softline ОБЛАКО

От ChatGPT к AI в бизнесе:
какие задачи требуют собственной GPU-инфраструктуры в облаке

Козлов Алексей, Зацепин Роман

Менеджеры продуктов в Софтлайн Облако

Повестка

1. Причины популярности ИИ

2. Когда ИИ стоит внедрять в организации

3. Архитектура ИИ моделей

4. ИИ модели: типы, основные характеристики, стоимость

5. ИИ агенты: виды и место в корп. ИИ решении

6. ИИ решения на практике: примеры

7. Портфель продуктов Софтлайн облако

8. Демонстрация решения

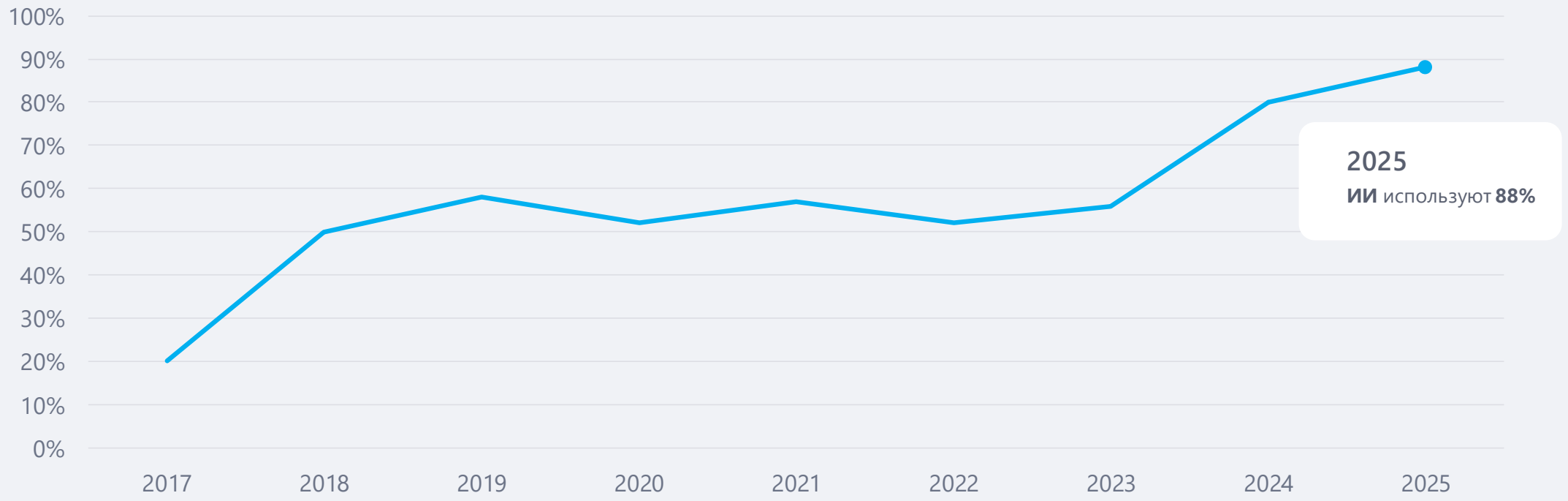
9. Графические ускорители и сценарии использования

Опрос:
Для каких задач вы используете / планируете использовать ИИ

1. Повседневные рабочие задачи (чат-боты)
2. Коммерческое ПО (или устройства) собственной разработки
3. Оптимизация собственной инфраструктуры или бизнес процессов
4. Написание кода
5. Для всего. Не принимаю ни одного решения без совета ИИ
6. Свой вариант

ИИ - от узкой специализации в массы

Доля компаний, использующих ИИ хотя бы в одной бизнес-функции*



*По данным McKinsey

Причины резкого роста популярности ИИ

Запуск ChatGPT V4

Первый в истории продукт, набравший 100 млн пользователей за 2 месяца. Интерфейс чата сделал мощный ИИ понятным и доступным каждому без обучения.

Демократизация через открытые модели

Появление открытых и относительно компактных моделей (Llama, Mistral) позволило бизнесу любого размера запускать ИИ на своих серверах, резко снизив порог входа и зависимость от вендоров.

Доказанная экономическая эффективность

Компании быстро увидели измеримый ROI: сокращение времени на создание контента на 60–80%, снижение затрат на первую линию поддержки, ускорение цикла разработки ПО в 2–3 раза.

Резкое увеличение вычислительных мощностей GPU

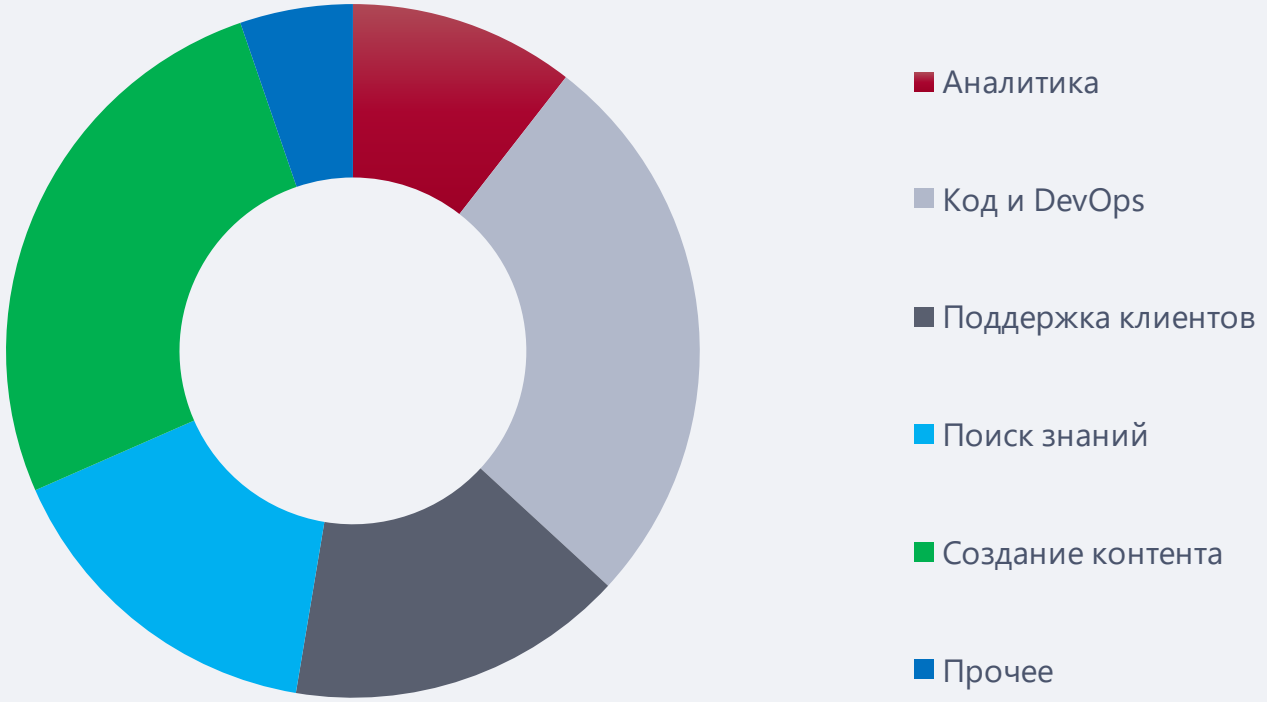
Внедрение тензорных ядер и аппаратной поддержки вычислений низкой точности увеличивало ИИ-производительность в 3–5 раз, радикально сокращая время обучения.

Инфраструктурный взрыв

Рекордные инвестиции гиперскейлеров ($250 млрд в 2025 г.) в GPU и AI-облака привели к многократному росту производительности и падению стоимости инференса.

Секторы – лидеры по использованию ИИ

Доля компаний, использующих ИИ хотя бы в одной бизнес-функции*



*По данным McKinsey

Только около 1/3 организаций смогли масштабировать ИИ за пределы пилотов!

