турбо Kodacode

Готовая облачная платформа –

как быстро запускать ИИ-модели в прод?

О чем расскажем сегодня

Вызовы запуска ИИ-моделей

Что такое Inference Platform

Возможности платформы

Что «под капотом» ?

Кейсы

Опыт КОДА

турбо

Kodacode

вебинар
10 июня

Денис Боженко
Ведущий менеджер продуктов
Группа сервисов
искусственного интеллекта
ТУРБО

Максим Глотов
Руководитель группы
архитектуры и проектирования
облачных решений
ТУРБО

Дмитрий Змитрович
Основатель
КОДА

13 лет эволюции

2012

Ростелеком

2013

DataLine

SafeData

2023

РТК-ЦОД

Публичное облако

С 2025

турбо

турбо_облако

5 федеральных округов

20+
площадок

13+
лет в облачном провайдинге

500 000+
vCPU

>40 ПБ
storage

«Облако рядом с клиентом»

Облачные сервисы на базе георапределенной сети
дата-центров уровня Tier III с отказоустойчивой инфраструктурой

Все востребованные на рынке виды облаков

- Публичные
- Частные
- Облако КИИ / 152 / ГИС
- Гибридные
- Зарубежные и российские вендоры

- «Турбо Облако» построено на базе географически распределенной сети собственных дата-центров уровня Tier III на территории РФ
- Все ключевые элементы платформы продублированы для отказоустойчивости
- Ресурсы каждого клиента изолированы на уровне виртуализации и на сетевом уровне с помощью VLAN
- Дополнительная сетевая защита обеспечена за счет Firewall и VPN-туннелей
- Поддерживаются все типы связности: оптика, стандартные VPN, ГОСТ VPN и другие варианты
- Безопасность инфраструктуры «Турбо Облака» подтверждена сертификатами ФСБ, ФСТЭК и аттестатом облачной платформы

Как запустить ИИ-модель?

→ 3 варианта

BareMetal с GPU

Сервис по аренде выделенного физического сервера, куда можно поставить GPU ускорители

→ **Клиент платит фикс. цену в месяц за конфигурацию сервера**

Виртуальная инфраструктура с GPU

Сервис по аренде виртуальных ресурсов с виртуализированными GPU ресурсами

→ **Клиент платит фикс. цену в месяц за объем потребляемых виртуальных ресурсов**

Inference

Сервис, который позволяет запускать ИИ-модели на облачных мощностях с GPU с автоскейлингом

→ **Клиент платит только за время работы модели в нашем облаке**

Как обычно разворачивают модель

ШАГ РАЗВЕРТЫВАНИЯ МОДЕЛИ	BareMetal	Виртуальная инфраструктура с GPU	Что-то новое
Загрузка модели	Клиент	Клиент	Клиент
Настройка мониторинга и логирования	Клиент	Клиент	Провайдер
Авто масштабирование модели	Клиент	Клиент	Совместно
Настройка контейнера, эндпоинта, параметров запуска	Клиент	Клиент	Совместно
Настройка балансировщика запросов	Клиент	Клиент	Провайдер
Установка и настройка Kubernetes	Клиент	Клиент	Провайдер
Настройка сети и безопасности	Клиент	Провайдер	Провайдер
Настройка GPU и среды выполнения	Клиент	Провайдер	Провайдер
Подготовка инфраструктуры / серверов	Провайдер	Провайдер	Провайдер

Вызовы для клиента

Запуск модели из коробки	**Клиенту нужно самому собирать inference-стек** Нужны DevOps/ML-инженеры для ручной настройки
Нативное масштабирование	**Клиенту сложно экономично держать переменную нагрузку** Нереализуемо на BareMetal / ВМ с GPU / собственной инфраструктуре
Масштабирование до 0	**Клиент переплачивает за простой** Включи/выключи ручками ресурсы
Мониторинг работы модели	**Клиент нужно сделать самому** Видно инфраструктуру, но не видно поведение модели как сервиса

А как мы видим процесс →

Клиент

1

Выбирает необходимую ему модель

Клиент

2

Указывает GPU-ресурсы для запуска модели

Клиент

3

Устанавливает триггеры и параметры для автоматического масштабирования

Провайдер

4

Разворачивает модель и берет на себя эксплуатацию, включая балансировку запросов

Клиент

5

Получает URL и пользуется моделью как готовым сервисом. Может встраивать в свои сервисы

турбо

Inference Platform

Inference Platform

PaaS-платформа для развёртывания и эксплуатации моделей искусственного интеллекта

Готовая среда для запуска ИИ-моделей
Инфраструктура для развертывания моделей
и контейнеров без дополнительной настройки

Стабильная работа под нагрузкой
Умная балансировка запросов с учётом
загрузки и производительности инстансов

Автоматическое масштабирование
Увеличение и уменьшение ресурсов в зависимости от количества запросов

Простая интеграция с проектами
Готовый сервис с доступом по URL для интеграции
в существующие бизнес-приложения без изменения архитектуры

Оптимизация затрат
Поминутная тарификация и автоматическое освобождение
ресурсов при отсутствии нагрузки (serverless-подход)

Подробнее по ссылке

Inference Platform—
3 модуля под разные задачи

01

Запуск ML-моделей без настройки среды

Загрузите модель из Hugging Face Hub, S3-хранилища или приватного реестра контейнеров

→

Платформа

автоматически определит нужную среду исполнения и запустит модель

02

Запуск собственных контейнеров с GPU

Если у вас есть собственная логика инференса, специфические зависимости или нестандартные фреймворки

→

Платформа

позволяет запустить любой контейнер с доступом к GPU. Образ загружается из публичного или приватного реестра

03

Хранение кэша моделей

Модели весом десятки и сотни гигабайт при первом запуске требуют времени на загрузки

→

Платформа

кеширует модели после первой загрузки для сокращения времени холодного старта при последующих запусках инстансов

