Локальный ИИ без дата-центра

Практический гид по компактной локальной инфраструктуре для российского МСБ

От первого пилота до защищённого контура

Алексей Торицын — независимый эксперт

Версия 1.0 · 13 июля 2026 года · www.toritsyn.com



Рисунок. Компактная инфраструктура как основа локального ИИ

О документе

Этот отчёт отвечает на практический вопрос: можно ли развернуть полезный корпоративный ИИ без серверной комнаты, сложного проекта на десятки миллионов рублей и постоянной зависимости от внешнего облака. Короткий ответ — да. Но только если начинать не с названия устройства, а с бизнес-задачи, профиля нагрузки и требований к данным.

Исследование объединяет и заново проверяет два независимых обзора рынка. Цены, характеристики, программная совместимость, нормативные утверждения и кейсы перепроверены по актуальным источникам. Там, где рынок не даёт сопоставимых нагрузочных тестов, приведены диапазоны для планирования, а не ложные точные цифры.

Отчёт не является юридическим заключением и не заменяет нагрузочный пилот. Цены зафиксированы на 13 июля 2026 года и могут быстро меняться из-за курса, доступности памяти, логистики и условий параллельного импорта.

Как читать отчёт

- **Часть I** предназначена для владельцев бизнеса и руководителей высшего звена. Здесь — выводы, варианты закупки, экономика и решение «что делать сейчас».
- **Часть II** предназначена для ИТ-директора и архитектора. Здесь — технические ограничения, стеки, безопасность, масштабирование и методика расчёта нагрузки.
- **Приложения** содержат расширенный список, допущения и проверяемые источники.

Содержание

1. Решение в двух страницах
2. Что изменилось на рынке
3. Отобранные решения и победители по сценариям
4. Пять закупочных пакетов
5. Экономика и совокупная стоимость владения
6. Отраслевые сценарии
7. План пилота 30/60/90
8. Методика и ограничения
9. Технические карточки платформ
10. Модели, память и производительность
11. Пользователи и нагрузка
12. Технологические стеки
13. Изолированный и гибридный контуры
14. Масштабирование на 1–2 года
15. Безопасность и регулирование
16. Реальные кейсы
17. Риски и закупочный контрольный список
18. Расширенный список и основания отбора
19. Сценарии нагрузки по размеру компании
20. Протокол нагрузочного пилота
21. Референсные технологические стеки
22. Модель совокупной стоимости владения и чувствительность
23. Контроль качества и приёмка
24. Глоссарий
25. Источники

Часть I. Решение для бизнеса

Выводы, варианты закупки, экономика и 90-дневный план пилота



1. Решение в двух страницах

Компактная локальная инфраструктура в 2026 году перестала быть экспериментом энтузиастов. Устройство размером с книгу уже может обслуживать корпоративный поиск по документам, ассистентов сотрудников, распознавание текста, распознавание речи и ограниченные мультиагентные процессы. Системы со 128 ГБ общей памяти способны загрузить модель на 70 млрд. параметров в 4–5-битном формате. Однако «модель запустилась» ещё не означает «сервис готов для компании».

Главная развилка проходит между тремя ресурсами:

1. **Память** определяет, какая модель и какой контекст вообще поместятся.
2. **Пропускная способность памяти и зрелость среды выполнения** определяют скорость ответа.
3. **Эксплуатация** определяет, будет ли система стабильно работать через полгода: резервные копии, обновления, права доступа, мониторинг и ответственность за инциденты.

ГЛАВНЫЙ ВЫВОД

Для большинства компаний лучший первый шаг — 90-дневный пилот на Mac mini M4 Pro с 64 ГБ объединённой памяти и накопителем 1 ТБ. В российской рознице такая конфигурация обычно стоит 250–300 тыс. руб. Она сохраняет полезный запас для одновременной работы Docker, Ollama, n8n, LocalAI и LangChain. Система AMD со 128 ГБ оправданна, когда пилот подтвердил потребность в модели на 70 млрд. параметров; высокая параллельность требует станции с CUDA, а регулируемый производственный контур — полноценного серверного проекта.

Пять выводов, которые влияют на инвестиционное решение

Первое. Практичный бюджет первого пилота составляет 250–300 тыс. руб., но рабочая система стоит дороже самого компьютера. К устройству добавляются накопитель или сетевое хранилище, ИБП, резервное копирование, внедрение и время администратора. Трёхлетняя совокупная стоимость владения пилотом обычно находится в диапазоне 0,65–1,05 млн. руб.

Второе. 128 ГБ общей памяти — это не эквивалент 128 ГБ профессиональной видеопамяти. AMD Ryzen AI Max+ 395 и NVIDIA GB10 позволяют разместить большую модель, но их память имеет пропускную способность около 256–273 ГБ/с. RTX 5090 и RTX PRO 6000 достигают 1 792 ГБ/с, поэтому намного быстрее обслуживают небольшие и средние модели, помещающиеся в видеопамяти. [1][3][6][10]

Третье. Mac mini M4 Pro с 64 ГБ памяти и накопителем 1 ТБ — наиболее сбалансированный вариант для первого пилота. Он работает тихо, потребляет мало энергии и даёт зрелую среду MLX/Metal. Конфигурация оставляет запас для Docker, Ollama, n8n, LocalAI и LangChain, поэтому пилот не упирается в память сразу после подключения поиска, базы данных и автоматизации. Ограничения остаются: память и внутренний накопитель нельзя модернизировать, CUDA не поддерживается, а модель на 70 млрд. параметров не имеет достаточного запаса для промышленной эксплуатации. Конкретные предложения и наличие меняются; на дату отчёта рабочий ориентир российского рынка — 250–300 тыс. руб., хотя отдельные продавцы выставляют цену выше. [1][2][30][31]

Четвёртое. Мультиагентность резко увеличивает нагрузку. Один запрос пользователя может породить 3–10 обращений к модели, поиск, повторное ранжирование и вызовы инструментов. Устройство, комфортно обслуживающее десять простых чат-сессий, может поддерживать лишь две-три активные многошаговые агентные задачи.

Пятое. Локальное размещение не равно соответствию 152-ФЗ. Оно помогает выполнить требование локализации и снижает риск передачи данных наружу, но не заменяет правовые основания обработки, политику, разграничение доступа, журналирование, защиту информационной системы и процессы реагирования. [17][27][28]

2. Что изменилось на рынке

Рынок разделился на пять практических классов.

Класс	Ориентир цены в РФ	Что покупает бизнес	Основное ограничение
Компактный пилот	180–300 тыс. руб.	Mac mini M4 Pro, младшие AMD-конфигурации	Ограниченная параллельная нагрузка
Большая общая память	230–450 тыс. руб.	Framework Desktop, Minisforum MS-S1 Max и аналоги на Ryzen AI Max+ 395	Более сложный Linux/ROCm-контур, сервис
Персональный ИИ-суперкомпьютер	500–950 тыс. руб.	DGX Spark, ASUS Ascent GX10, Dell Pro Max GB10	Цена и совместимость с архитектурой Arm отдельных компонентов
Высокая скорость на платформе CUDA	580 тыс.–2,0 млн. руб.	Рабочая станция с RTX 5090 или RTX PRO 6000	Энергия, шум, цена и ограниченный объём видеопамяти
Российский серверный контур	от 1,5 млн. руб., по запросу	YADRO/Rikor/Aquarius в актуальной конфигурации с графическим ускорителем	Проектная закупка, не «коробка»

У этого сегмента нет единой корректной статистики объёма: аналитики смешивают периферийный ИИ, рабочие станции, компьютеры для ИИ и серверы. Поэтому в отчёте не используется эффектная, но плохо применимая цифра «размер рынка». Для заказчика важнее три наблюдаемых факта: появление 128-ГБ настольных систем AMD, выход компактной платформы NVIDIA GB10 и резкий рост цен на память и профессиональные графические ускорители в 2026 году. [3][4][6][7][11]

Российский рынок живёт по другой логике. Apple, NVIDIA GB10 и значительная часть графических ускорителей поступают через параллельный импорт; цена зависит не только от рекомендованной цены производителя и курса, но и от наличия, доставки, гарантии и риска канала. Одновременно российские серверные производители дают сервис и возможность собрать доверенный контур, однако публичные цены и точные конфигурации графических ускорителей обычно отсутствуют.

