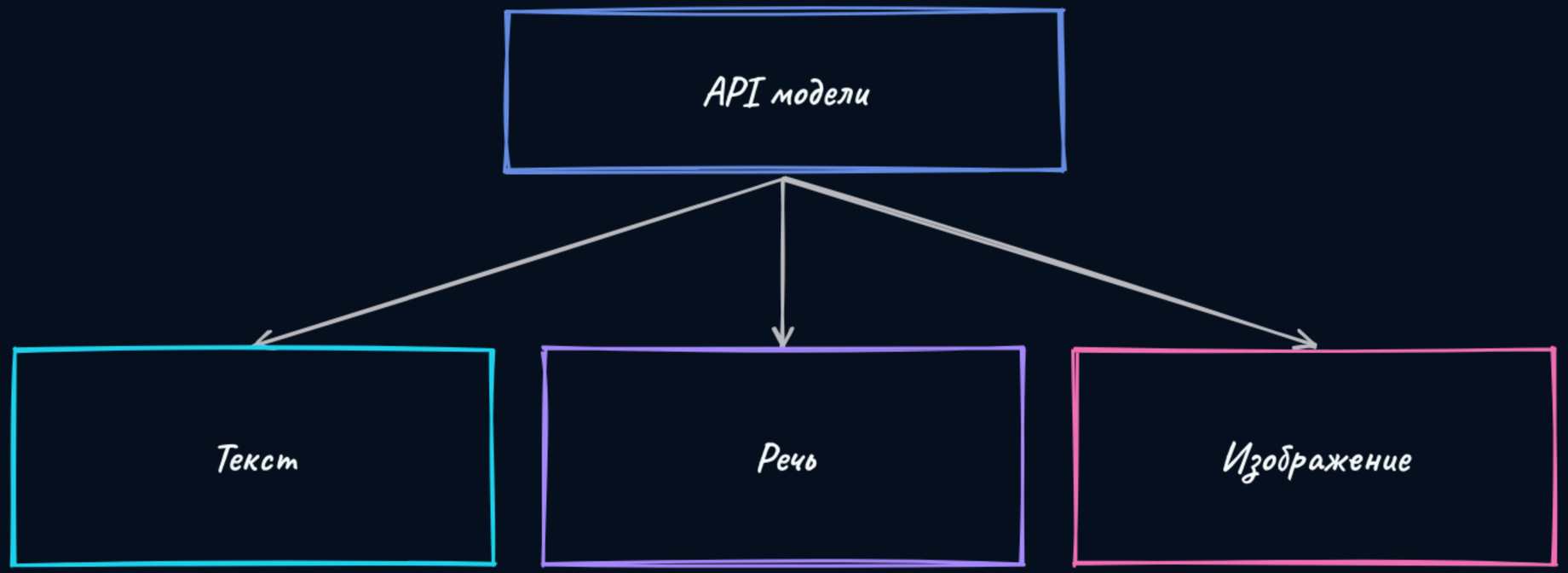
Этап 3 2016–2019



Инструменты снижают барьер; команда и пайплайн остаются

Foundation models

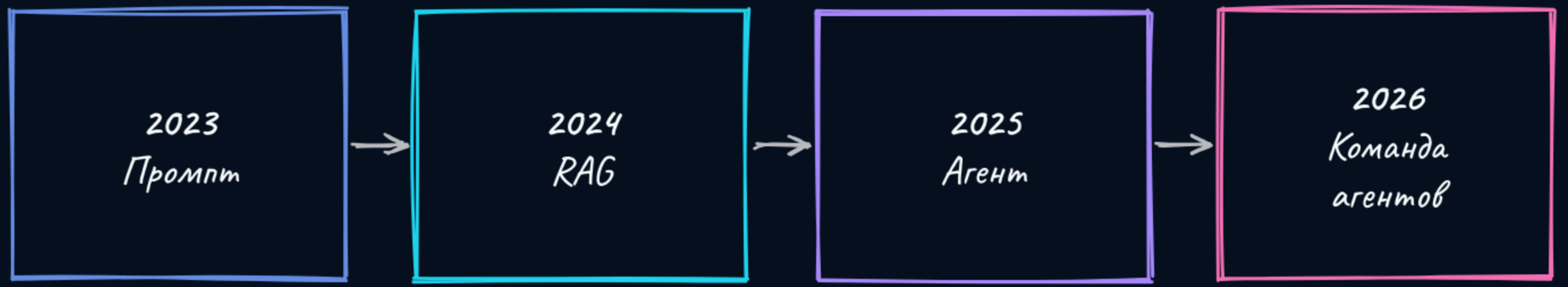
Этап 4 2020–2022



Продукт соединяет контекст, модель и интеграции

GenAI-приложения

Этап 5 2023-н.в.