Inference Platform — возможности сервиса

Автоматическое масштабирование
Увеличение и уменьшение ресурсов в зависимости от количества запросов

Scale to Zero
В случае отсутствия нагрузки мы можем скейлить модель в 0 с нулевой тарификацией

До 8 GPU на контейнер
Использование от 1 до 8 графических ускорителей NVIDIA H200 SXM для увеличения вычислительной мощности

Multi-node inference
Возможность запуска модели на нескольких серверах

Загрузка собственных моделей
Поддержка собственных моделей и контейнерных образов без ограничений платформы

Поддержка ML-фреймворков
Работа с vLLM, Ollama, Diffusers и SGLang с возможностью адаптации под конкретные задач

HTTP / HTTPS endpoint
Приватный или доступный через Интернет URL для каждого инстанса с авторизацией через API-ключ и поддержкой Open AI-совместимого API для удобной интеграции без изменений архитектуры

Аппаратное ядро платформы

Intel Xeon Platinum 8480+ x2

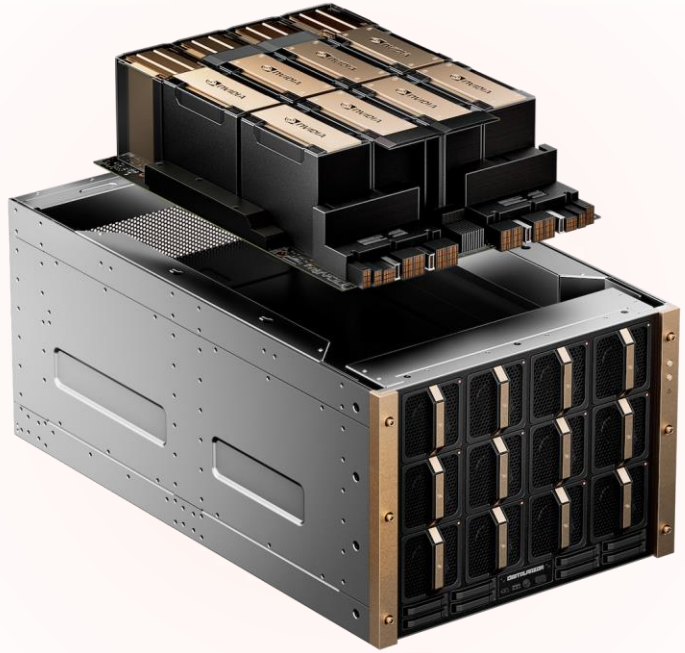
2Tb RAM

32 Tb NVMe SSD

400 GB/s Connectx7 x8

GPU H200 x8

- NVSwitch
- Non-blocking Infiniband Fabric
- GPU Direct RDMA



HGX Supermicro H200

Scale-in Scale-Out

Возможность расширения другими хостами

Multi GPU инференс

Multi Node инференс

Serving layer

Multi Node Inference

Scale to zero

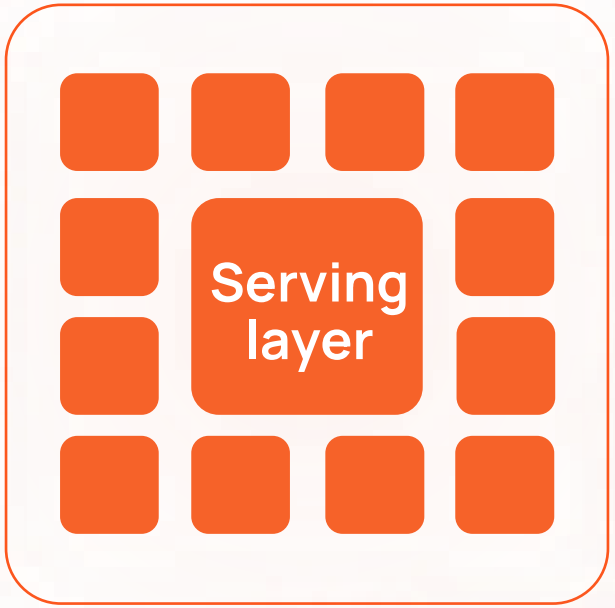
vLLM

Triton

SGLang

Diffusers

Уже доступны



Высокая скорость холодного старта при масштабировании

Поддержка кастомных runtime

До 8

экземпляров масштабирование

Prefill и Decode

на разных ресурсах

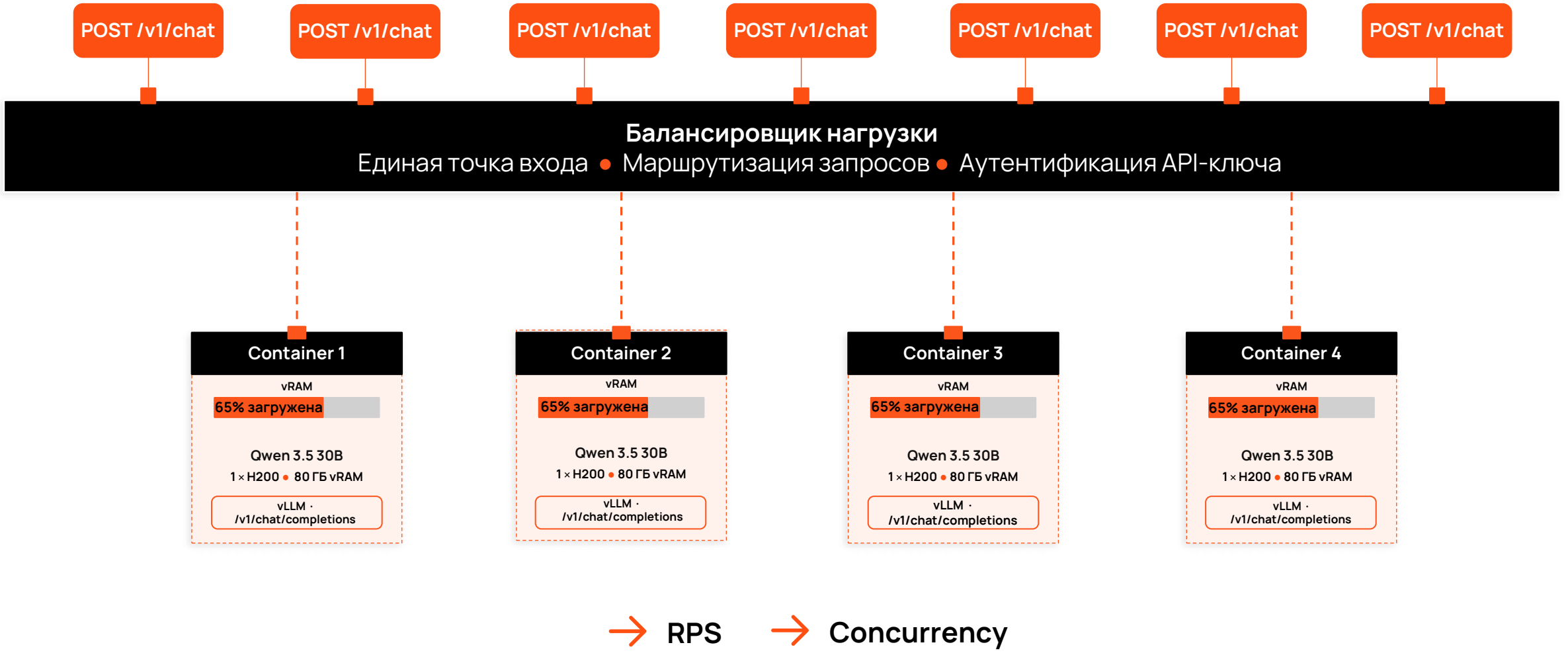
До 8x

GPU H200

на один инференс

Конвейер runtime

Балансировка запросов и автоскейлинг



→ RPS → Concurrency

ModelRun Operator и Docker Run Operator

Запуск клиентских моделей

vLLM · OpenAI API

SGLang · батчинг

Ollama · Ollama API

Diffusers · изображения

API

OpenAI-совместимый /v1/chat/completions или Ollama API
Клиент получает типизированный эндпоинт под задачу

ВОЗМОЖНОСТИ

- ✓ Автоскейлинг по RPS / Concurrency (HPA)