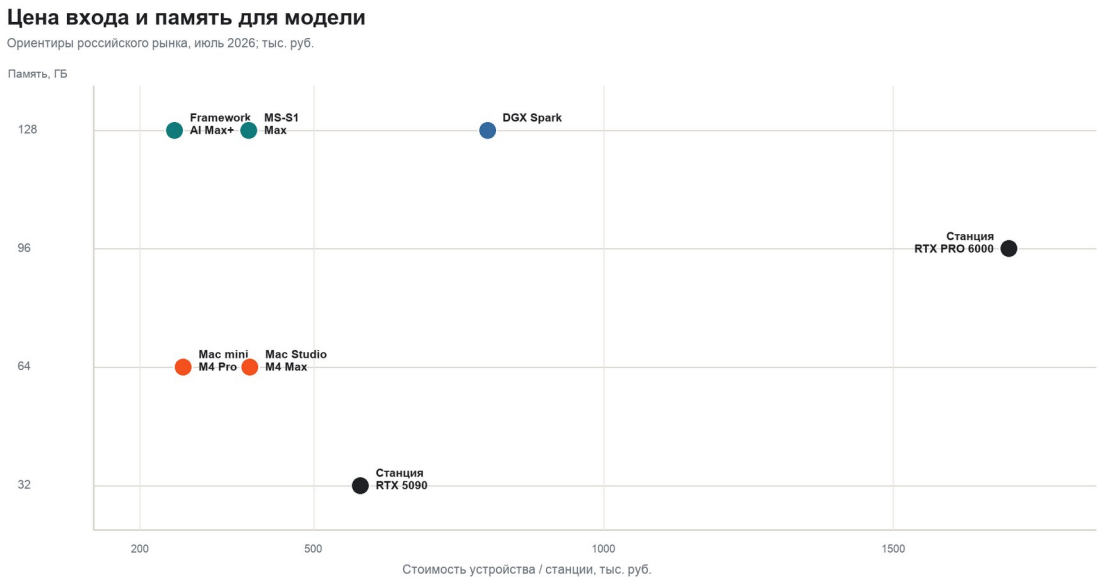


Рисунок. Цена и доступная память: разные способы купить запас под модель

3. Отобранные решения и победители по сценариям

Отобранные решения

Решение	Рабочая конфигурация	Цена в РФ / оценка	Для чего подходит лучше всего	Ключевое ограничение
Mac mini M4 Pro	64 ГБ / 1 ТБ	250–300 тыс. руб.	Первый пилот, поиск по корпоративным документам, ассистенты	Память не расширяется, CUDA не поддерживается
Framework Desktop AI Max+	128 ГБ / 2 ТБ	220–300 тыс. руб. при импорте	Большой объём памяти за умеренную цену, эксперименты с моделями на 70 млрд. параметров	самостоятельная сборка, слабая локальная поддержка
Minisforum MS-S1 Max	128 ГБ / 2 ТБ	около 388 тыс. руб.; наличие меняется	Модели на 70 млрд. параметров, несколько сервисов, два порта 10 Гбит/с	Память распаяна, Linux требует компетенции
Mac Studio M4 Max	64 ГБ / 1 ТБ	350–430 тыс. руб.	Тихий быстрый Mac-узел, обработка текста, изображений и медиаконтента	Ограниченная текущая память, нет CUDA
DGX Spark / GB10	128 ГБ / 4 ТБ	500–950 тыс. руб.	CUDA-совместимая разработка, дообучение, агенты	архитектура Arm, высокая российская наценка
Рабочая станция с RTX 5090	32 ГБ видеопамяти / 96 ГБ оперативной памяти	от 581 тыс. руб.	Высокая скорость моделей на 8–30 млрд. параметров, обработка изображений, речи и видео	Модель на 70 млрд. параметров не помещается целиком; ускоритель потребляет до 575 Вт
RTX PRO 6000 Blackwell	96 ГБ видеопамяти	1,4–2,0 млн. руб. за станцию	Промышленное формирование ответов, модели на 70 млрд. параметров, высокая параллельность, коррекция ошибок памяти и аппаратная изоляция ресурсов	Цена и энергопотребление ускорителя 300–600 Вт
Российский сервер с графическим ускорителем	Конфигурация по ТЗ	от 1,5–3,0 млн. руб.	Регулируемый контур, сервис, резервирование	Нужен проект и проверка конкретного состава

Цены — ориентиры среза, а не оферта. Например, российская карточка MS-S1 Max показывала 387 500 руб. за 128 ГБ / 2 ТБ без наличия и 331 399 руб. за доступную 64-ГБ версию; готовая станция с RTX 5090, 96 ГБ оперативной памяти и накопителем 2 ТБ предлагалась за 581 200 руб. DGX Spark стоил у NVIDIA 4 699 долл. США, а российская импортная карточка доходила до 939 590 руб. [5][7][12][13]

Победители без попытки назвать одного абсолютного лидера

- **Лучший первый пилот:** Mac mini M4 Pro, 64 ГБ / 1 ТБ, с бюджетом 250–300 тыс. руб.
- **Лучший объём памяти за деньги:** Framework Desktop 128 ГБ; при приоритете российского канала — Minisforum MS-S1 Max 128 ГБ.
- **Лучший доступ к CUDA в компактном формате:** DGX Spark или другой GB10-узел.
- **Лучшая скорость для небольших и средних моделей:** рабочая станция с RTX 5090.
- **Лучший одноузловой вариант для промышленной эксплуатации:** RTX PRO 6000 Blackwell 96 ГБ.
- **Лучший путь для регулируемого контура:** актуальный российский сервер с согласованным графическим ускорителем, ОС и средствами защиты — после проверки сертификатов всего решения.

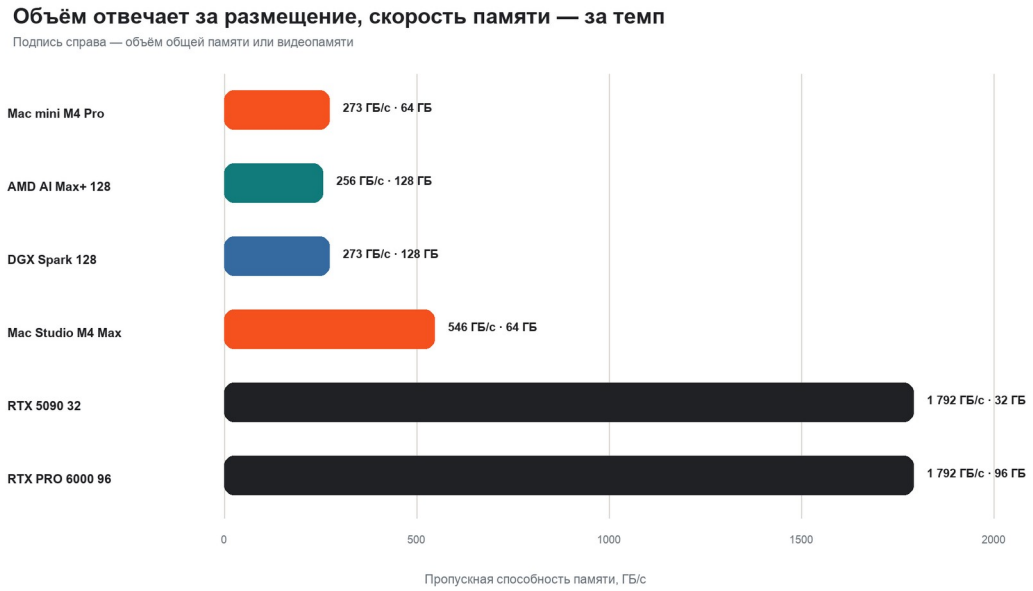


Рисунок. Память определяет размер модели, пропускная способность — темп ответа

4. Пять закупочных пакетов

Пакет 1. «Пилот без серверной»

Состав: Mac mini M4 Pro 64 ГБ / 1 ТБ, внешний твердотельный накопитель 2 ТБ или небольшое сетевое хранилище, ИБП, резервный диск.

Первоначальные затраты: 250–300 тыс. руб.

Нагрузка: 10–30 активных пользователей за день, обычно 2–4 параллельные сессии на 8–14 млрд. параметров или 1–2 на 20–32 млрд. параметров.

Задачи: поиск по корпоративным документам с формированием ответа, внутренний ассистент, краткие резюме, черновики, классификация и расшифровка записей.

Не подходит: модель на 70 млрд. параметров в промышленной эксплуатации, высокая параллельная нагрузка, сертифицированный защищённый контур.

Пакет 2. «Большая модель в офисе»

Состав: Minisforum MS-S1 Max либо Framework Desktop с 128 ГБ, твердотельный накопитель 2–4 ТБ, сетевое хранилище и резервное копирование, ИБП.