Нужно ли мне инвестировать/внедрять ИИ? Экспресс-оценка

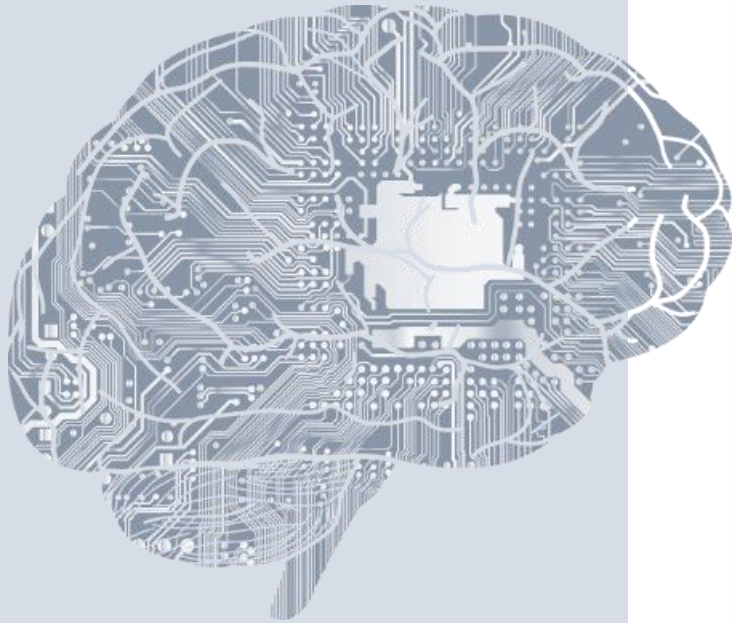
Вопрос	Да/Нет
Есть много ручной работы с информацией?	☐
Есть повторяемые процессы?	☐
Есть накопленные данные?	☐
Есть кадровый дефицит?	☐
Ошибки дорого стоят?	☐
Планируется рост бизнеса?	☐
Есть измеримые KPI?	☐

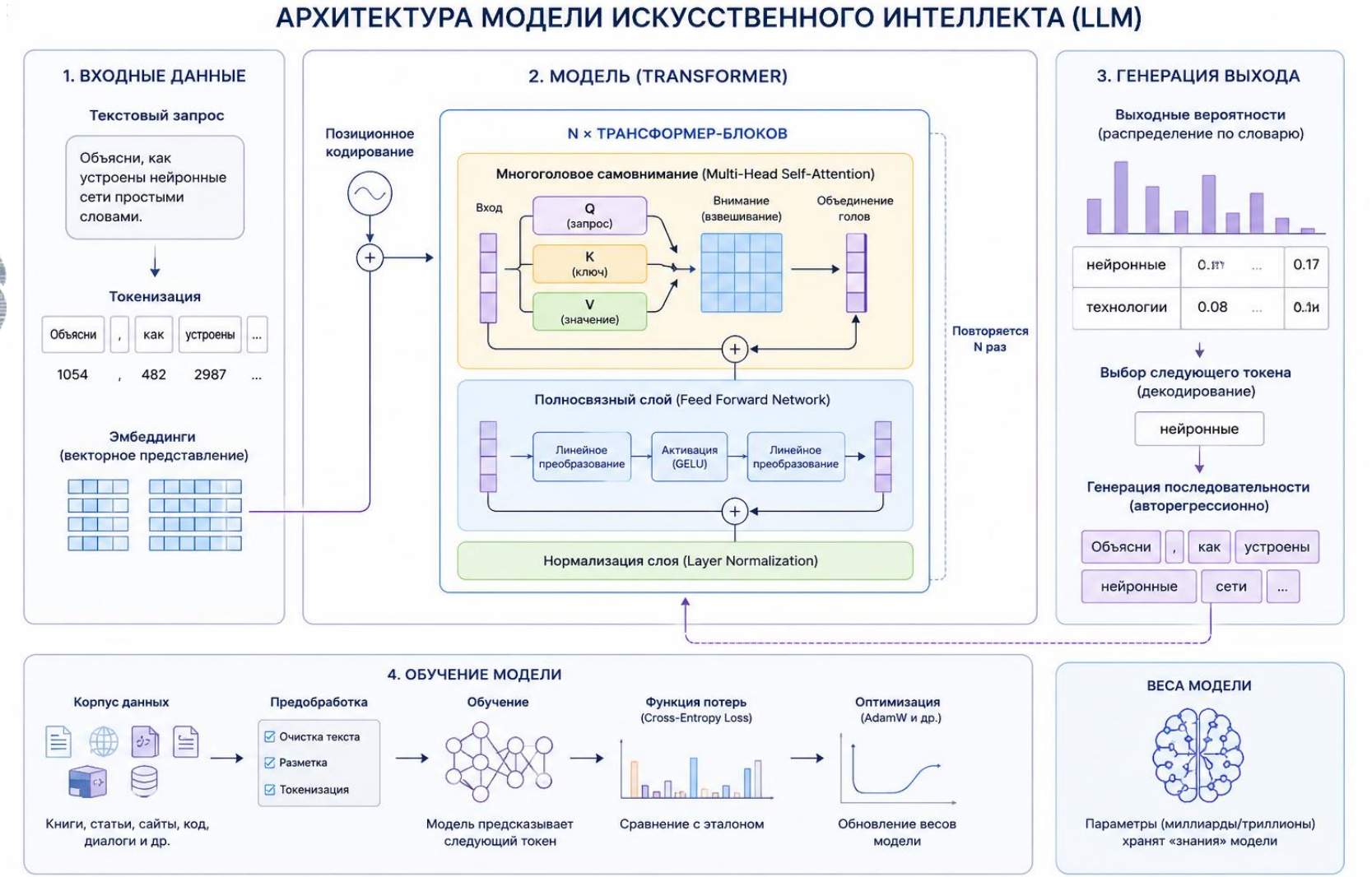
Если ИИ способен сократить хотя бы 10–15% операционных затрат или повысить производительность на 20–30%, инвестиции обычно окупаются в течение 6–18 месяцев.

Главные ошибки при внедрении ИИ в компании

1. Старт без чёткой бизнес-метрики /задачи
2. Внедрение ради технологии, а не для решения боли
3. Игнорирование данных и их качества
4. Отсутствие ответственного владельца и кросс-функциональной команды
5. Завышенные ожидания и «магия AGI»
6. Пренебрежение людьми и изменениями
7. Отсутствие цикла непрерывных улучшений
8. Попытка сразу автоматизировать сложные сквозные процессы

Модель ИИ (LLM) – мозг любого ИИ решения





Какую модель ИИ выбрать?