- ✓ Scale-to-zero (min реплик = 0)
- ✓ Cold start при первом запросе после паузы
- ✓ Model Cache: модель не скачивается заново

Запуск клиентских контейнеров

Container Registry

Образ со встроенной моделью

API

Любой кастомный HTTP-эндпоинт внутри образа
Платформа проксирует запросы без ограничений на формат

КОГДА ИСПОЛЬЗОВАТЬ

- ✓ Нестандартный рантайм или кастомный код
- ✓ Модель не из HuggingFace / S3
- ✓ Специфичная логика препроцессинга
- ✓ Возможно своя дообученная модель

- ✓ ModelRun Operator и Docker Run Operator — это собственные компоненты платформы
- ✓ Они оркестрируют целевое состояние inference-runtime
- ✓ Они управляют большим количеством платформенных сущностей и переходов состояний
- ✓ Они обеспечивают self-healing и повторяемый контроль жизненного цикла

Как уже используют Inference Platform

	Аудит	Внутренний ИТ	Kodacode
Задача	Работа с объемной закупочной документацией и финансовой отчетностью: нет эффекта от классических методов извлечения данных типа RAG, так как теряется много важного контекста = некорректная сводка по документации = невозможность принимать решения	Сделать поиск по внутренней документации (RAG) Ускорить разработку Сделать решение для вайбкодинга: например, подключить opencode к LLM	Необходим гибкий сервис по запуску моделей с автоматическим масштабированием в зависимости от пользовательской нагрузки
Решение	**Мультинодовый инференс от Турбо:** размещение большой модели с длинным контекстным окном, в рамках которого помещается вся необходимая документация без потери контекста Большая модель размещена на нескольких серверах	**Инференс от Турбо:** размещение моделей с масштабированием до 0 при отсутствии нагрузки	**Модели клиента в нашем сервисе:** Размещение с автоматическим скейлингом в зависимости от нагрузки
Модели	▪ **DeepSeek v3.2 (16GPU H200)**	▪ **Qwen3.5-122B-A10B-AWQ (1x H200)** ▪ **gpt-oss:20b (70Гб памяти H200)**	▪ **GLM 5.1 (16xH200)** ▪ **Qwen3.5 122B-A10B (8xH200)**
	Позитивные отзывы	**Позитивные отзывы**	**Позитивные отзывы**