Первоначальные затраты: 320–480 тыс. руб.

Нагрузка: 25–60 активных пользователей за день; 4–8 параллельных быстрых запросов, 2–4 запроса средней тяжести, 1–2 запроса к модели на 70 млрд. параметров.

Задачи: качественный поиск по корпоративным документам с генерацией ответа, анализ договоров, несколько локальных моделей, агенты с очередью задач.

Не подходит: строгий целевой уровень обслуживания без резервного узла; эксплуатация без Linux-компетенции.

Пакет 3. «Компактный CUDA-контур»

Состав: DGX Spark или ASUS GX10, сеть 10 Гбит/с, сетевое хранилище, ИБП.

Первоначальные затраты в России: 600 тыс.–1,05 млн. руб. в зависимости от канала.

Нагрузка: 40–100 активных пользователей за день; 6–12 быстрых, 3–6 средней тяжести или 1–3 тяжёлых параллельных сессии как плановый диапазон.

Задачи: разработка и перенос стека NVIDIA, NIM и TensorRT-LLM, экономичная адаптация моделей, обработка текста и изображений, мультиагентные сервисы.

Не подходит: закупка только ради рекламного «1 квадриллион операций с плавающей запятой в секунду» без проверки целевой модели.

Пакет 4. «Высокая параллельность»

Состав: рабочая станция под Linux с RTX 5090 и 32 ГБ видеопамяти, 96–128 ГБ оперативной памяти, накопитель 2–4 ТБ, сеть 10 Гбит/с, ИБП, отдельное сетевое хранилище.

Первоначальные затраты: 650–850 тыс. руб.

Нагрузка: 50–150 активных пользователей за день; 12–25 параллельных сессий на 7–14 млрд. параметров либо 4–8 на 20–30 млрд. параметров после настройки пакетной обработки запросов.

Задачи: клиентская поддержка, код, распознавание текста и изображений, распознавание и синтез речи, большие очереди документов.

Не подходит: размещение модели на 70 млрд. параметров целиком в 32 ГБ видеопамяти; тихое офисное размещение без подготовки.

Пакет 5. «Регулируемая промышленная эксплуатация»

Состав: российский сервер высотой 1–2 монтажных единицы или сервер в напольном корпусе, профессиональный графический ускоритель с 48–96 ГБ видеопамяти, оперативная память с коррекцией ошибок, дисковый массив, резервные блоки питания, доверенная ОС, совместимые средства защиты, отдельное резервное хранилище.

Первоначальные затраты: обычно от 1,5–3 млн. руб.; точная цена появляется только после ТЗ.

Нагрузка: определяется выбранным графическим ускорителем и целевым уровнем обслуживания, а не брендом шасси.

Задачи: финансы, ПДн, критичные процессы, централизованный доступ, аудит.

Важно: запись в реестре российского оборудования не сертифицирует автоматически всю ИИ-систему.

5. Экономика и совокупная стоимость владения

Цена устройства — плохой показатель стоимости проекта. На горизонте трёх лет доминируют внедрение и эксплуатация. Электроэнергия для мини-компьютера почти незаметна, а для станции с графическим ускорителем становится видимой, но всё равно обычно уступает стоимости квалифицированного специалиста.

Пакет	Оборудование и инфраструктура	Внедрение	Эксплуатация и поддержка	Электроэнергия	Совокупная стоимость владения за 3 года
Пилот, Mac mini	255 тыс. руб.	180 тыс. руб.	410 тыс. руб.	8 тыс. руб.	0,65–1,05 млн. руб.
Офис, MS-S1 Max	470 тыс. руб.	240 тыс. руб.	738 тыс. руб.	18 тыс. руб.	1,10–1,80 млн. руб.
CUDA, DGX Spark	1 020 тыс. руб.	240 тыс. руб.	900 тыс. руб.	25 тыс. руб.	1,70–2,80 млн. руб.
Высокая скорость, RTX 5090	681 тыс. руб.	300 тыс. руб.	1 060 тыс. руб.	95 тыс. руб.	1,70–2,70 млн. руб.

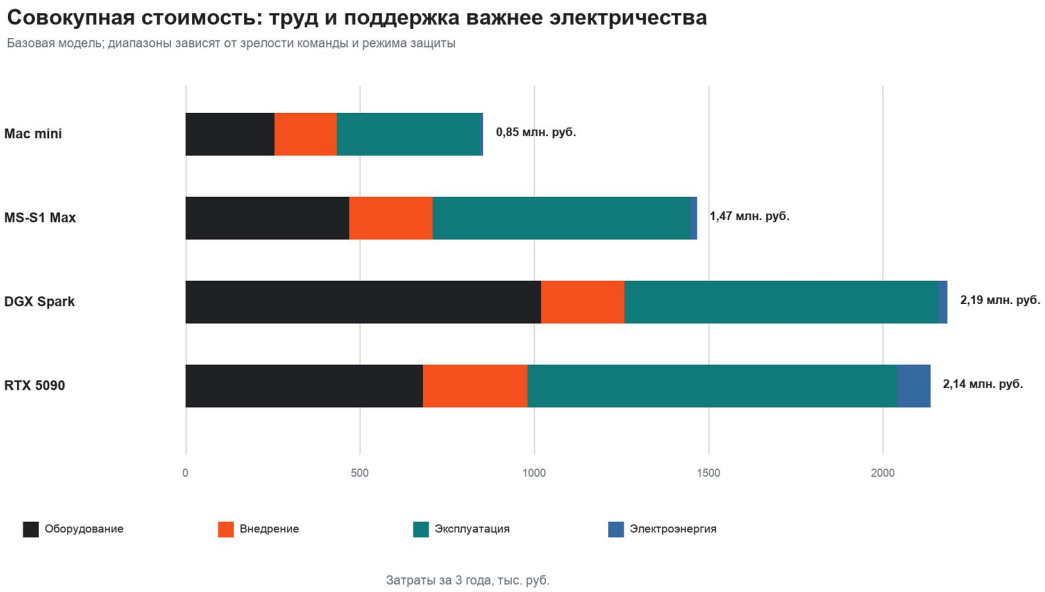


Рисунок. В трёхлетней совокупной стоимости владения труд и поддержка часто дороже электричества

Базовая модель использует тариф 9 руб./кВт·ч, 250 рабочих дней, отдельный период фоновой обработки и ставку специалиста 3 000 руб./час. Диапазон отражает разную зрелость команды. Если инфраструктуру ведёт уже существующий сотрудник и процессы отлажены, нижняя граница достижима. Если интеграцию выполняет подрядчик, нужен полностью изолированный контур и нет внутренней компетенции, реалистична верхняя.

Локальный контур не всегда дешевле облака. Он выигрывает, когда данные нельзя передавать наружу, запросы стабильны и велики, оборудование загружено, а команда умеет его обслуживать. При редких запросах, высокой потребности в самых сильных моделях и отсутствии чувствительных данных облако экономичнее. Гибрид позволяет локально выполнять массовые и чувствительные задачи, а редкие сложные запросы отправлять во внешний сервис после классификации и удаления персональных данных.

6. Отраслевые сценарии

Строительство

Самый быстрый возврат дают поиск по проектной документации, сравнение версий, извлечение требований из договоров и нормативов, подготовка ответов на замечания и распознавание текста актов. Мультимодальная модель может помогать с фотографиями и схемами, но она не должна самостоятельно принимать решения о безопасности или соответствии проекта. Для пилота подходит Mac mini M4 Pro 64 ГБ / 1 ТБ; для массового распознавания и работы с информационными моделями зданий нужен узел NVIDIA.

Пищевая промышленность

Полезны поиск по корпоративным документам с генерацией ответа по ХАССП, техническим регламентам и СТО, проверка сопроводительных документов, расшифровка сменных журналов и визуальный контроль упаковки. Реальное время на линии — отдельная задача на периферии: её нельзя смешивать с офисным чат-ботом на одном узле. Для линии нужны камеры, детерминированный конвейер компьютерного зрения и резервирование.

Логистика

Приоритеты: международные транспортные накладные, ТТН и обычные накладные, поиск расхождений, краткое изложение переписки, помощь диспетчеру, объяснение причин задержки. Мультиагентная маршрутизация должна работать в режиме рекомендации; изменения в системе управления перевозками и финансовые действия требуют подтверждения человеком.

Ритейл

Наиболее зрелые задачи — карточки товаров, анализ обратной связи, база знаний для персонала, классификация обращений и периферийная аналитика магазина. Для сети выгоднее разделить центральную языковую модель с поиском по документам и локальные узлы компьютерного зрения, чем ставить большую модель в каждом магазине.