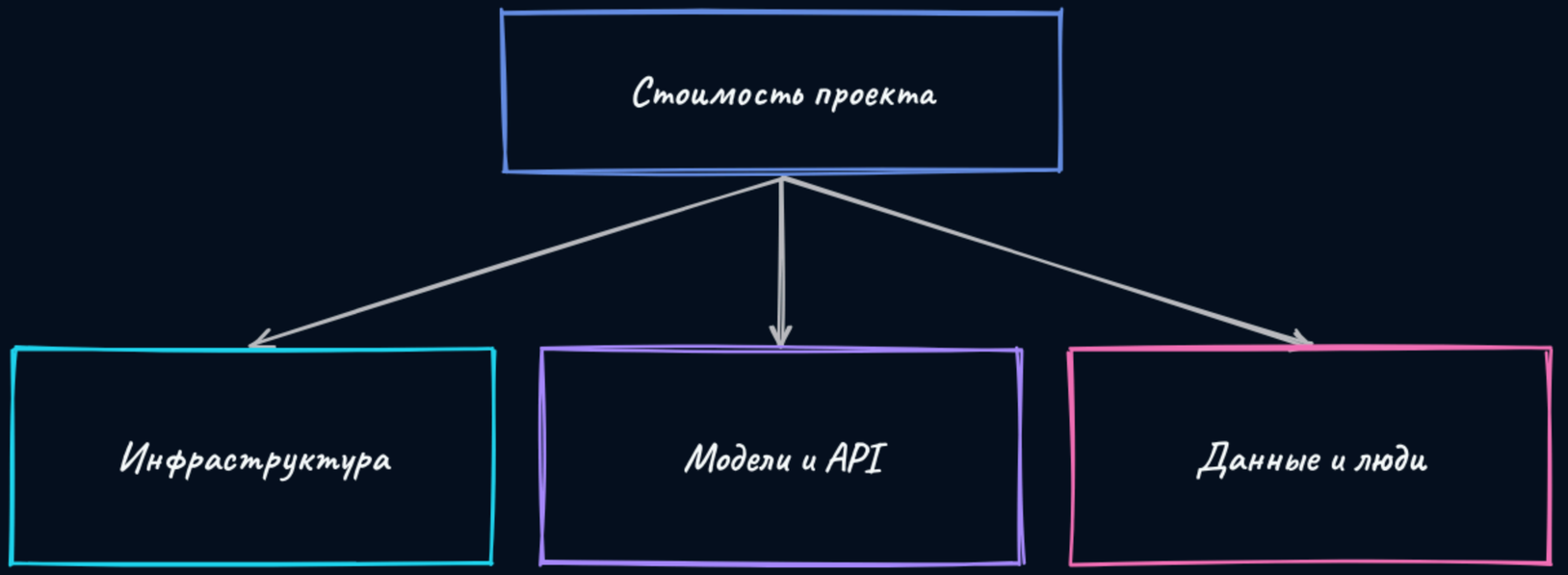


Роли сжимаются до tiny teams

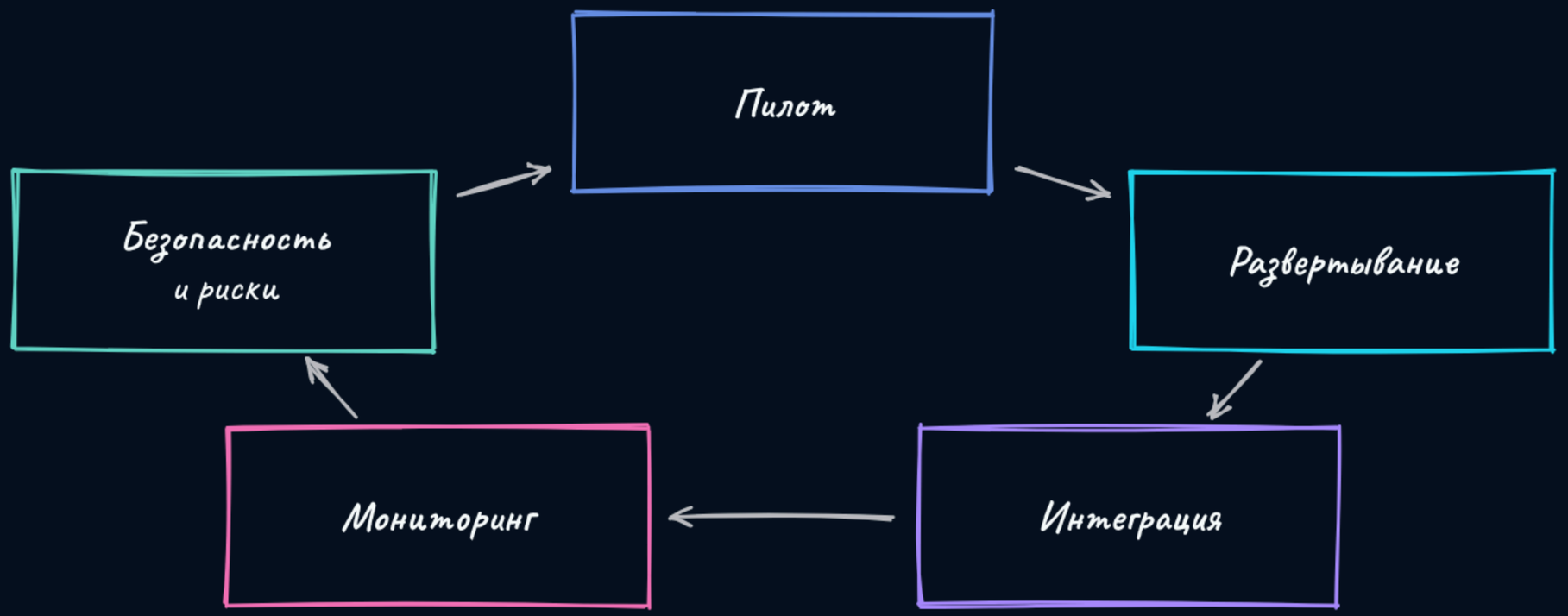
Раньше	Пилот сегодня
7–10 ролей	2–3 человека
Данные · ML	AI + MLOps
Инфраструктура	Домен + продукт
Интеграция	

Совмещение ролей не отменяет ответственность

Оценка затрат и скрытые статьи