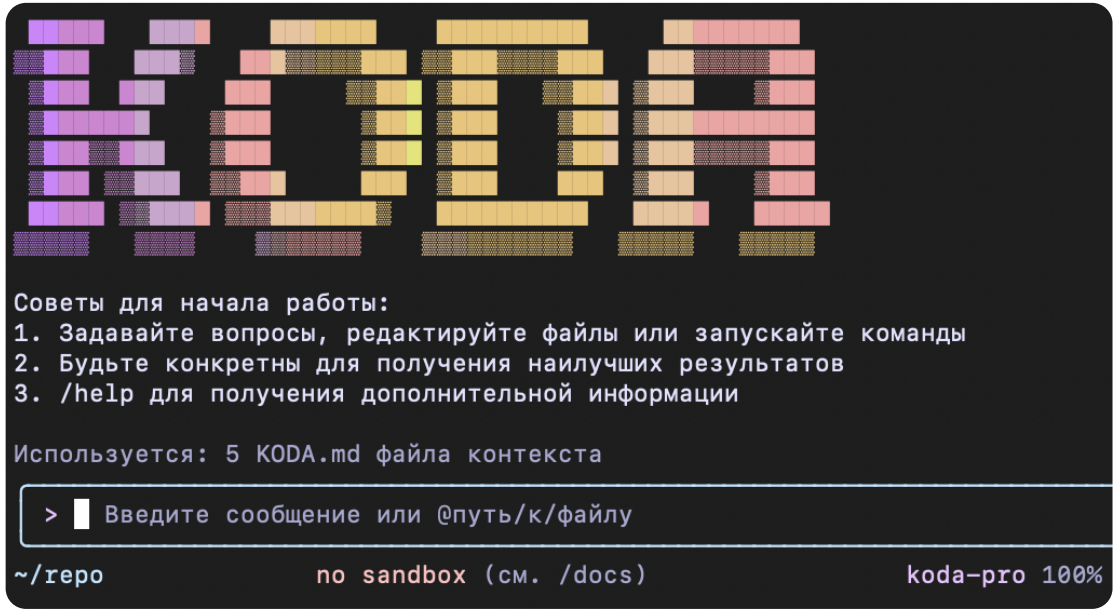
Kodacode

kodacode.ru

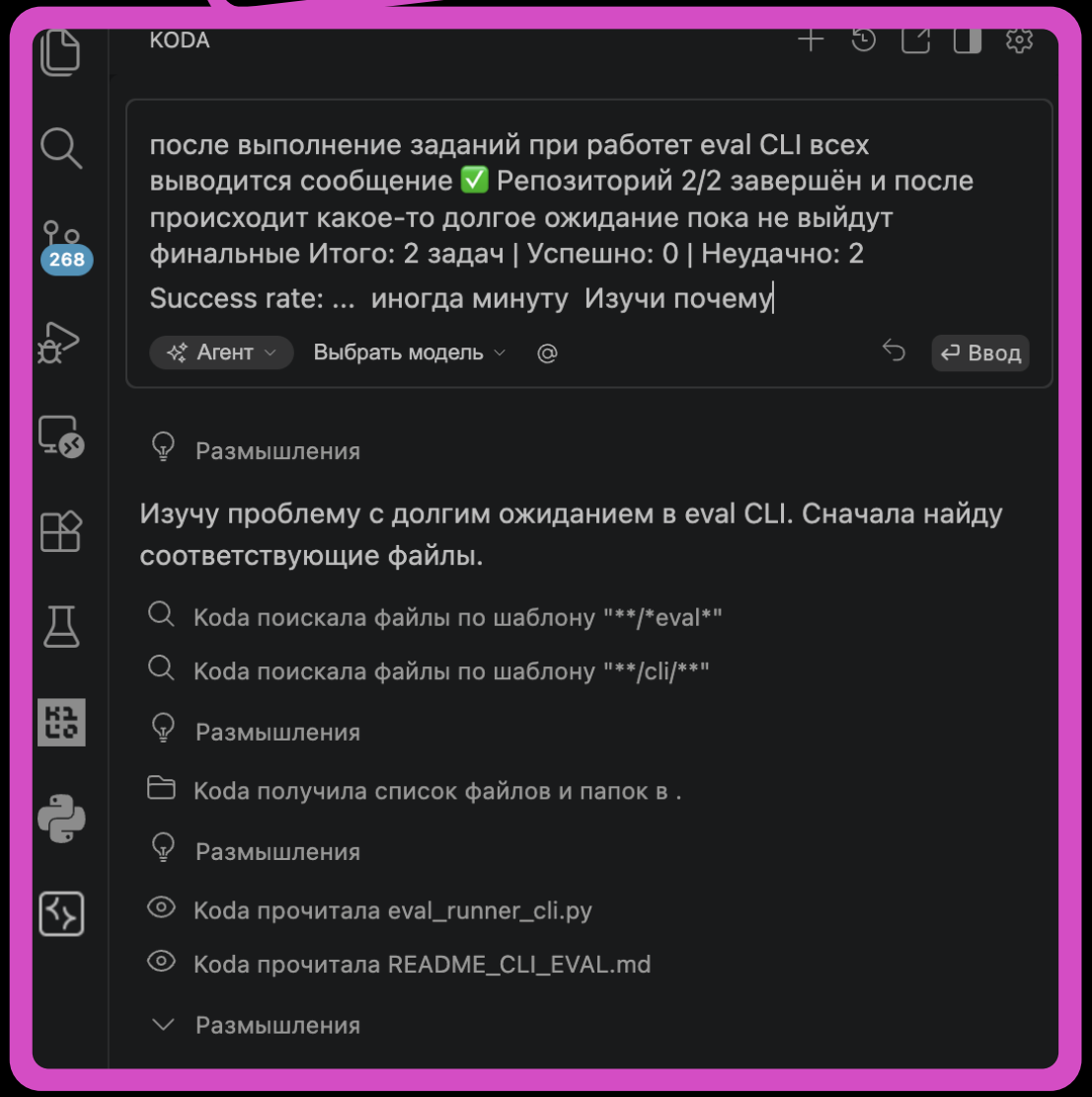
AI-ассистент для разработчиков с агентским режимом и поддержкой современных моделей

KodaCode отлично понимает русский язык, работает без VPN и доступен в VS Code, JetBrains IDE и терминале (CLI) — для всего цикла разработки.

Скачать | Документация



Kodacode



Kodacode — ИИ-ассистент разработчика

Kodacode

Kodacode

Наш ИИ-ассистент риентирован
на полный цикл разработки
ПО внутри компаний.

Основан на лучшем доработанном
опенсорсе и дообученненных моделях

По базовым компонентам
мы уже близки к уровню Cursor

Команда — авторы GigaCode
и многих моделей для русского
языка и кода от Сбера, которые
в свое время были SOTA

Некоторые модели по прежнему
SOTA в своем классе. Участвовали
в первых претрейнах GigaChat

Наше решение лучше поддерживает
работу на русском языке

Только факты

Kodacode

7 млрд
токенов в день генерируют
наши LLM-модели
Koda base и Koda pro

95 000
запросов в день
прилетает
на наши модели

Вся инфраструктура —
в России

Что говорят наши пользователи?