Финансы

Начинать стоит с внутреннего поиска по нормативным документам, черновиков, проверок соответствия требованиям и ассистента первой линии — без автономных решений по кредитам, платежам и блокировкам. Рекомендации Банка России от 16 июня 2026 года требуют собственной модели угроз и политики ИБ для ИИ; в критичных процессах регулятор рекомендует подтверждение человеком. [18]

7. План пилота 30/60/90

Дни 1–30: доказать полезность

- выбрать один процесс и владельца результата;
- классифицировать данные и определить, что нельзя передавать в облако;
- развернуть одну модель, поиск по корпоративным документам с генерацией ответа, интерфейс и журналирование;
- загрузить 100–500 типовых документов;
- создать 50–100 проверочных вопросов с эталонными ответами;
- подключить 5–10 пилотных пользователей.

Критерии успеха: не «токены в секунду», а доля полезных ответов, время сотрудника на задачу, число ошибок, стоимость запроса и отсутствие утечек.

Дни 31–60: проверить реальную нагрузку

- добавить модель повторного ранжирования и права на документы;
- измерить время до первого токена, скорость генерации, очередь и 95-й процентиль задержки;
- включить 2–3 реальных интеграции только на чтение;
- проверить внедрение вредоносной инструкции, ошибочные права и восстановление из резервной копии;
- сравнить лёгкую и более качественную модели.

Дни 61–90: принять инвестиционное решение

- нагрузочно протестировать пик и усиление нагрузки многошаговыми агентами;
- посчитать фактическую совокупную стоимость владения и экономию времени;
- оформить модель угроз, регламенты и владельца сервиса;
- выбрать: оставить один узел, добавить резервный, перейти на CUDA или остановить проект.

РЕШЕНИЕ О ПРОДОЛЖЕНИИ ИЛИ ОСТАНОВКЕ

Масштабировать стоит только сценарий, который даёт измеримую экономию, проходит тест качества на собственных данных и выдерживает пиковую нагрузку. Покупка «железа впрок» до этого момента увеличивает риск, а не зрелость проекта.

Часть II. Технический отчёт

Платформы, нагрузка, технологические стеки, безопасность и масштабирование



8. Методика и ограничения

В исходных исследованиях были сильные идеи, но также типичные артефакты машинного анализа: одна и та же цифра повторялась без первичного источника, рекламные показатели в триллионах операций в секунду смешивались с токенами в секунду, а старый сервер назывался актуальным решением 2026 года. Эти элементы удалены.

Для спецификаций использованы страницы производителей; для ПО — документация проектов; для российских цен — доступные карточки продавцов; для законодательства — правовые акты и материалы регуляторов. Маркетинговые кейсы отмечены как описание поставщика. Нагрузка приведена диапазонами, потому что корректный результат зависит от модели, квантизации, длины контекста, среды выполнения, пакетной обработки запросов и целевого времени ответа.

Курс доллара на 13 июля 2026 года — около 76,66 руб. по публикации, основанной на данных Банка России. Он используется только как справочная база: российская цена включает логистику, НДС, гарантию и риск поставки. [14]

9. Технические карточки платформ

Apple Mac mini M4 Pro, 64 ГБ / 1 ТБ

Это основной выбор автора для первого пилота. Старшая конфигурация M4 Pro включает 14-ядерный процессор, 20-ядерный графический ускоритель и объединённую память с пропускной способностью 273 ГБ/с. Устройство работает тихо, занимает мало места и поддерживает Thunderbolt 5; при необходимости можно выбрать сетевой порт 10 Гбит/с. [1][2]

На дату отчёта российские продавцы предлагают конфигурацию 64 ГБ / 1 ТБ примерно за 250–300 тыс. руб.; отдельные предложения могут быть дороже. Наличие и происхождение поставки нужно подтверждать перед оплатой. Ориентир подкрепляют карточка iQmac с ценой 289 990 руб. и рыночный срез Price.ru; предложение около 257 890 руб. встречалось без наличия. [30][31]

Главное практическое преимущество 64 ГБ — запас для всей среды пилота, а не только для весов модели. На одном узле можно одновременно разместить Docker, Ollama, n8n, LocalAI, LangChain, базу поиска по документам и средства наблюдения. Для моделей на 8–14 млрд. параметров остаётся комфортный запас; модели на 20–32 млрд. параметров требуют разумной квантизации и ограничения параллельных сессий. Модель на 70 млрд. параметров технически можно запускать в отдельных режимах, но для промышленной эксплуатации запаса недостаточно.

Ограничения принципиальны: память и внутренний накопитель нельзя модернизировать, CUDA не поддерживается, а отказ устройства останавливает все размещённые на нём сервисы. Поэтому резервная копия, внешний накопитель или сетевое хранилище и документированное восстановление обязательны уже на пилоте.

Framework Desktop и Minisforum MS-S1 Max, 128 ГБ

Обе системы используют Ryzen AI Max+ 395 с Radeon 8060S, общей памятью LPDDR5x и пропускной способностью около 256 ГБ/с. Framework Desktop привлекателен ценой и ремонтопригодной конструкцией; MS-S1 Max дороже, но предлагает два сетевых порта 10 Гбит/с, четыре порта USB4 и более серверную компоновку. [3][4]

ROCm 7.2.1 официально поддерживает Ryzen AI Max+ 395 под Ubuntu 24.04. Это заметное улучшение относительно 2025 года, но конкретную модель, формат весов и среду выполнения всё равно нужно проверять до закупки. Для GGUF-моделей наиболее предсказуемым вариантом остаётся llama.cpp через HIP или Vulkan. [15][16]

Модель на 70 млрд. параметров в четырёх- или пятибитном формате помещается с запасом под систему и умеренный кэш контекста. Основное ограничение — скорость генерации и параллельная нагрузка. Два узла разумнее использовать как независимые копии сервиса, а не как попытку объединить память для одной модели.

Mac Studio M4 Max, 64 ГБ

Версия с 16-ядерным процессором и 40-ядерным графическим ускорителем обеспечивает до 546 ГБ/с — примерно вдвое больше M4 Pro. Это заметно ускоряет языковые модели, ограниченные пропускной способностью памяти. Наличие старших конфигураций меняется; объём памяти и срок поставки нужно фиксировать в коммерческом предложении. [9]

Mac Studio оправдан, когда компании нужен тихий узел на macOS, высокая скорость обработки текста и изображений, а 64 ГБ достаточно. Если приоритет — модель на 70 млрд. параметров или стандартный промышленный стек CUDA, система AMD со 128 ГБ либо DGX Spark обычно логичнее.

Практическая конфигурация: 64 ГБ / 1 ТБ, внешний быстрый накопитель для моделей и временных данных, сетевое хранилище для документов и резервных копий, сеть 10 Гбит/с при вынесенном хранилище. PostgreSQL, Qdrant и подготовку документов можно оставить на том же узле во время пилота, но при росте нагрузки их лучше вынести.

Путь роста: второй Mac Studio добавляет пропускную способность и позволяет обновлять один узел без полного простоя. Память двух Mac не становится единой памятью для одной модели. Если 64 ГБ недостаточно, разумнее перейти к системе AMD или GB10 со 128 ГБ либо к профессиональному ускорителю с большой видеопамятью.

NVIDIA DGX Spark и другие системы GB10

GB10 объединяет 20-ядерный процессор Arm, графический ускоритель Blackwell, 128 ГБ общей памяти с пропускной способностью 273 ГБ/с, сеть 10 Гбит/с и интерфейс ConnectX-7 со скоростью 200 Гбит/с. NVIDIA заявляет возможность размещать модели до 200 млрд. параметров на одном устройстве и до 405 млрд. на паре. Это предел вместимости, а не обещание комфортной интерактивной скорости. [6]

Преимущество платформы — CUDA, TensorRT-LLM, NIM, PyTorch и единый путь переноса на серверы NVIDIA. Ограничения — архитектура Arm, теплопакет 140 Вт, высокая цена в России и необходимость проверять сторонние контейнеры. Обновления 2026 года улучшили управление несколькими DGX Spark и позволили отключать Wi-Fi и Bluetooth в UEFI, что полезно для изолированных контуров. [8]

Практическая конфигурация: один DGX Spark с накопителем 4 ТБ, подключением 10 Гбит/с к сетевому хранилищу и отдельным резервным копированием. До закупки следует проверить весь набор зависимостей, включая распознавание текста, разбор документов, драйверы базы данных, агентные компоненты и средства наблюдения. Один компонент, доступный только для x86, способен перечеркнуть преимущество готовой системы.