Тип модели	Назначение	Применение в бизнесе	Проприетарные модели	Открытые модели
БОЛЬШИЕ ЯЗЫКОВЫЕ (LLM)	Генерация текста, диалоги, рассуждения, написание кода	Чат-боты, AI-агенты, автогенерация отчётов, копирайтинг, колл-центры	GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), Gemini 1.5 Pro (Google)	Llama 3.1 405B (Meta), Qwen2.5-72B (Alibaba), DeepSeek-V2, Mistral Large 2
НЕБОЛЬШИЕ ЯЗЫКОВЫЕ (SLM)	Та же текстогенерация при низких требованиях к GPU и задержке	On-premise RAG-системы, ассистенты на одном GPU, работа в закрытом контуре без интернета	— (обычно не поставляются как API)	Phi-3-mini (3.8B), Phi-3-small (7B) от Microsoft; Qwen2.5-7B/14B; Gemma 2 (9B/27B) от Google; Llama 3.2 (3B/1B) от Meta
ЭМБЕДДИНГ-МОДЕЛИ	Перевод текста в числовой вектор для семантического поиска	Умный поиск по внутренним документам, подбор товаров, выявление дубликатов, основа для RAG	OpenAI text-embedding-3-large, Cohere Embed v3	intfloat/multilingual-e5-large, BAAI/bge-m3, all-MiniLM-L6-v2 (через Sentence-Transformers)
МУЛЬТИМОДАЛЬНЫЕ	Одновременное понимание текста, изображений, аудио, видео	Разбор сканов накладных и паспортов, анализ фото с производства, описание товаров по снимкам	GPT-4o / GPT-4V, Gemini 1.5 Pro (видео), Claude 3.5 (с изображениями)	LLaVA 1.6, CogVLM2, Qwen-VL
ГЕНЕРАТИВНЫЕ ДЛЯ ИЗОБРАЖЕНИЙ	Создание картинок, логотипов, макетов по текстовому описанию	Маркетинговые креативы, прототипирование дизайна, визуализация товаров	Midjourney V6/V7, DALL-E 3, Adobe Firefly	Stable Diffusion XL (SDXL), Stable Diffusion 3 (SD3), FLUX.1
КОМПЬЮТЕРНОЕ ЗРЕНИЕ (CV)	Детекция объектов, сегментация, распознавание лиц, OCR	Контроль качества на конвейере, пропускные системы, оцифровка документов, мониторинг безопасности	— (узкоспециализированные решения в составе платформ)	YOLOv8 / YOLOv9, Meta SAM 2 (сегментация), TrOCR, PaddleOCR, YOLO-NAS
РЕЧЕВЫЕ (STT / TTS)	Преобразование речи в текст и обратно	Транскрибация звонков, голосовые ассистенты, озвучка обучающих материалов	Whisper (OpenAI, через API), ElevenLabs (TTS)	Whisper.cpp / faster-whisper (STT), NVIDIA Riva (on-premise STT/TTS), Coqui TTS, XTTS-v2
КОДОГЕНЕРИРУЮЩИЕ	Написание, объяснение, отладка кода	Copilot для разработчиков, автоматическое создание SQL-запросов, миграция легаси-кода	GitHub Copilot (Codex)	Code Llama (Meta), StarCoder 2, DeepSeek-Coder, Codestral

ИИ модели – основные характеристики

Характеристика	Что показывает	Примеры моделей
КАЧЕСТВО РАССУЖДЕНИЙ	Способность решать сложные задачи, писать код, анализировать документы	Высокое: GPT-5, Claude Opus, Gemini 2.5 Pro
МОДАЛЬНОСТЬ	Работа с текстом, изображениями, аудио и видео	Текст: DeepSeek-R1; Текст+Изображения: Claude Sonnet; Полная мультимодальность: Gemini 2.5 Pro
РАЗМЕР КОНТЕКСТА	Объем данных, обрабатываемый за один запрос	~128K: GPT-5; ~200K: Claude Sonnet; до 1–2M: Gemini 2.5 Pro
СКОРОСТЬ РАБОТЫ	Скорость генерации ответа	Быстрые модели: GPT-5 Mini, Llama 3 8B, Qwen 3 8B
ПОДДЕРЖКА ЯЗЫКОВ	Работа на разных языках	GPT, Claude, Gemini, Qwen — **100+ языков**
ИНТЕГРАЦИЯ И АГЕНТЫ	Подключение к API, БД, ERP, CRM	GPT + Agents, Claude + MCP, Gemini + Google Workspace
СТОИМОСТЬ	Стоимость эксплуатации	От локальных open-source моделей (Llama, Qwen) до премиальных GPT и Claude
БЕЗОПАСНОСТЬ	Возможность локального размещения	On-Premise: Llama 3 70B, Qwen 3, DeepSeek

ИИ модели – проприетарные или с открытым кодом

Критерий	Проприетарные модели	Открытые модели
ДОСТУП К ВЕСАМ И КОДУ	**Веса и архитектура закрыты**. Доступ только через API (иногда через партнёрские лицензии).	**Веса опубликованы, архитектура известна**, часто открыт и код для инференса/обучения.
СПОСОБ ИСПОЛЬЗОВАНИЯ	**Как облачный сервис:** вы отправляете запросы на сервер провайдера.	**Можно скачать и запустить на своём сервере,** в своём облаке или даже на локальной машине.
СТОИМОСТЬ	**Оплата за токены (входные/выходные)** или фиксированная подписка. Дорого при больших объёмах.	**Бесплатное использование (лицензия Apache 2.0, MIT и т.п.).** Вы платите только за «железо» и электричество.
КОНТРОЛЬ ДАННЫХ И КОНФИДЕНЦИАЛЬНОСТЬ	**Данные уходят вовне.** Есть риски: провайдер может логировать запросы (даже с оговорками об отключении). Для строгих регуляторов (GDPR, 152-ФЗ) может быть недопустимо.	**Данные не покидают ваш контур.** Полный контроль над трафиком и хранением. Подходит для чувствительных корпоративных или персональных данных.
ПРОИЗВОДИТЕЛЬНОСТЬ И ТОЧНОСТЬ	**Часто лидируют в бенчмарках** (огромные бюджеты на обучение, постоянные обновления). Могут быть более «отполированными».	**Быстро догоняют:** Llama 3, Qwen 2.5 уже на уровне GPT-4 во многих задачах. Требуют грамотного хостинга для быстрого инференса.
КАСТОМИЗАЦИЯ И ДООБУЧЕНИЕ	**Ограниченный файн-тюнинг через API** (не все провайдеры разрешают, и только определёнными методами). Нельзя менять архитектуру.	**Полная свобода:** можно дообучать (LoRA, полный файн-тюнинг), квантовать, дистиллировать, модифицировать под свои задачи.
НАДЁЖНОСТЬ И ПОДДЕРЖКА	**SLA, круглосуточная поддержка**, гарантированная доступность.	**Ответственность лежит на вас.** Нужна своя экспертиза по развёртыванию и обслуживанию.
ПРОЗРАЧНОСТЬ И АУДИТ	**«Чёрный ящик»:** нельзя проверить, на каких данных обучали, нет ли скрытых предвзятостей.	**Полностью прозрачны:** можно проанализировать датасеты (если они открыты), провести аудит безопасности.
ПРИМЕРЫ	**GPT-4o (OpenAI), Claude 3.5 (Anthropic), Gemini 1.5 (Google)**	**Llama 3.1 (Meta), Qwen 2.5 (Alibaba), Mistral Large 2 (Mistral AI), DeepSeek-V2**

ИИ модель: арендовать и не разориться

Критерий	ВМ с GPU (self-hosted на облаке)	Открытая модель по токенам (Managed API)
ЭКОНОМИКА	**Фиксированная плата в месяц.** Выгодно при высокой и стабильной нагрузке (сотни млн токенов/мес) или при специфических требованиях.	**Оплата за токены.** Выгодно при малых, нерегулярных нагрузках (пилоты, прототипы, сезонные всплески).
КОНТРОЛЬ И КАСТОМИЗАЦИЯ	**Полный:** дообучение (LoRA/full FT), квантование, изменение промптов, управление версиями.	**Ограничен:** обычно только промпт-инжиниринг и параметры инференса. Дообучение либо невозможно, либо дорого.
ПРИВАТНОСТЬ ДАННЫХ	**Максимальная:** данные не покидают вашу ВМ. (Но нужно учитывать, что сама ВМ находится в облаке провайдера).	**Зависит от провайдера:** данные передаются на его API. Даже при отключённом логировании риски утечки выше, чем на своей ВМ.
АДМИНИСТРИРОВАНИЕ	**Требуется:** установка драйверов, фреймворков (vLLM, TGI), мониторинг, масштабирование. Нужны DevOps/MLOps-компетенции.	**Не требуется:** провайдер управляет инфраструктурой, обновляет модель, обеспечивает доступность.
ПРОИЗВОДИТЕЛЬНОСТЬ И ЗАДЕРЖКА	При грамотной оптимизации (TensorRT-LLM, vLLM) можно достичь низких и предсказуемых задержек. **Вы управляете SLA сами.**	**Провайдер гарантирует SLA,** но возможна латентность сети и «шумные соседи».
МАСШТАБИРОВАНИЕ	**Ограничено мощностью арендованной ВМ.** При росте нужно вручную добавлять узлы или переезжать на более крупный инстанс.	**Автоматическое**: API-провайдер горизонтально масштабирует под нагрузку.
ОБНОВЛЕНИЕ МОДЕЛИ	**Требует ручного развёртывания новой версии** (скачать веса, перезапустить инференс).	**Автоматически применяется провайдером** (опционально можно зафиксировать версию).