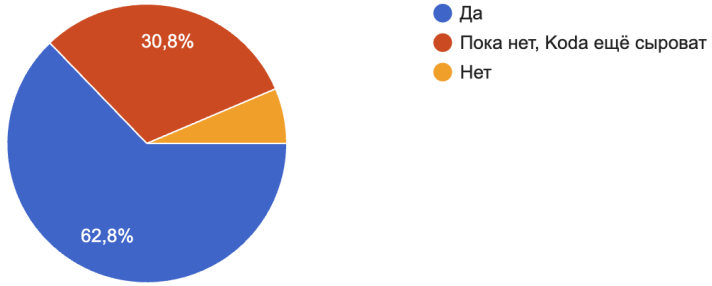
Kodacode

Использовал ли раньше аналогичные инструменты?

78 ответов

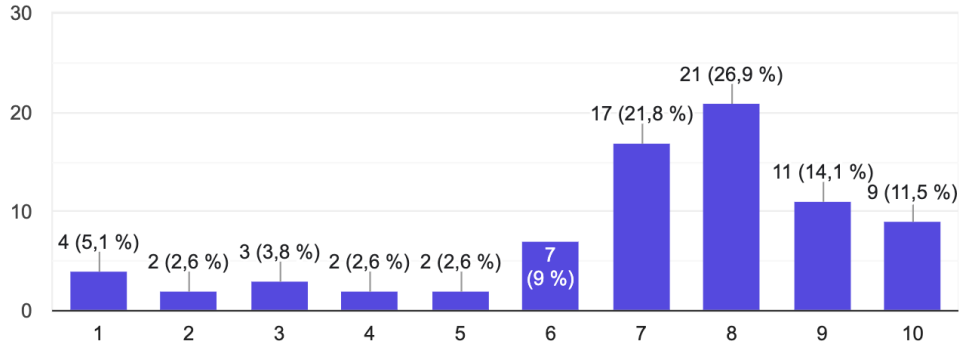
Можешь ли порекомендовать Koda знакомым или как внутренний инструмент для разработчиков в своей компании?

78 ответов



По шкале от 1 до 10, насколько удобным и полезным оказался Koda в твоей работе, где:

78 ответов



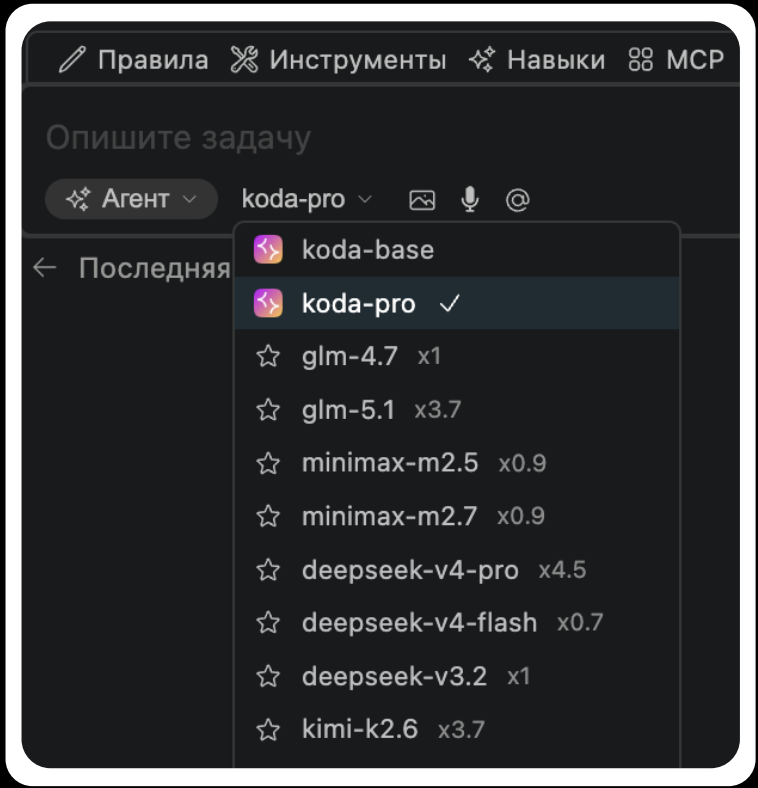
Модели и инфраструктура «под капотом»

Kodacode

8 моделей
развернутых
в России

B2C
GPT, Opus и другие —
через внешних провайдеров

B2B
только
наши модели



Модель	Размер	Нагрузка запросов в день	Базовые ресурсы	Инф-ра в пиках нагрузки
Koda base	122B	60к	2 H200	4 H200
Koda pro	1T	30к	8 H200	16 H200
KodaCode Completion	1.3B	250к	H200	H200
Koda NextEdit	1.5B	100k		
Koda embedder	0.6B	50к		
Koda reranker	0.6B	10к		
Apply model	1.7B	1к		
ASR распознавание речи		1k		

Как мы выбираем облачного провайдера

Kodacode

- Надежность инфраструктуры: нам нельзя упасть даже на 1 час
- Нужна возможность быстро масштабироваться на пиковую нагрузку с тарификацией за фактическое использование
- Нам предпочтителен кластер с infiniband для мультинод разворачивания большой LLM
- Мы быстро растем и нужна возможность масштабироваться на бОльшую инфраструктуру в дальнейшем
- Нужен облачный провайдер, которому доверяют: компании, которые используют наш SaaS, беспокоятся, чтобы их данные «не улетали» за рубеж и в целом хотят понимать, где могут оказаться их данные