Путь роста: вторая система GB10 может работать как независимая копия сервиса; ConnectX-7 также позволяет тестировать связанные режимы. Перенос на сервер NVIDIA остаётся отдельным этапом и требует повторных измерений. DGX Spark оправдан, если компания действительно использует экосистему CUDA, а не только стремится разместить максимально большую модель.

Рабочая станция с RTX 5090

32 ГБ памяти GDDR7, пропускная способность 1 792 ГБ/с и энергопотребление до 575 Вт дают высокую скорость для моделей, которые целиком помещаются в видеопамяти. Модели на 7–14 млрд. параметров хорошо подходят для многопользовательской работы; модели на 20–30 млрд. параметров остаются рабочим вариантом. Для модели на 70 млрд. параметров приходится переносить часть весов в оперативную память, из-за чего скорость резко падает. [12][29]

Для бизнеса важнее не сама карта, а качество сборки: охлаждение под длительную нагрузку, 96–128 ГБ оперативной памяти, надёжный блок питания, сеть 10 Гбит/с, гарантия сборщика и резервный узел. Игровой компьютер без круглосуточного нагрузочного теста не следует автоматически считать сервером.

Практическая конфигурация: Linux, 96–128 ГБ оперативной памяти, твердотельный накопитель 2–4 ТБ, блок питания и корпус с запасом по охлаждению, сеть 10 Гбит/с, ИБП и отдельное хранилище. Основная модель на 20–32 млрд. параметров целиком остаётся в видеопамяти; небольшая модель-маршрутизатор может работать на том же ускорителе при строгом контроле памяти. Запросы к модели на 70 млрд. параметров лучше направлять на другой узел или в разрешённое облако.

Путь роста: две станции с RTX 5090 масштабируют число запросов и дают резерв на обновление. Они не образуют единую видеопамять объёмом 64 ГБ для одного процесса. При устойчивой потребности в модели на 70 млрд. параметров следующий шаг — профессиональный ускоритель с 96 ГБ видеопамяти, а не третья игровая карта в офисе.

RTX PRO 6000 Blackwell, 96 ГБ

96 ГБ памяти GDDR7 с коррекцией ошибок, пропускная способность 1 792 ГБ/с, аппаратная изоляция до четырёх вычислительных экземпляров и варианты энергопотребления 300 или 600 Вт делают карту мостом между рабочей станцией и сервером. Она вмещает модель на 70 млрд. параметров с большим запасом и может обслуживать несколько моделей. Официальная цена выросла до 13 250 долл. США, поэтому это уже следующий уровень, а не стартовый вариант. [10][11]

Практическая конфигурация: рабочая станция или серверное шасси, 192–256 ГБ оперативной памяти с коррекцией ошибок, накопитель 4–8 ТБ, сеть 10 или 25 Гбит/с и контроль температуры и питания. Аппаратная изоляция ресурсов не заменяет резервный узел: отказ карты останавливает все размещённые на ней процессы.

Путь роста: второй профессиональный ускоритель или второй узел добавляют после профилирования. Для формирования ответов чаще выгоднее две независимые копии сервиса с балансировщиком; распределение одной модели между ускорителями выбирают лишь тогда, когда она не помещается на одной карте. На этом уровне нужны формальный целевой уровень обслуживания, запасная часть или договор замены и проверка тепловых ограничений шасси.

Российский серверный класс

Актуальный пример — YADRO VEGMAN R220 G3: российский сервер корпоративного класса, допускающий установку полноразмерных графических ускорителей и включённый в реестр. Публичная конфигурация и цена зависят от проекта. Старый Aquarius T60 D19, встречавшийся в одном исходном исследовании, исключён как нерелевантный ориентир 2026 года. [19]

Российский бренд шасси важен для поставки и сервиса, но соответствие требованиям определяется всей системой: конкретным ускорителем, операционной системой, средствами защиты, конфигурацией, моделью угроз и эксплуатационными процессами.

Практическая конфигурация: техническое задание должно фиксировать модель, длину контекста, параллельную нагрузку, тип ускорителя, объём оперативной памяти с коррекцией ошибок, дисковый массив, сеть, резервные блоки питания, операционную систему, средства защиты, допустимую потерю данных, время восстановления и место ремонта. Поставщик предоставляет полный состав поставки и результаты испытаний целевой среды. Цена шасси без ускорителя и цена готового контура несопоставимы.

Путь роста: серверный класс масштабируется через второй узел, балансировку, отдельное хранилище и регламент переключения. Резерв должен быть совместим по драйверам и моделям, а не просто иметь свободный слот расширения. Для регулируемой системы приёмка включает документы на каждую значимую компоненту и проверку восстановления.

10. Модели, память и производительность

Почему нельзя сравнивать только триллионы операций в секунду

Пиковое число операций зависит от точности вычислений, разреженности и типа операции. Скорость языковой модели часто ограничена чтением весов из памяти. Поэтому RTX 5090 с 32 ГБ может быть намного быстрее системы со 128 ГБ общей памяти на модели с 14 млрд. параметров, но не вместить модель на 70 млрд. параметров без переноса части весов в оперативную память.

Класс модели	Примерный объём весов в четырёхбитном формате	Рекомендуемая память узла	Типичная роль
Плотная модель, 7–14 млрд. параметров	5–10 ГБ	16–32 ГБ	быстрый чат, классификация, инструменты
Плотная или экспертная модель, 20–32 млрд. параметров	14–24 ГБ	32–64 ГБ	основной корпоративный ассистент
70–80 млрд. параметров	40–55 ГБ	64–128 ГБ	сложный анализ, низкая параллельность
100–200 млрд. параметров или крупная экспертная модель	65–120+ ГБ	128 ГБ и более	лаборатория, специализированные задачи

К весам добавляются среда выполнения, кэш, контекст и параллельные последовательности. Конфигурация «ровно по размеру файла» непригодна для промышленной эксплуатации.

Практическая матрица

Платформа	8–14 млрд. параметров	20–32 млрд. параметров	70–80 млрд. параметров	Дообучение
Mac mini 64	отлично	хорошо, 1–3 сессии	не рекомендуется	адаптация небольших моделей
AMD 128	отлично	хорошо	работает при низкой параллельности	адаптация модели с оговорками
Mac Studio 64	отлично	отлично	возможно на грани	адаптация через MLX
DGX Spark 128	отлично	отлично	хорошо после оптимизации	сильная сторона
RTX 5090 32	лучшая скорость	отлично	только с переносом части весов	адаптация в пределах видеопамяти
RTX PRO 6000 96	промышленная эксплуатация	промышленная эксплуатация	промышленная эксплуатация	полноценный рабочий вариант

11. Пользователи и нагрузка

Фраза «поддерживает 50 пользователей» ничего не говорит без описания режима. В отчёте разделены четыре показателя: зарегистрированные пользователи, активные пользователи за день, одновременно активные сессии и одновременно формируемые ответы.

Три профиля

Профиль	Контекст / ответ	Частота	Требование	Подходящий класс
Лёгкий	2 тыс. токенов / ответ 400 токенов	10–20 запросов в день	первый токен менее чем за 3 с, скорость от 15 токенов в секунду	7–14 млрд. параметров
Поиск по документам	6 тыс. токенов / ответ 700 токенов	20–40 запросов в день	первый токен менее чем за 5 с, скорость от 10 токенов в секунду	20–32 млрд. параметров
Многошаговая агентная задача	8–32 тыс. токенов, 3–10 обращений к модели	5–15 задач в день	очередь, аудит, подтверждение сотрудником	20–70 млрд. параметров и инструменты