ИИ агенты: работники вместо советников

ИИ-агенты – это цифровые сотрудники будущего

ЧТО ТАКОЕ ИИ АГЕНТ?

ИИ-агент – это автономная система, которая понимает цель, планирует действия, использует инструменты, и выполняет задачи для достижения результата с минимальным участием человека

ДЛЯ ЧЕГО НУЖНЫ ИИ АГЕНТЫ?

ИИ-агенты ~~VS~~ модели ИИ

Агент без модели – скрипт, модель – собеседник

МОДЕЛИ ИИ (LLM И ДР.)

- ✓ Генерируют ответы на запросы
- ✓ Работают в рамках одного запроса и контекста
- ✓ Не обладают самостоятельностью
- ✓ Не используют инструменты и системы
- ✓ Требуют постоянных инструкций от пользователя

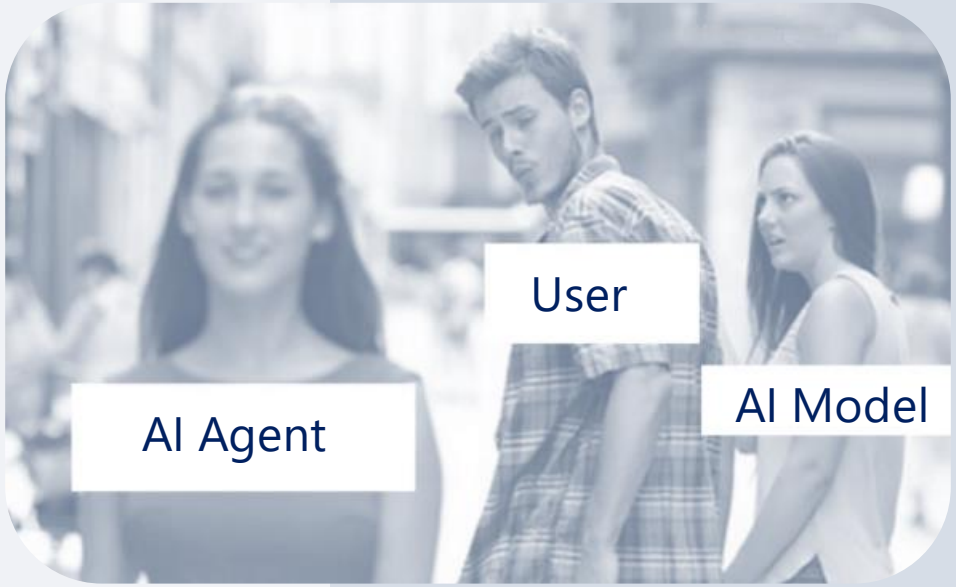
Пример: GPT-4, Claude, Llama 3

VS

ИИ-АГЕНТЫ

- ✓ Генерируют ответы на запросы
- ✓ Работают в рамках одного запроса и контекста
- ✓ Не обладают самостоятельностью
- ✓ Не используют инструменты и системы
- ✓ Требуют постоянных инструкций от пользователя

Пример: AutoGPT, Devin, Copilot, Jasper



ИИ агенты: виды и назначение 1/2

1. ПЕРСОНАЛЬНЫЕ АССИСТЕНТЫ

Помогают в повседневных задачах, планировании, поиске информации, управлении временем

ChatGPT

Microsoft Copilot

Google Gemini

Reclaim AI

2. КОРПОРАТИВНЫЕ АГЕНТЫ

Автоматизируют бизнес-процессы, работают с корпоративными данными, системами и документами

Salesforce Agentforce

Microsoft Copilot Studio

UIPath Autopilot

IBM watsonx Orchestrate

3. АГЕНТЫ ДЛЯ РАЗРАБОТКИ И IT-ОПЕРАЦИЙ

Помогают писать код, тестировать, анализировать ошибки, управлять инфраструктурой

GitHub Copilot

Cursor

Devin

Amazon CodeWhisperer

ИИ агенты: виды и назначение 2/2

4. АНАЛИТИЧЕСКИЕ АГЕНТЫ

Анализируют данные, строят отчеты, находят инсайты и делают прогнозы

DataRobot

ThoughtSpot Sage

Qlik AutoML

Hex

5. КРЕАТИВНЫЕ АГЕНТЫ

Создают тексты, изображения, видео, презентации и другой контент

Jasper

Midjourney

Canva

Runway

6. ОТРАСЛЕВЫЕ АГЕНТЫ

Решают специализированные задачи в медицине, финансах, производстве, логистике и др.

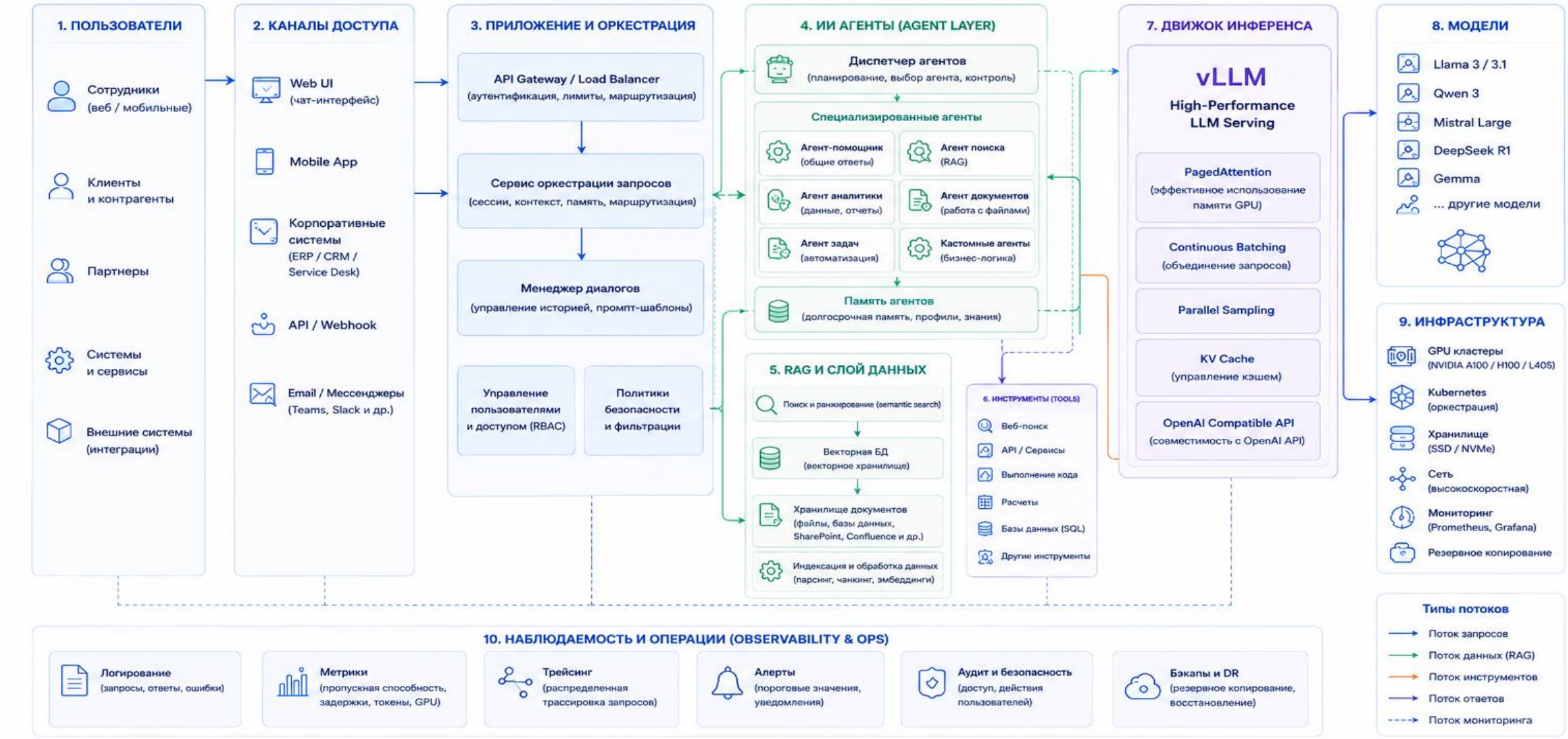
K Health

Harvey AI

Blumberg GPT

Augmentir

Архитектура корпоративного ИИ решения (AI ассистент)