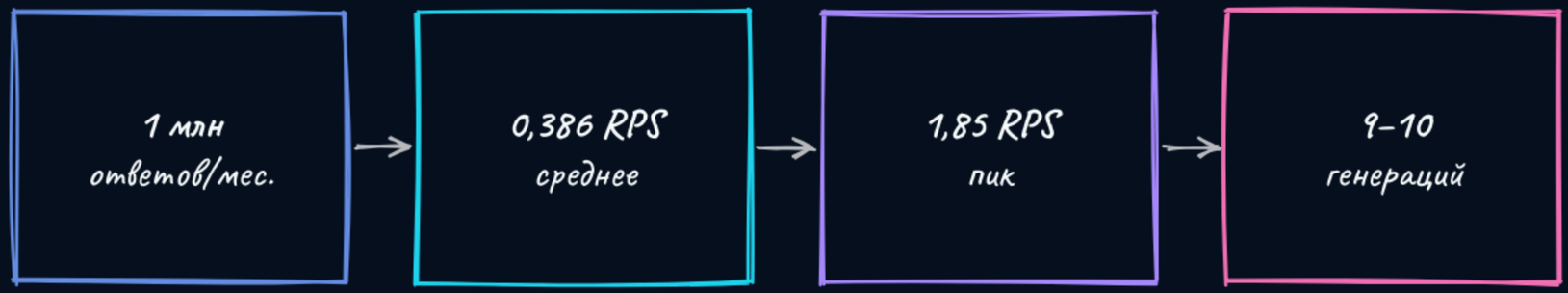
Техническая сторона



TCO support-бота



Из параметров в нагрузку



Параллельность ≈ пиковый RPS × 5 секунд

OpenAI gpt-5.2: стоимость

Вход	Выход	Всего
1,4 млрд токенов $1,75/млн $2 450	180 млн токенов $14/млн $2 520	$4 970/мес. $0,0249/диалог