[недавно] Нам нельзя упасть даже на час

Kodacode

Мы арендовали BareMetal с 8-ю H200 под нашу основную pro-модель

[недавно] Нам нельзя упасть даже на час

Kodacode

Мы арендовали BareMetal с 8-ю H200 под нашу основную pro-модель

- → Сгорела 1 GPU
  - → На 7 GPU уже не запустишь
    - → Только через день нам дали другую машину на замену — из тех, что были в наличии
      - → А настройка новой машины занимает время (1-2 часа)

[недавно] Нам нельзя упасть даже на час

Kodacode

Мы арендовали BareMetal с 8-ю H200 под нашу основную pro-модель

→ Сгорела 1 GPU

→ На 7 GPU уже не запустишь

→ Только через день нам дали другую машину на замену — из тех, что были в наличии

→ А настройка новой машины занимает время (1-2 часа)

Итого
1,5 дня

[недавно] Нам нельзя упасть даже на час

Kodacode

Мы арендовали BareMetal с 8-ю H200 под нашу основную pro-модель

→ Сгорела 1 GPU

→ На 7 GPU уже не запустишь

→ Только через день нам дали другую машину на замену — из тех, что были в наличии

→ А настройка новой машины занимает время (1-2 часа)

Итого
1,5 дня

→ 15 мин.

У нас не было тогда SaaS-пользователей и мы переключились за 15 минут на зарубежного облачного провайдера
И ПОЛЬЗОВАТЕЛИ НИЧЕГО НЕ ЗАМЕТИЛИ

Но сейчас у нас много SaaS пользователей

Отсутствие сервиса даже на день
может сильно пройтись по репутации

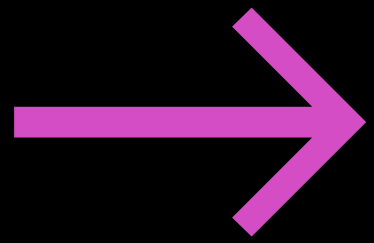
Решение:

Но сейчас у нас много SaaS пользователей

Kodacode

Отсутствие сервиса даже на день
может сильно пройтись по репутации

Решение:



Кластер

Он не сгорит весь

И перенос запуска на кластере
на другую ноду займет минуты, а не часы

Масштабирование от нагрузки — это важно

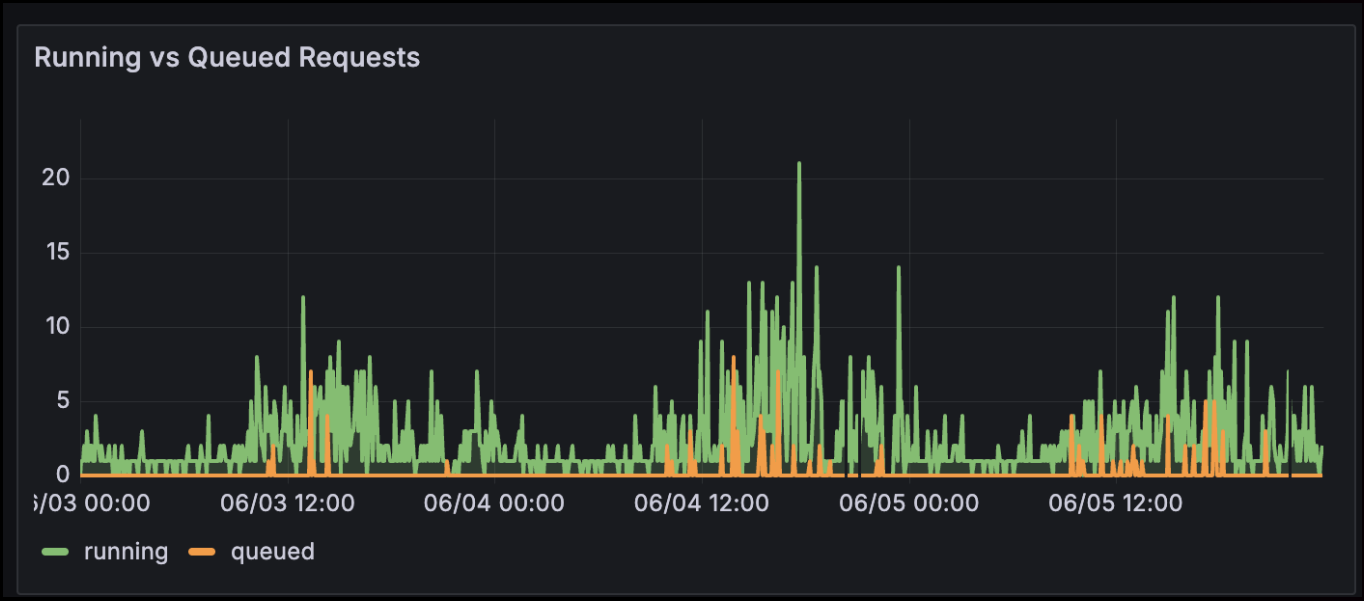
Kodacode

Нагрузка неравномерная

Нам нужна возможность масштабироваться
в течение дня на бОльшее число GPU:
очень важно для больших моделей Base и Pro

На пике инференс может «захлебнуться»

… и общая производительность GPU сильно падает.
Решение: ограничить max число запросов на инстанс модели
и «излишки» направлять на другой недавно поднятый инстанс



Возможность масштабироваться
автоматически в зависимости
от уровня одновременных запросов
и платить только за фактическое
время экономит нам деньги

Не нужно оплачивать максимальный
объем инфраструктуры full time —
когда держим её «про запас»
для пиковых нагрузок

Мультинод развертывание моделей

Kodacode

- → Развернули сначала Koda Pro на 8 H200
  - → Нагрузка выросла и нужно было масштабировать
    - → Поднять еще одну целую ноду и отправлять излишний трафик на нее — это плохо: модели весят много и занимают большую часть памяти GPU

Мультинод развертывание моделей

Kodacode

- → Развернули сначала Koda Pro на 8 H200
  - → Нагрузка выросла и нужно было масштабировать
    - → Поднять еще одну целую ноду и отправлять излишний трафик на нее — это плохо: модели весят много и занимают большую часть памяти GPU