Сценарии использования ИИ для бизнеса: кейс из практики Софтлайн облако

Заказчик – крупный агрохолдинг

Задача – оптимизация корпоративных сценариев: юридический анализ договоров, краткое описание предмета, сверка Word и сканированного PDF, заполнение карточки документа примерно на 50 реквизитов, корпоративный чат для сотрудников

Набор ресурсов – 2×VM с 1×NVIDIA H200, 48 vCPU, 384 GB RAM, 1024 GB Disk Tier 1. Backup 2048 GB

Модели ИИ – Qwen/Qwen3-235B-A22B-Instruct-2507-FP8, Qwen/Qwen3-Coder-Next, Openai/whisper-large-v3-turbo

Достигнуты цели:

- Ускорение процесса работа сотрудников с помощью ИИ в 5 бизнес-сценариях;
- Дообучение моделей ИИ до уровня, достаточного для последующего этапа интеграции в бизнес-процессы

Софтлайн облако - виртуальные машины с GPU

AI/ HPC/VDI

Флагманские GPU для LLM и HPC

- Виртуальные Машины **с 1-2-4-8** GPU **H200**

ФИКСИРОВАННЫЕ КОНФИГУРАЦИИ			
SKU	**GPU, qt**	**vCPU**	**vRAM, gb**
H200-1.12.128	1	12	128
H200-1.24.192		24	192
H200-2.24.256	2	24	256
H200-2.48.384		48	384
H200-4.48.512	4	48	512
H200-4.96.768		96	768
H200-8.96.1024	8	96	1024
H200-8.192.1536		192	1536

ПРОИЗВОЛЬНЫЕ КОНФИГУРАЦИИ					
SKU	**GPU, qt**	**vCPU**		**vRAM, gb**	
		min	**max**	**min**	**max**
H200-1.12-48.128-192	1	12	48	128	192
H200-2.24-96.256-384	2	24	96	256	384
H200-4.48-96.512-768	4	48	96	512	768
H200-8.96-192.1024-2000	8	96	192	1024	2000

GPU для ИИ, рендеринга и перекодирования

Через контур open stack/ЛК

- Публичный контур + возможность выделенного контура под проект
- **До 8 GPU на BM**/хост для обучения и инференса, сетевые тома с репликацией, сетевое хранилище, IP, сеть 1–10 Gb, базовые образы
- Линейка ускорителей**: H100 80GB, A100 80GB; RTX 6000 96GB**, L4, A10, A2, T4; **RTX 5090/4090/3090/2080**
- Защита от DDoS и облачный файрвол
- Intel Xeon 3-го и 5-го поколения, **1 ЦОД Tier III в Москве,** жёсткая **сегментация L2/L3 (VLAN/VXLAN)**
- 24×7 техподдержка

Пилоты до 2-3 дней, > по согласованию

Демонстрация ИИ модели и агентов

Графические ускорители и сценарии использования

B300 / B200

Blackwell / Blackwell Ultra для AI-фабрики и самых тяжелых ИИ-контуров

Не “ускоритель для эксперимента”, а производственный уровень ИИ-инфраструктуры

- обучение, дообучение и инференс с длинным контекстом
- reasoning-модели, ИИ-агенты, мультимодальный ИИ
- крупные конфигурации с несколькими GPU конфигурации и ИИ-фабрика (ИИ-фабрика (AI Factory))

B300

Blackwell Ultra, 2.3 TB GPU-памяти с 8 картами

B200

1,44 TB GPU-памяти с 8 картами

HBM + NVLink

критично для больших моделей

Доступность у нас

Публичное облако, выделенные серверы

Инфраструктура. Надёжная. Защищённая.

softline® ОБЛАКО

Бизнес-кейс: AI Factory - производственный контур для корпоративного ИИ

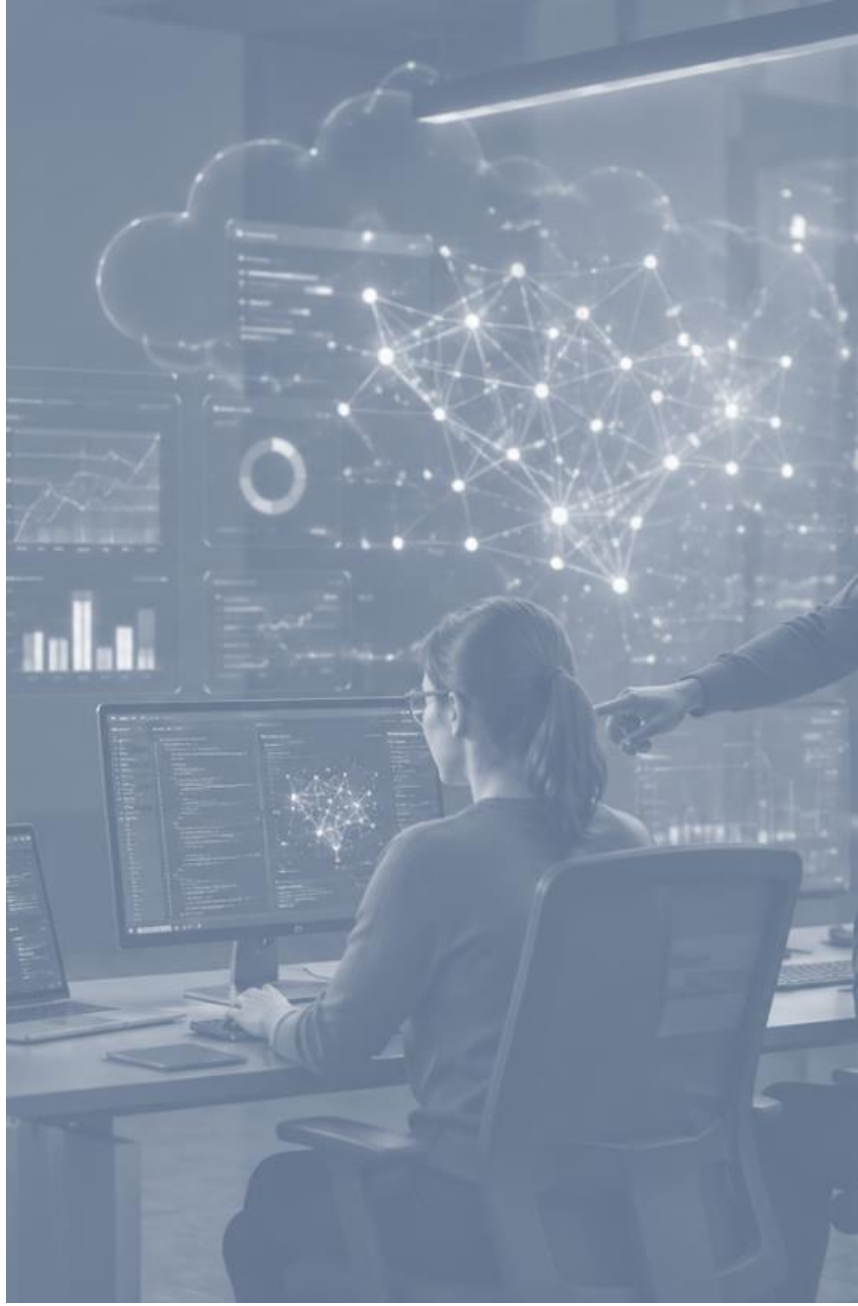
Единая платформа для внутренних и клиентских ИИ-сервисов

Сценарии клиента:

- RAG по внутренним базам знаний и документам
- корпоративные ассистенты и ИИ-агенты
- генерация, суммаризация, классификация
- инференс для большого числа пользователей

Когда нужен этот класс

ИИ становится сервисом в промышленной эксплуатации: важны пропускная способность, изоляция данных, управляемость и масштабирование.