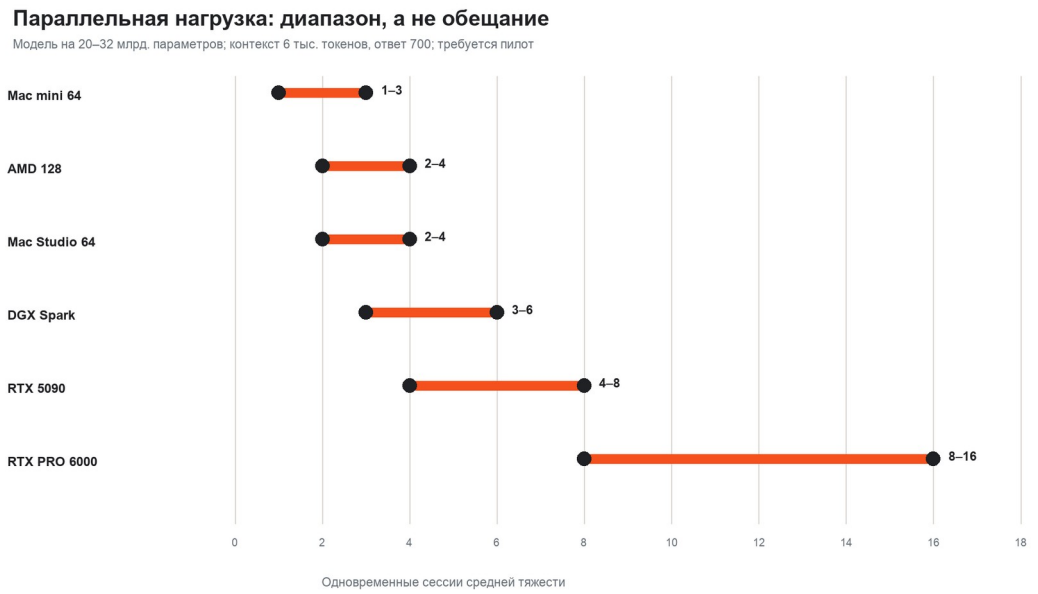


Рисунок. Плановое число одновременных запросов для рабочего профиля; подтвердить пилотом

Для обычной офисной работы активных пользователей за день часто в 5–10 раз больше пиковых генераций. Для агентов коэффициент ниже: одна пользовательская задача занимает модель дольше и создаёт несколько вызовов. Поэтому для Mac mini разумно планировать 10–30 активных пользователей за день, а не обещать «50 пользователей»; для RTX 5090 — 50–150 активных пользователей за день на быстрых моделях после настройки пакетной обработки запросов.

12. Технологические стеки

Минимальный пилот на Apple

macOS → MLX, llama.cpp, Ollama или LocalAI → Open WebUI → Qdrant либо PostgreSQL с pgvector → модуль разбора документов → модель повторного ранжирования → LangGraph или LangChain для ограниченных процессов → Langfuse или OpenTelemetry для трассировки. Docker удобен для сервисов, а n8n — для управляемой автоматизации. MLX нередко быстрее при непосредственном запуске в macOS.

Linux на AMD со 128 ГБ

Ubuntu 24.04 → ROCm 7.2+ → llama.cpp через HIP или Vulkan либо проверенный сервер формирования ответов → Qdrant или PostgreSQL → Open WebUI → LangGraph → Prometheus и Grafana. vLLM следует включать только после проверки целевой модели и встроенного ускорителя: официальная поддержка ROCm не означает одинаковую зрелость всех вычислительных операций.

Промышленный стек NVIDIA

Ubuntu или DGX OS → CUDA → vLLM, SGLang, TensorRT-LLM либо NIM → LiteLLM как шлюз → Qdrant или pgvector → Redis для очереди → LangGraph → OpenTelemetry или Langfuse → Prometheus и Grafana → Keycloak или корпоративная система единого входа.

Что не нужно ставить в первый день

Kubernetes, Milvus и сложная мультиагентная платформа сами по себе не делают решение зрелым. На одном узле достаточно systemd или Docker Compose, PostgreSQL и понятного резервного копирования. k3s или Kubernetes нужны, когда появляются несколько узлов, несколько команд и подтверждённая потребность в обновлениях без простоя и высокой доступности.

13. Изолированный и гибридный контуры



Рисунок. Изолированный контур: данные, модели и журналы остаются внутри периметра

Полностью изолированный контур

Пользователь проходит единый вход и проверку роли. Запрос поступает в шлюз, затем проходит проверку на утечку данных и попадает в корпоративный поиск и локальную модель. Инструменты агента доступны только по списку разрешённых действий; запись или отправка данных требует подтверждения сотрудником. Модели и контейнеры переносятся через контролируемый шлюз, проверяются, подписываются и помещаются во внутренний реестр. Журналы и резервные копии не покидают периметр.

Изолированный контур дороже не из-за оборудования как такового. Дополнительные затраты создают контролируемые обновления, антивирусная проверка, внутренний реестр, управление ключами и замена внешних сервисов локальными аналогами.



Рисунок. Локальный маршрут по умолчанию: чувствительные запросы остаются внутри, разрешённые могут обращаться к облаку

Локальный маршрут с резервным обращением к облаку

До отправки наружу каждый запрос классифицируется. Персональные данные, коммерческая тайна, внутренние документы и идентификаторы всегда обрабатываются локально. Для разрешённых задач данные сокращаются и обезличиваются; маршрутизатор учитывает бюджет, доступность и требуемый класс модели. Ответ внешнего сервиса маркируется, журналируется и проходит итоговую проверку.

Резервный маршрут нельзя строить по принципу «если локальная модель не справилась, отправить всё в облако». Правила должны быть формализованы, передача по умолчанию запрещена, а защита от внедрения вредоносных инструкций — проверена на тестовых атаках.

14. Масштабирование на 1–2 года



Рисунок. Масштабирование: сначала разделить роли, затем добавлять узлы

1. **0–3 месяца:** один узел формирования ответов, отдельная резервная копия и один бизнес-процесс.
2. **3–9 месяцев:** вынести PostgreSQL или Qdrant и подготовку документов на отдельный узел, добавить очередь и мониторинг.
3. **9–18 месяцев:** добавить независимую копию сервиса, балансировку и отдельные модели для быстрых и сложных задач.
4. **18–24 месяца:** при устойчивой нагрузке перейти на профессиональный графический ускоритель или серверный контур.

Группа мини-компьютеров обычно масштабирует число независимых запросов, но не память одной модели. ConnectX-7 у GB10 создаёт специальный канал для пары устройств; обычная сеть 10 Гбит/с или Thunderbolt не превращает несколько Mac или AMD-систем в эквивалент единого ускорителя.

15. Безопасность и регулирование

Что даёт локальный контур

- первичное хранение данных внутри компании;
- контроль сетевого периметра;
- возможность отключить телеметрию и внешние программные интерфейсы;
- собственные журналы и ключи;
- меньшую зависимость от доступности облака.

Чего локальный контур не даёт автоматически

- правового основания обработки персональных данных;
- корректного согласия и опубликованной политики;
- сертификации всей системы;
- защиты от внутренних нарушителей;
- изоляции документов для разных ролей;
- безопасного выполнения действий агентами;
- качества и объяснимости решений.

С 1 июля 2025 года часть 5 статьи 18 Федерального закона № 152-ФЗ запрещает при сборе персональных данных граждан России использовать для записи, систематизации, накопления, хранения, уточнения и извлечения базы за пределами страны, кроме предусмотренных законом случаев. Штраф до 500 млн. руб. относится не к любому нарушению: он применяется к определённым повторным нарушениям и рассчитывается как оборотный в установленных пределах. [17][27]

Для финансовых организаций дополнительно применимы документы Банка России, включая положения 683-П и 757-П для соответствующих типов организаций и рекомендации 3-МР от 16 июня 2026 года по информационной безопасности при разработке и применении ИИ. Требования ФСТЭК и ФСБ зависят от категории информации, статуса системы и использования криптографии; универсального правила «МСБ обязан сертифицировать языковую модель» нет. [18]

Минимальный контрольный список безопасности

- реестр данных и моделей, контрольные суммы и подписи файлов весов;
- единый вход, ролевой доступ, права на уровне документов и отдельные служебные учётные записи;
- запрет действий по умолчанию и подтверждение сотрудником для платежей, изменений и внешних сообщений;
- шифрование дисков, резервных копий и ключей;
- внутренний реестр моделей и контролируемое обновление;
- аудит запросов, найденных фрагментов, ответов и вызовов инструментов;

- защита от внедрения вредоносных инструкций и вывода данных наружу;
- проверка контейнеров и зависимостей;
- проверенное восстановление и порядок действий при инциденте.

16. Реальные кейсы и чему они учат МСБ

Публичных кейсов, где российское МСБ раскрывает точную конфигурацию, модель, параллельную нагрузку и экономический результат, почти нет. Это характерная черта молодого рынка. Поэтому ниже производственные примеры отделены от выводов, которые можно перенести в проекты малого и среднего бизнеса.