НО так как в Турбо у нас кластер,
то, конечно же, мы подняли
модель сразу на 16 H200

Мультинод развертывание моделей

Kodacode

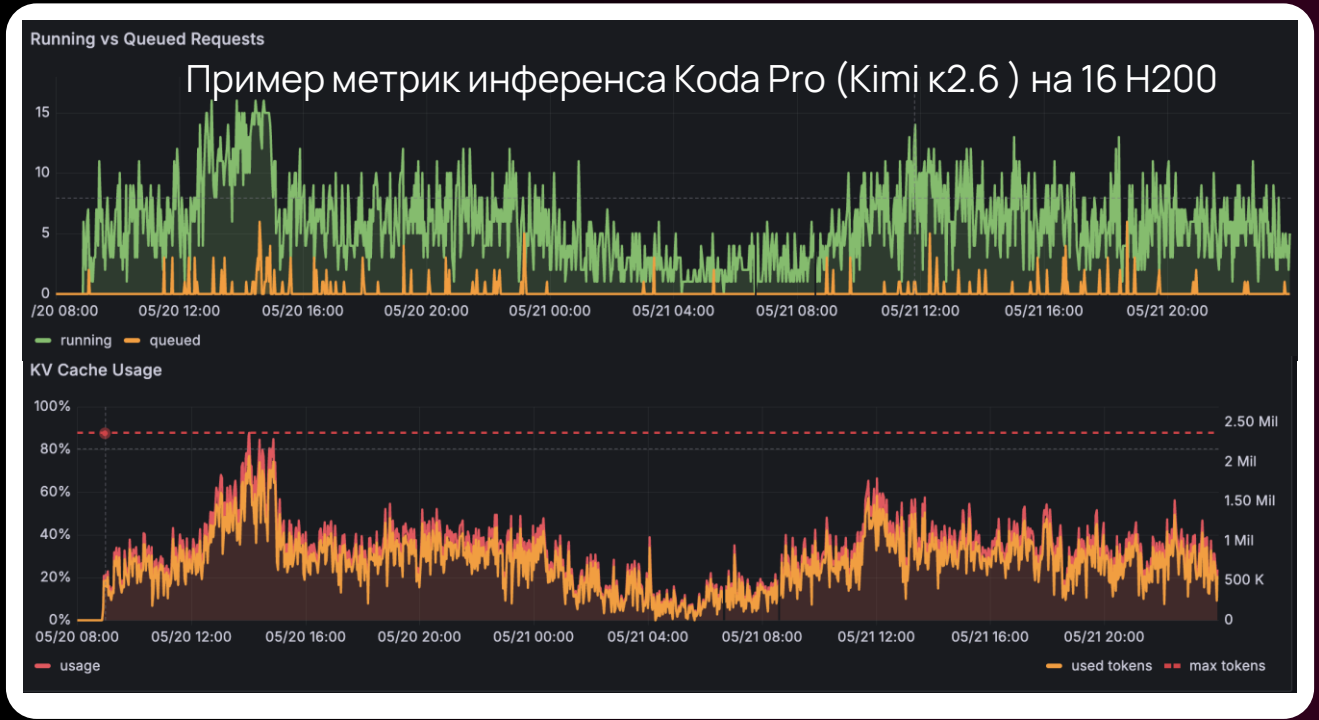
→ Развернули сначала Koda Pro на 8 H200

→ Нагрузка выросла и нужно было масштабировать

→ Поднять еще одну целую ноду и отправлять излишний трафик на нее — это плохо: модели весят много и занимают большую часть памяти GPU

НО так как в Турбо у нас кластер, то, конечно же, мы подняли модель сразу на 16 H200

→ кластер решает ☺



Почему выбрали Inference Platform?

Kodacode

Кластер, Кластер, Кластер …

- Дает много гибкости в высоконагруженном инференсе
- Автоскейлинг для моделей
- Тарификация за фактическое использование