Enterprise ИИ: H200 / H100 / A100

Класс для промышленной эксплуатации LLM, машинного обучения, компьютерного зрения, анализа данных и HPC

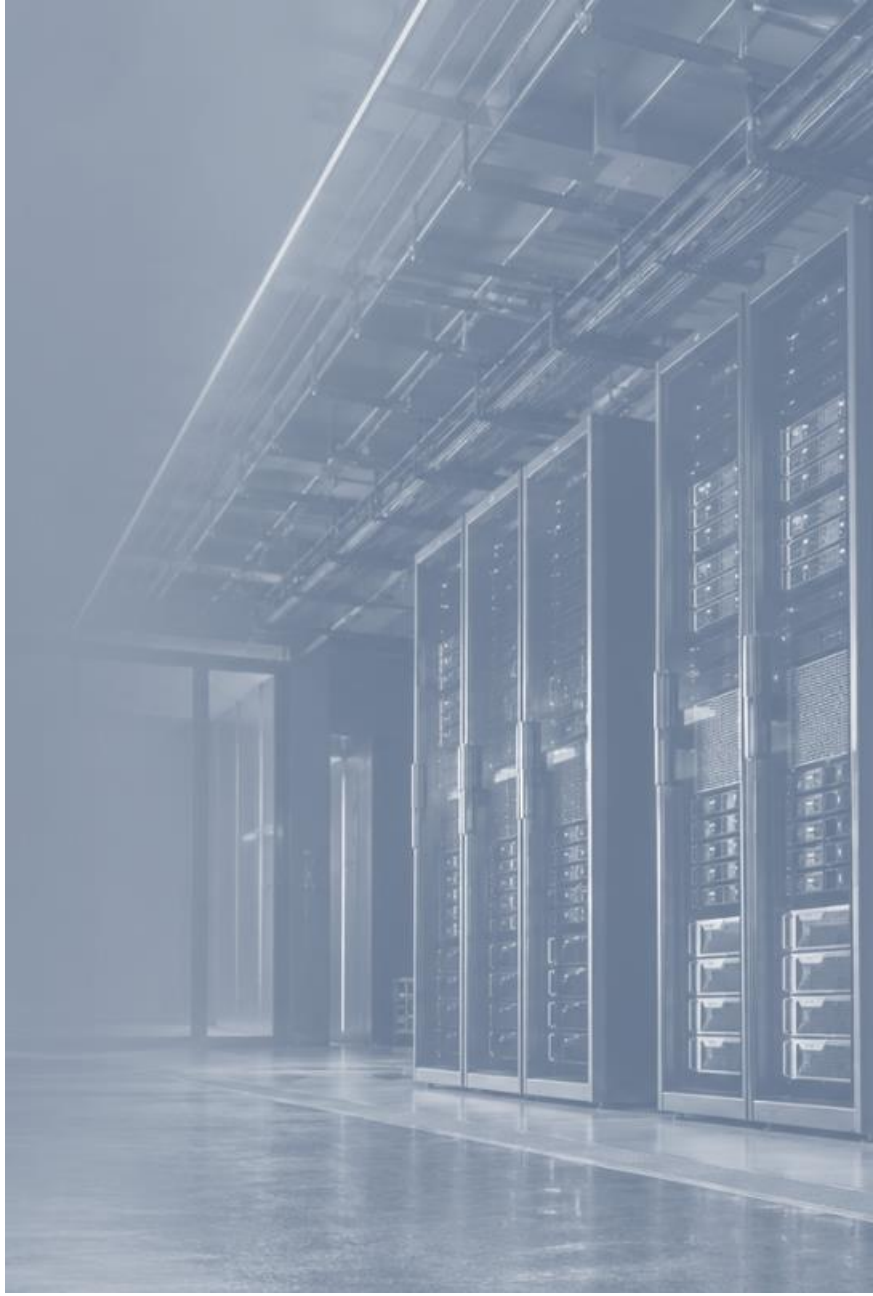
H200	141 GB HBM3e: крупный инференс, дообучение, HPC и модели с большим контекстом
H100 / H100 NVL	GPU для промышленной эксплуатации ИИ: обучение, инференс, transformer-модели, конфигурации с несколькими GPU.
A100 40/80 GB	Производительный ускоритель для машинного обучения/HPC и задач с понятной экономикой

Почему отдельный класс

HBM-память и серверная архитектура важны там, где RTX-класс ограничен по памяти, связности между GPU или стабильности под длительной нагрузкой

Доступность у нас

Публичное облако, выделенные серверы



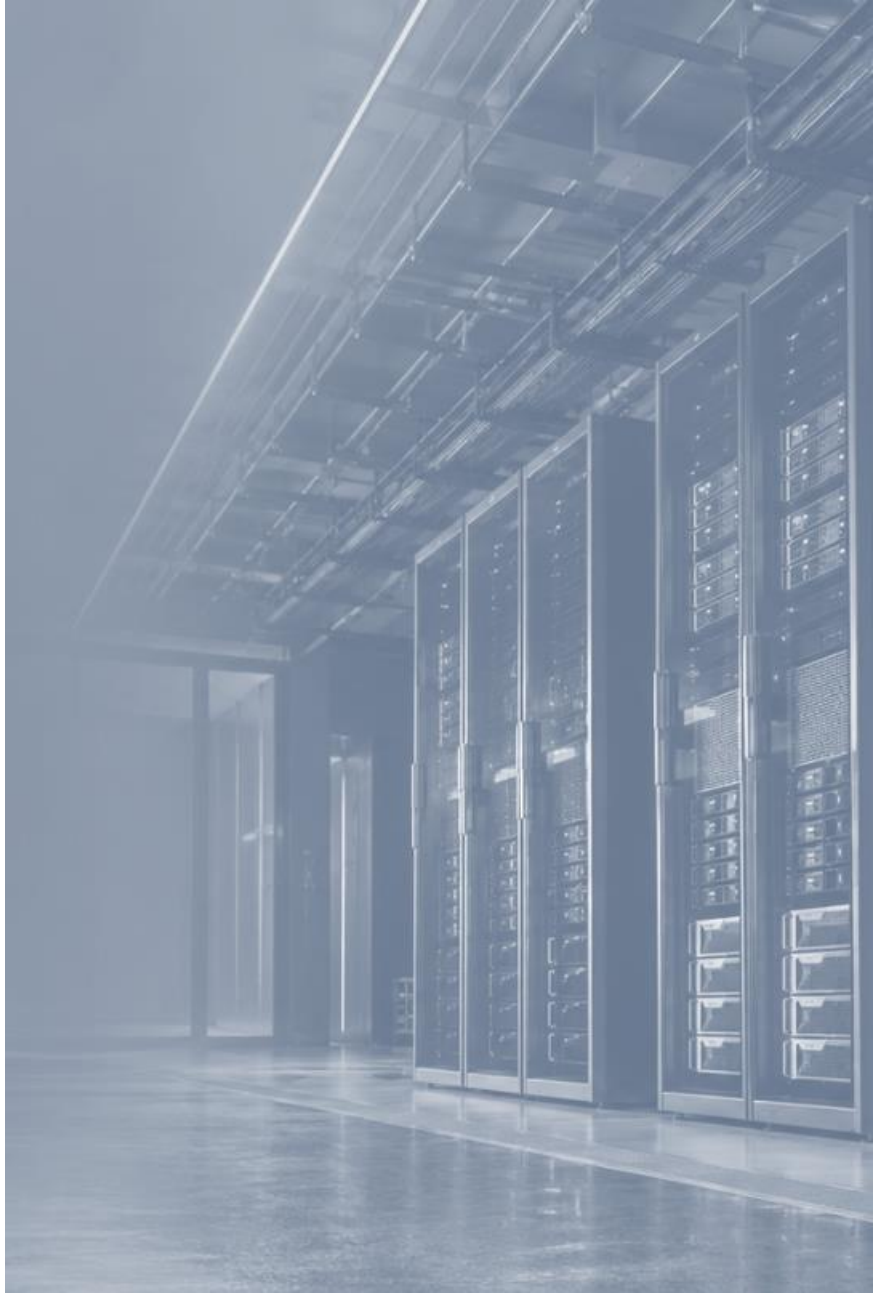
Бизнес-кейс: LLM в промышленной эксплуатации / RAG