Отрасль / организация	Что сделано	Заявленный результат	Урок для МСБ
Строительство, Bouygues Construction	RTX PRO и Omniverse, цифровые двойники и моделирование площадок	интеграция данных в 5 раз быстрее	обработка изображений и трёхмерных сцен требует ускорителя, а не обычного офисного мини-компьютера
Муниципальные услуги, Onetera	локальное формирование ответов на RTX PRO 6000, анализ норм, карт и разрешений	обработка тысяч документов в защищённом контуре	близкий пример для строительной документации
Ритейл, Lowe’s	локальные ускорители, небольшие вычислительные узлы, ассистент и компьютерное зрение	более 1 750 цифровых двойников магазинов	центральную языковую модель и локальное компьютерное зрение лучше разделять
Ритейл, AiFi / Choice Market	локальное компьютерное зрение для магазина без касс	продажи среди пользователей решения выросли на 20%	периферийная аналитика может давать прямой операционный эффект
Пищевая промышленность	компьютерное зрение для контроля упаковочной линии	промышленный контроль качества	обработка изображения в реальном времени требует отдельного детерминированного контура
Финансы, Banco Galicia	локальная платформа OpenShift и контейнерная обработка естественного языка	около 40% экономии операционных затрат	управление и эксплуатационная платформа важнее одной модели
Государственные финансы Франции, ThinkDeep	локальный мультиагентный поиск по документам на NVIDIA DGX	2 дня → 2 минуты; 1 000 часов в месяц для 100 пользователей	агенты дают эффект при контролируемых данных и строгих ограничениях действий

Источники кейсов — страницы компаний и поставщиков. Численные результаты приведены как заявления участников кейса, а не как независимый аудит. [20][21][22][23][24][25][26]

17. Риски и закупочный контрольный список

Главные риски

Слишком большая модель. Качество растёт не линейно, а задержка и стоимость — быстро. Сначала сравните модели на 8–14, 20–32 и 70 млрд. параметров на собственных задачах.

Память без скорости. 128 ГБ позволяют загрузить модель, но не гарантируют комфортную многопользовательскую работу.

Один узел без восстановления. Резервный образ, запасной накопитель и проверенная процедура восстановления часто важнее ещё 10% скорости.

Неподдерживаемый стек. До закупки проверяйте операционную систему, архитектуру Arm или x86, среду выполнения и формат весов.

Параллельный импорт. Фиксируйте серийные номера, срок гарантии, исполнителя ремонта и порядок действий при отказе через год.

Что запросить у поставщика

1. Полный состав поставки: память, накопители, сетевые адаптеры, блок питания и графический ускоритель.
2. Результаты длительного нагрузочного теста и температурный режим.
3. Условия гарантии в России и срок замены.
4. Совместимость с выбранной операционной системой и средой выполнения.
5. Возможность отключить беспроводные интерфейсы и телеметрию.
6. Для российского сервера — реестровую запись и сертификаты именно требуемой конфигурации.
7. Для промышленной эксплуатации — результат теста целевой модели с заданным контекстом и параллельной нагрузкой.

Итоговая рекомендация

Покупать сейчас: Mac mini M4 Pro 64 ГБ / 1 ТБ за 250–300 тыс. руб. для контролируемого пилота. Конфигурация оставляет запас для Docker, Ollama, n8n, LocalAI и LangChain. Систему MS-S1 Max или Framework со 128 ГБ стоит выбирать, если пилот доказал потребность в модели на 70 млрд. параметров.

Тестировать перед покупкой: DGX Spark и рабочую станцию с RTX 5090 — их ценность сильно зависит от целевой модели, среды выполнения и профиля запросов.

Проектировать отдельно: регулируемый серверный контур и систему с RTX PRO 6000.

Избегать: 16-ГБ систем для корпоративного поиска по документам, старых «серверов ИИ» без современного ускорителя, безымянных мини-компьютеров со 128 ГБ без поддержки и покупки нескольких узлов без понятной схемы масштабирования.

Практические приложения

Матрицы выбора, протокол испытаний, TCO и критерии приёмки

18. Расширенный список и основания отбора

В основной список вошли платформы, которые дают достаточный объём памяти, воспроизводимый программный стек, понятный путь закупки и основания ожидать поддержку в течение 1–2 лет. Устройства с привлекательной ценой, но неясной гарантией, охлаждением или совместимостью оставлены в расширенном списке.

Платформа / класс	Статус	Почему включена или исключена	Что проверить перед закупкой
Mac mini M4 Pro 64 ГБ / 1 ТБ	основной выбор для пилота	простая установка, тихая работа, запас для полного программного стека	фактическую конфигурацию, порт 10 Гбит/с, происхождение и гарантию продавца
Mac Studio M4 Max 64 ГБ	основной список	примерно вдвое выше пропускная способность памяти, зрелая среда macOS	доступность памяти, срок поставки, достаточно ли 64 ГБ
Framework Desktop 128 ГБ	основной список с оговоркой	выгодная цена за 128 ГБ и ремонтопригодная конструкция	импорт, блок питания, локальный сервис, образ Linux
Minisforum MS-S1 Max 128 ГБ	основной список	128 ГБ, два порта 10 Гбит/с, компактный корпус	наличие версии 128 ГБ, шум, гарантию в России
DGX Spark	основной список	наиболее цельный компактный стек CUDA	совместимость контейнеров с Arm, российскую цену
ASUS Ascent GX10 / Dell Pro Max GB10	альтернативы	тот же класс GB10 с другим сервисным каналом	местного поставщика, состав поставки, условия поддержки
Рабочая станция с RTX 5090	основной список	высокая скорость в нижнем сегменте CUDA	охлаждение при круглосуточной работе, блок питания, гарантию сборщика
Рабочая станция с RTX PRO 6000	следующий уровень	96 ГБ памяти с коррекцией ошибок и высокая параллельность	энергопотребление конкретной версии, цену и срок поставки
Beelink GTR9 Pro и аналоги AI Max+	расширенный список	интересная платформа, но меньше уверенности в корпоративном сервисе	микропрограмму, память, гарантию, совместимость ROCm
GMKtec EVO-X2 и аналоги	расширенный список	конкурентная цена за 128 ГБ	длительный нагрузочный тест и канал поставки
YADRO VEGMAN R220 G3	проектный класс	российский сервер, сервис и конфигурация по техническому заданию	конкретный ускоритель, операционную систему, средства защиты, реестр и сертификаты
Старые серверы без современного ускорителя	исключены	возраст шасси не компенсирует слабый ускоритель	не покупать по слову «ИИ» в названии
Системы с 16 ГБ	исключены для корпоративного поиска	мало места для модели, контекста и сервисов	допустимы как рабочее место разработчика

Отбор поставщика в России

Коммерческое предложение должно фиксировать не только название устройства. Нужны серийные номера и происхождение, точный объём памяти, накопитель и сетевой адаптер, версия микропрограммы, комплект питания, срок гарантии, адрес ремонта, срок замены и возможность получить резервный узел. Для параллельного импорта отдельно уточняется, кто исполняет гарантию, если продавец прекратит работу.

Цена по карточке сравнима только с ценой такой же готовности. Корпус без памяти и операционной системы нельзя ставить в один ряд с собранной станцией, а импортную цену без доставки и налогов — с российской поставкой с гарантией.

Признаки рискованного предложения

- указана только серия процессора, но нет точного объёма и типа памяти;
- цена относится к корпусу без комплектующих, хотя описание создаёт впечатление готовой станции;
- нет длительного теста с выбранной моделью и средой выполнения;
- гарантия ссылается на производителя, но канал ремонта в России не назван;
- поставщик обещает объединить память нескольких мини-компьютеров без описания распределённого формирования ответов;
- производительность приведена в триллионах операций или токенах в секунду без модели, контекста, длины ответа и параллельной нагрузки;
- срок поставки не отделён от срока замены неисправного узла.

Если два предложения отличаются на 10–15%, но только одно содержит полный состав, протокол испытаний и понятный сервис, управленчески более дешёвым часто оказывается именно оно: простой и повторная настройка быстро съедают первоначальную экономию.

19. Сценарии нагрузки по размеру компании

Микробизнес и небольшая команда: 10–30 сотрудников

Пользователи: 5–15 участников пилота, 5–10 активных пользователей за день, 1–2 одновременных ответа.

Типовой процесс: внутренний поиск, ответы по базе знаний, подготовка документов, расшифровка встреч.

Профиль: модель на 8–14 млрд. параметров, контекст 2–6 тыс. токенов, 100–500 документов, до 2 млн. страниц в год при фоновой обработке с очередью.

Инфраструктура: Mac mini M4 Pro 64 ГБ / 1 ТБ; при стандарте Linux — система AMD с 64 –128 ГБ.