gpt-5.2: экономика токенов

OpenAI API: CAPEX/OPEX/ Команда/Скорость

Финансы	Команда	Срок
CAPEX ≈ 0 LLM ≈ $5 тыс./мес. Платформа $0,5–5 тыс.	Запуск: 5–7 FTE Работа: 1,5–3 FTE	MVP: 4–6 недель Прод: 8–12 недель

Qwen self-host: SLA 5 сек

Qwen-14B · INT4

~90 ток/с
180 ток: 2,0 с
~2 запроса/GPU
6–8 GPU

Qwen-32B · INT4

~45 ток/с
180 ток: 4,0 с
~1 запрос/GPU
≈12 GPU

Пик ≈10 ответов, SLA 5 с; с батчингом GPU понадобится меньше

Self-host: CAPEX, OPEX, команда

CAPEX	Команда	Срок
14B: $120–220 тыс.	Запуск: 8–12 FTE	12–20+ недель
32B: $200–350+ тыс.	Работа: 3–5 FTE	Плюс поставка

Амортизация $180 тыс. ≈ $5 тыс./мес.; OPEX: колокация и эксплуатация GPU

Qwen на арендованных GPU

Кейс · Шаг 6 GPU аренда

Финансы

CAPEX ≈ 0

OPEX: GPU, сеть, хранение

Дороже своего железа при высокой загрузке

Реализация

Быстрее закупки серверов

Нужны MLOps/SRE

Риск: цена и доступность GPU

CAPEX ниже, OPEX выше

OPEX аренды GPU



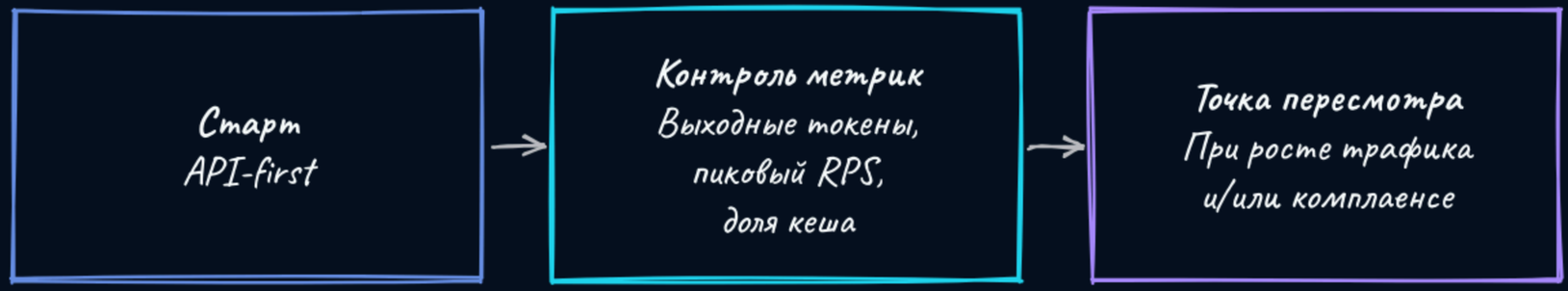
Ставка — по прайс-листу cloud.ru, февраль 2026; сеть и резерв сверху

Три стратегии развертывания

	API	Аренда GPU	Свои GPU
CAPEX	Низкий	Низкий	Высокий
OPEX	~$5 тыс. LLM/мес.	1,17–2,33 млн ₽/мес.	Инфраструктура + команда
Запуск	Быстро	Средне/быстро	Медленнее
Контроль	Ниже	Высокий	Максимум

При текущей нагрузке — API-first

Управленческий вывод по кейсу



Обязательное размещение на своих серверах меняет выбор

Материалы

Источники Ссылки

API, GPU, свой хостинг

- OpenAI API: документация и цены
- ↗cloud.ru: прайс-лист GPU, февраль 2026
- Hugging Face, Qwen: модели и скорость
- FinOps: юнит-экономика

Главные выводы