Корпоративный ИИ-ассистент с предсказуемой производительностью

Задачи:

- поиск по регламентам и базе знаний
- обработка обращений и документов
- генерация ответов, саммари и аналитики
- дообучение / адаптация под отраслевые данные

В промышленной эксплуатации важна не только GPU-карта: нужны CPU, RAM, хранение, сеть, безопасность, мониторинг и поддержка.



RTX PRO 6000 Blackwell

PCIe GPU с большим объёмом памяти для инференса, визуального ИИ и графических нагрузок

Когда выбирать этот класс

ИИ-инференс	LLM/RAG, ИИ-агенты, пакетный инференс и многопользовательские сервисы, когда нужна большая видеопамять на один GPU.
ИИ для изображений и видео / медиа	генерация и обработка изображений и видео, компьютерное зрение, медиапайплайны.
3D / цифровые двойники	рендеринг, визуализация, виртуальные сцены, инженерные модели и графические пайплайны.
Граница применения	Не подменяет HGX/HBM-класс для обучения крупных моделей; сильная сторона — инференс и гибрид ИИ + графика.

Исполнение у нас

Публичное облако, выделенные серверы

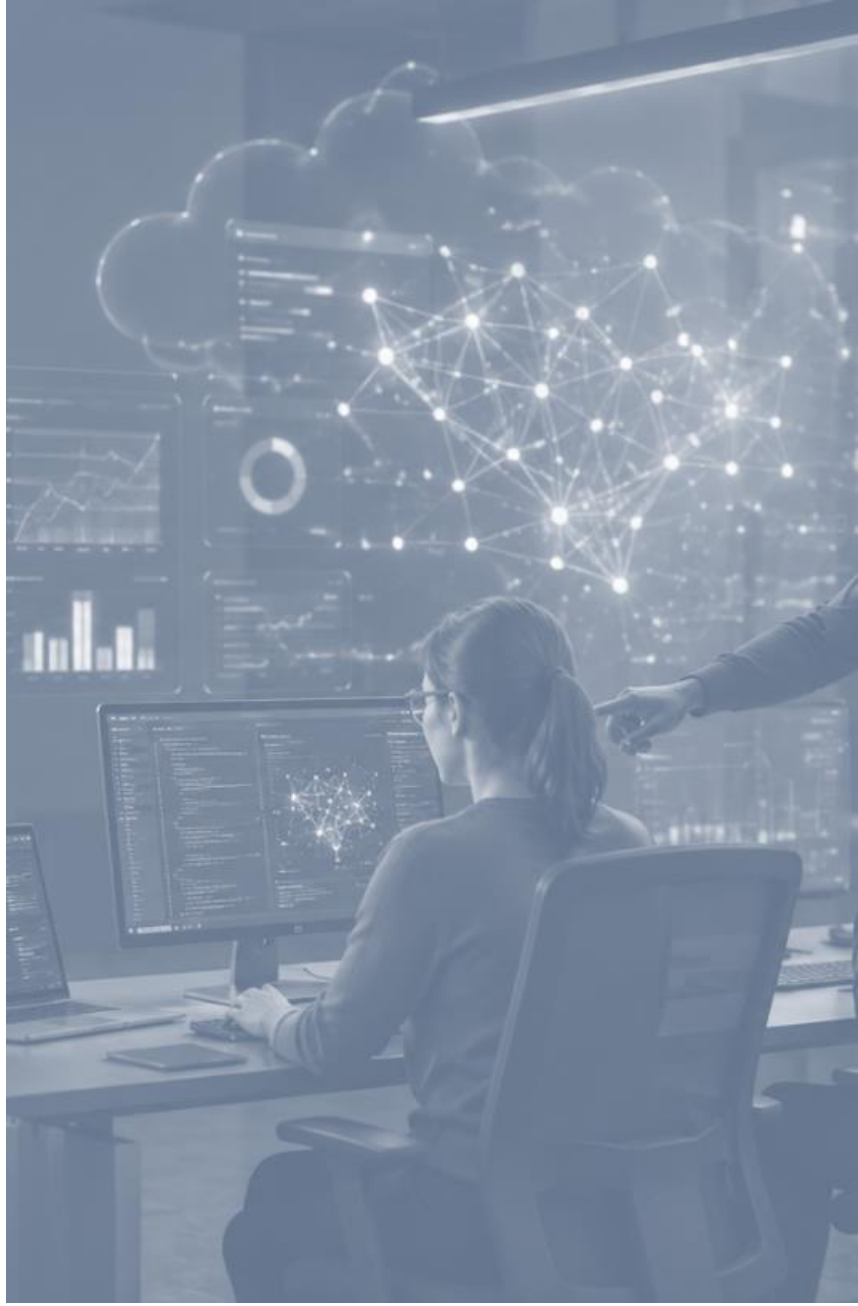
Инфраструктура. Надёжная. Защищённая.

softline® ОБЛАКО

Бизнес-кейс: ИИ-сервис с высокой нагрузкой на инференс

Стоимость токена, задержка и параллельные пользовательские запросы

- чат-боты, ассистенты и ИИ-агенты
- суммаризация и генерация контента
- обработка изображений и видео
- пакетный инференс и внутренние API



Когда RTX PRO 6000 уместнее

модель помещается в 96 GB, нагрузка в основном - инференс, RAG и адаптационное дообучение модели, а H100/H200 экономически избыточны

Профессиональные GPU: RTX 6000 Ada/ A40 / A5000 / A30 / L4

Проф. графика, VDI, рендеринг, видео и средний класс машинного обучения

RTX 6000 Ada	A40	A5000	A30	L4
48 GB: визуализация, машинное обучение, рендеринг	48 GB: визуальные вычисления, vGPU	24 GB: 3D, машинное обучение-разработка, рендеринг	24 GB HBM2: инференс, компьютерное зрение	24 GB: видео, инференс, VDI

Этот слой закрывает не «тяжёлое обучение LLM», а смешанные задачи: графика + вычисления + видео + удалённые рабочие места.