Облачный провайдер, которому доверяют данные

В России немного компаний, у кого можно взять кластер для инференса в аренду

Kodacode

kodacode.ru

AI INFERENCE

Запуск любых AI-моделей за несколько минут

Передайте модель — платформа развернет окружение и отдаст готовый API endpoint

Hugging Face · S3 · Private Repo · Fine-tuned

Автозагрузка из Hugging Face и S3

Просто укажите источник модели – платформа развернет ее сама

99,95%
SLA
Доступность с перерасчётом стоимости при нарушении

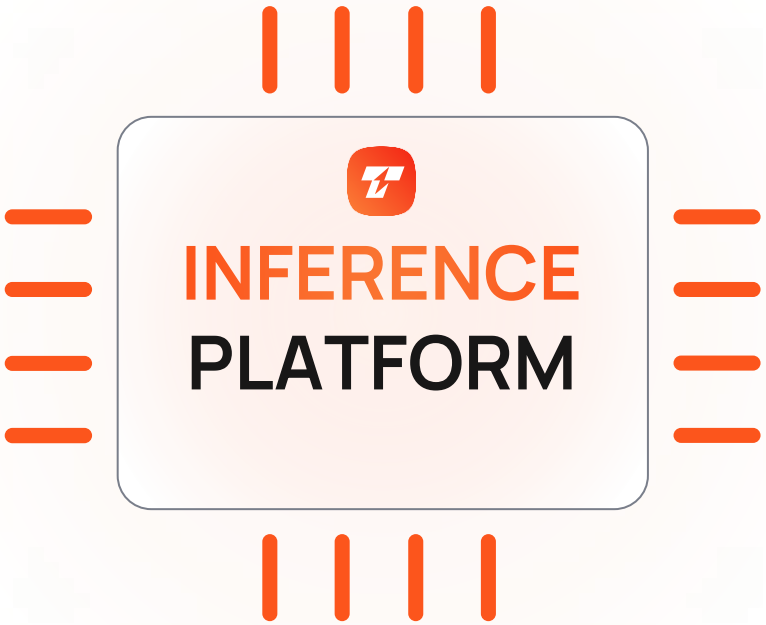
Без своего кластера

Кластер, Kubernetes, GPU, инженеры — на стороне платформы

5 ГБ
шаг выделения vRAM
GPU-память тарифицируется по факту использования

Scale-to-zero

В простое инстанс не тарифицируется. При трафике поднимается сам



ГБ/ч
только за потребление
не нужно резервировать сервер целиком

КЛАСТЕР
NVIDIA H200

ГБ/мес.
хранения кэша
Модель остается готовой к быстрому запуску

Автомасштабирование

Реплики растут по RPS и параллельным запросам. Вы задаёте min и max

Готовый endpoint

Модель, окружение и доступы собираются в готовый API-контур

1,2 ТБ
максимум на 1 контейнер
GPU-память тарифицируется по факту

Крупные модели

Multi-node inference и tensor parallelism распределяют модель по серверам.

Совместимость с OpenAI API

Для LLM-моделей достаточно поменять base URL — интеграция без переписывания клиентского кода

Прозрачная тарификация по выделенной GPU-памяти **за 1 ГБ памяти в час**

Как протестировать ИИ-сервисы бесплатно

1

Обратитесь к вашему аккаунт-менеджеру или подайте заявку с сайта TurboCloud.ru

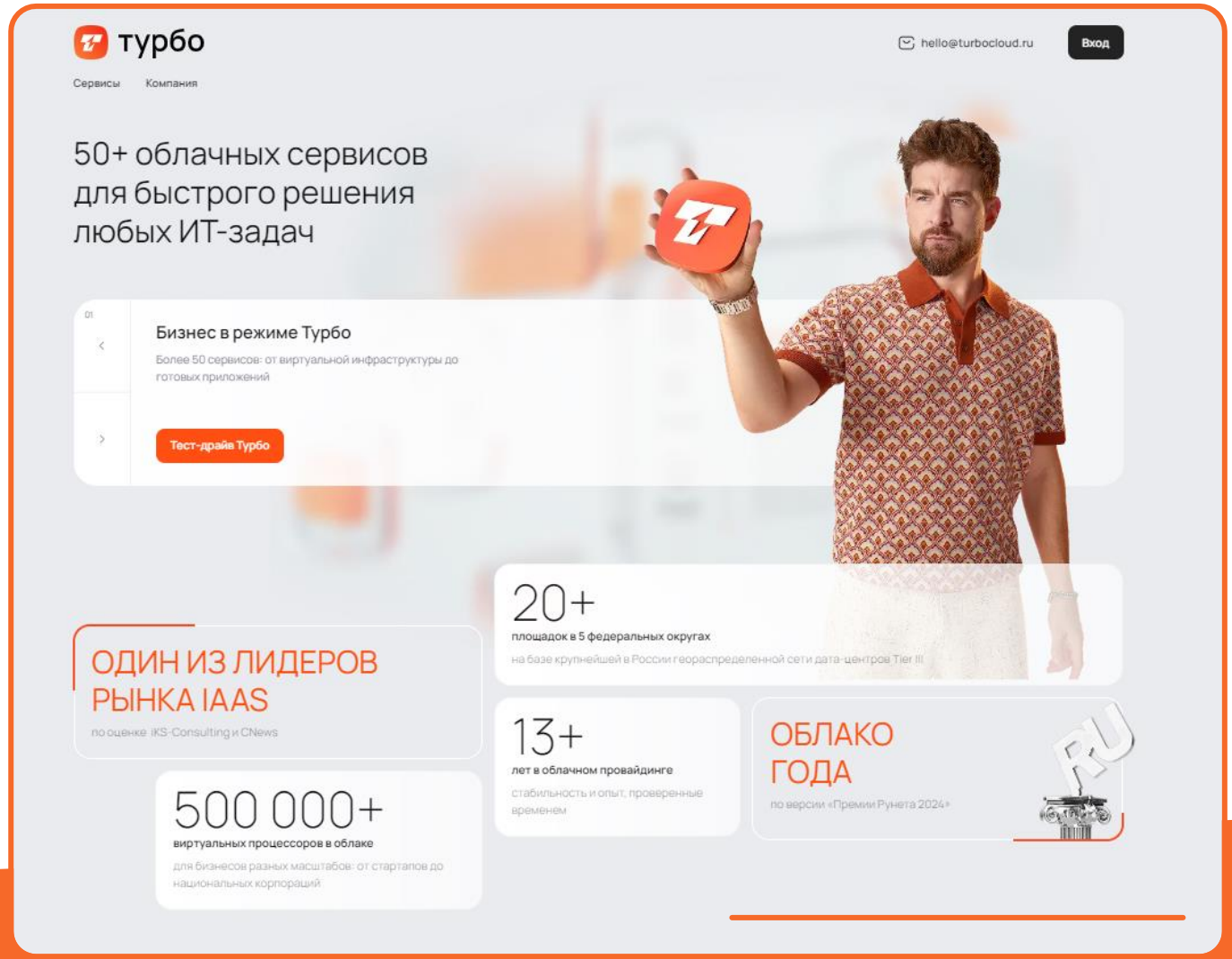
2

Заполните опросный лист, укажите модель, требуемые ресурсы

3

Получите доступ на 14 дней и тестируйте без ограничений

БЫСТРЫЙ СТАРТ С ТУРБО



TurboCloud.ru

турбо

Kodacode

Ваши вопросы?

Денис Боженко
Ведущий менеджер продуктов
Группа сервисов
искусственного интеллекта
ТУРБО

Максим Глотов
Руководитель группы
архитектуры и проектирования
облачных решений
ТУРБО

Дмитрий Змитрович
Основатель
КОДА