Доступность у нас → Публичное облако, выделенные серверы

Бизнес-кейс: инженерные команды и 3D-производство

GPU-рабочие места и рендеринг без закупки парка рабочих станций

- VDI и удалённые рабочие станции
- CAD / BIM / 3D-модели
- рендеринг и визуализация
- транскодинг и обработка видео

Массовый и legacy-слой GPU

RTX 5090 / 4090 / 3090 / 3080, T4, A2, A2000, 2080 Ti, GTX 1080

RTX 50/40/30	PoC, разработка, инференс, небольшие модели, рендеринг, пакетные задачи с хорошей экономикой
T4 / A2	энергоэффективный инференс, видеоаналитика, транскодинг, и начальный уровень ИИ
A2000 / 2080 Ti / GTX 1080	экономичный dev/test, графические задачи, совместимость со старым стеком и нерегулярные нагрузки

Профессиональный позиционирование

это не класс для тяжёлого обучения LLM; его ценность — быстрый старт, доступность, экономичность и широкий пул задач

Доступность у нас → Публичное облако, выделенные серверы

Бизнес-кейс: быстрый пилот ИИ / dev-test

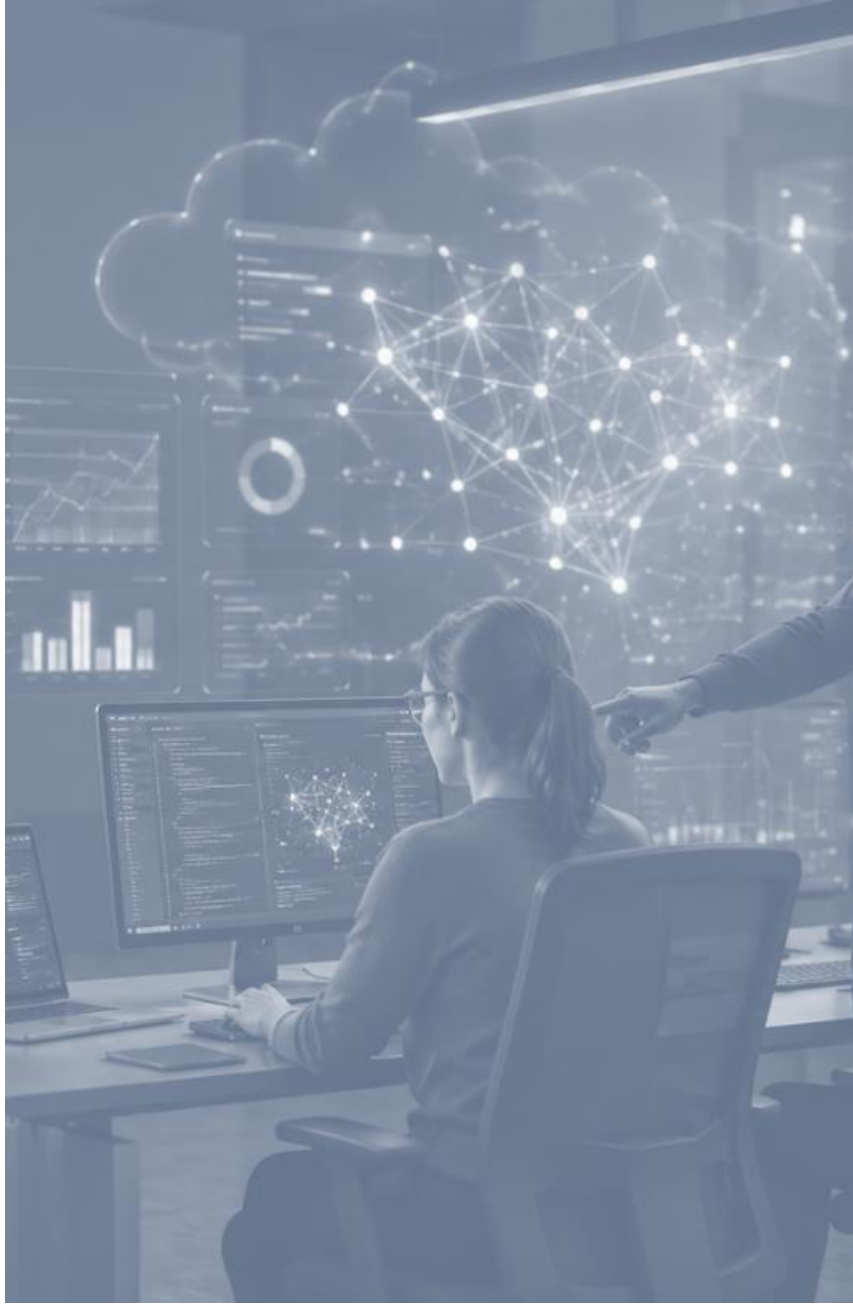
Проверить гипотезу до закупки или масштабирования

Задачи:

- запустить модель за дни, а не месяцы
- проверить качество и требования к памяти
- оценить стоимость инференса и хранения
- понять, когда переходить на A100 / RTX PRO 6000 / H200

Принцип выбора

GPU выбирается по задаче, размеру модели, объёму данных, SLA, требованиям к безопасности и размещению



Форматы в коммерческой части:
публичное облако с GPU / сервер с GPU в ЦОД / сервер с GPU на вашей площадке

Готовые образы для GPU-серверов

CUDA, драйверы, Docker/NGC и прикладной софт уже предустановлены

Быстрый старт	развернуть GPU-сервер с готовым образом вместо ручной установки ОС, драйверов, CUDA и фреймворков
GPU / AI-стек	Linux-образы с CUDA 13.2, NVIDIA-драйверы, Docker/NGC, OpenWebUI, LibreChat, OpenCode и AI-агенты
Прикладные сценарии	ComfyUI / Automatic1111 / Fooocus, Blender, 3D-пакеты, DaVinci Resolve, Adobe-видео/фото, Hashcat

Что получает клиент

- быстрее PoC и запуск среды разработки
- меньше ошибок при настройке CUDA/драйверов
- повторяемая конфигурация для команд и проектов

Доступность у нас → Для части конфигураций в публичном облаке

Готовые модели: попробовать до развёртывания

Публичные эндпоинты и API для теста LLM, мультимодальных и диффузионных моделей

Как использовать

- проверить качество модели до подбора GPU-конфигурации
- сравнить модели под русский язык, кодинг, агентские задачи и длинный контекст
- после теста перейти к постоянному контуру: GPU-сервер, свои веса, дообучение или private API

Примеры в каталоге

gemma-4-26B-A4B-it

reasoning, мультимодальность, русский язык, кодинг

Qwen3.5-35B-A3B

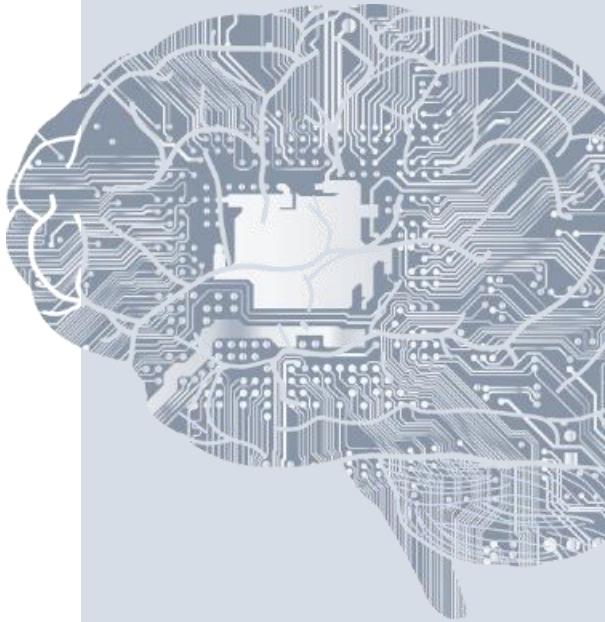
баланс качества и ресурсов, длинный контекст

Qwen3-Coder-Next

coding-агенты и длинный контекст 262K

NVIDIA Nemotron-3 Nano

агентские системы, контекст до 1M токенов



Каталог содержит свободно распространяемые LLM, мультимодальные и диффузионные модели; часть моделей можно попробовать через публичный эндпоинт

Вопрос/Ответ

CLOUD.SOFTLINE.RU

CLOUD@SOFTLINE.COM

8 495 232 0023

Инфраструктура. Надёжная. Защищённая.

softline® ОБЛАКО