Граница роста: появление двух независимых подразделений или регулярной очереди свыше 10 минут — повод вынести подготовку данных и базу на отдельный узел.

Малая компания: 30–100 сотрудников

Пользователи: 15–40 активных пользователей за день, 2–6 одновременных сессий средней тяжести, 3–10 обращений к модели на многошаговую задачу.

Типовой процесс: поиск по договорам и регламентам, ассистент отдела продаж или закупок, классификация документов, несколько интеграций только на чтение.

Профиль: основная модель на 20–32 млрд. параметров, отдельная модель на 7–14 млрд. для классификации и инструментов, векторный поиск и повторное ранжирование.

Инфраструктура: система AMD со 128 ГБ или DGX Spark; RTX 5090, если скорость важнее размещения модели на 70 млрд. параметров.

Граница роста: 95-й процентиль времени ответа выше целевого значения в течение 20% рабочего времени, очередь более пяти задач или потребность в обновлениях без простоя.

Средняя компания: 100–250 сотрудников

Пользователи: 40–100 активных пользователей за день, 5–16 одновременных сессий средней тяжести, фоновые очереди распознавания и индексации.

Типовой процесс: несколько областей поиска с разными правами, сервисная поддержка, анализ договоров, отделы в нескольких местах.

Профиль: быстрая модель для массовых запросов, более качественная модель для сложных задач, отдельные контуры обработки изображений и речи.

Инфраструктура: две станции с RTX 5090 как независимые копии сервиса либо профессиональный ускоритель с 96 ГБ; база данных и хранилище — на отдельном узле.

Граница роста: круглосуточная работа, несколько команд разработки или регулируемые данные переводят проект в серверный класс.

Растущий средний бизнес: 250–500 сотрудников

Пользователи: 100–250 активных пользователей за день, 10–40 одновременных быстрых сессий и 8–16 сессий средней тяжести, несколько фоновых очередей.

Типовой процесс: корпоративный портал ИИ, контакт-центр, анализ документов и изображений, агенты с подтверждаемыми действиями.

Профиль: несколько моделей, приоритеты очередей, ограничения частоты запросов и отдельные контуры данных.

Инфраструктура: RTX PRO 6000 или серверный ускоритель, два узла формирования ответов, отдельные базы данных, поиск и наблюдение, резервный контур.

Граница роста: компактные станции остаются лабораторией и периферийным слоем; центральный сервис становится серверной платформой.

Размер компании	Активные пользователи за день	Пиковые ответы	Одновременные многошаговые задачи	Разумный старт
10–30 сотрудников	5–10	1–2	1	Mac mini M4 Pro 64 ГБ
30–100	15–40	2–6	1–3	AMD 128 ГБ или DGX Spark
100–250	40–100	5–16	2–6	две RTX 5090 или профессиональный ускоритель
250–500	100–250	10–40 быстрых	4–12	RTX PRO или сервер с ускорителем

Диапазоны — модель планирования, а не результат испытаний. В контакт-центре 30 сотрудников могут создать больше нагрузки, чем 300 офисных пользователей. Размер компании задаёт исходную гипотезу; окончательное решение принимает нагрузочный тест.

20. Протокол нагрузочного пилота

Зафиксировать конфигурацию

В протокол заносятся точная модель, квантизация, размер контекста, среда выполнения и версия драйвера, параметры пакетной обработки, температура и режим питания, фоновые сервисы, формат векторных представлений, размер индекса, длина найденных фрагментов и параметры повторного ранжирования. Без этих данных результаты в токенах в секунду несопоставимы.

Собрать рабочий набор

1. 50–100 вопросов по внутренним документам с проверяемым ответом и ссылкой на источник.
2. 20 вопросов, на которые система должна отказаться отвечать из-за отсутствия данных.
3. 20 конфликтующих документов с версиями и датами.
4. 20 запросов с попыткой получить чужие документы или обойти инструкцию.
5. 10–20 длинных задач, запускающих инструменты, очередь или несколько агентов.
6. Набор типовых файлов: документы со слоем текста, сканы, таблицы, письма и изображения.

Измерять четыре слоя

Слой	Метрики	Минимальный критерий пилота
Качество	полнота, опора на источники, корректность ссылок, отказ при отсутствии данных	не хуже согласованного эталона сотрудника
Скорость	время до первого токена, токенов в секунду, медиана, 95-й процентиль, полное время ответа	целевое время выполняется на пике, а не только для одного запроса
Надёжность	ошибки, нехватка памяти, перегрев, восстановление, глубина очереди	8–24 часа длительного теста без деградации
Безопасность	ролевой доступ, утечки, вредоносные инструкции, неправомерные вызовы инструментов, аудит	нулевой доступ к запрещённым документам и действиям

Шесть прогонов

Прогон А: один пользователь, короткий контекст — базовая скорость.

Прогон Б: один пользователь, длинный контекст — предел памяти и время до первого токена.

Прогон В: 2, 4, 8 и 16 параллельных запросов — совокупная скорость и 95-й процентиль.

Прогон Г: рабочий день с переменной нагрузкой — очередь и стабильность.

Прогон Д: многошаговый процесс из 3–10 обращений к модели — усиление нагрузки и стоимость.

Прогон Е: отказ узла и восстановление — фактическое время восстановления и допустимая потеря данных.

Условия продолжения или остановки

Проект продолжается, если качество достигает бизнес-порога, 95-й процентиль укладывается в целевое время, права работают на уровне документа, нагрузка имеет 30% запаса, восстановление проверено, а экономия времени подтверждена владельцем процесса. Проект останавливается или пересматривается, если эффект держится только на демонстрационных вопросах, система требует постоянного ручного ухода разработчика или одна ошибка создаёт несопоставимый риск.

Карточка решения по пилоту

В итоговом протоколе должны остаться:

1. бизнес-процесс, владелец и измеримый исходный показатель;
2. версия документов, модели, квантования, среды выполнения и драйверов;
3. медиана, 95-й процентиль, совокупная скорость и глубина очереди;
4. доля принятых ответов, критические ошибки и причины отказа;
5. результаты проверки прав, вредоносных инструкций, вызовов инструментов и восстановления;
6. оценка совокупной стоимости владения на 24 месяца и цена следующего шага;
7. допущения, которые ещё требуют подтверждения.

Повторный прогон обязателен после замены модели, существенного изменения корпуса знаний, обновления драйвера или среды выполнения и появления новой пользовательской группы.

21. Референсные технологические стеки

Стек А. Безопасный поиск по корпоративным документам

Слой	Минимум	Когда усложнять
Интерфейс	Open WebUI или корпоративный портал	нужен собственный пользовательский опыт и сложные роли
Шлюз	FastAPI и LiteLLM	несколько моделей, квоты и маршрутизация
Формирование ответа	MLX, llama.cpp, Ollama или LocalAI	появляется пакетная обработка и формальное целевое время
Документы	проверенный модуль разбора и распознавание текста	сложные таблицы и версии
Поиск	PostgreSQL с pgvector или Qdrant	десятки миллионов фрагментов
Повторное ранжирование	локальная модель совместного кодирования	базового поиска недостаточно
Управление доступом	обратный прокси-сервер и Keycloak	много подразделений
Наблюдение	структурные журналы и OpenTelemetry	нужен Langfuse и передача событий в центр безопасности
Резервное копирование	снимок базы, индекса и конфигурации	нужна внешняя автономная копия

Стек Б. Мультиагентный процесс

Оркестратор агентов описывает конечный процесс: получить задачу, определить класс данных, извлечь документы, вызвать разрешённые инструменты, запросить подтверждение, записать результат и журнал. Практичный состав: LangGraph или LangChain для состояния; Redis или RabbitMQ для очереди; отдельные служебные учётные записи; изолированная среда для кода; список разрешённых инструментов; подтверждение сотрудником для записи; трассировка каждого обращения к модели и инструменту.

Стек В. Фоновая обработка документов

Подготовка данных отделяется от интерактивного формирования ответов. Очередь принимает файлы, антивирус проверяет вложение, модуль разбора извлекает структуру, распознавание обрабатывает сканы, документ делится на фрагменты, векторные представления строятся пакетно, а индекс обновляется атомарно. Пользователь видит состояние и версию индекса. Ночную обработку можно выполнять, не отнимая ресурсы у дневного чата.

Стек Г. Речь и изображения

Распознавание и синтез речи, распознавание текста и мультимодальные модели имеют разные профили нагрузки. Для встреч допустима фоновая расшифровка; для контакт-центра нужен потоковый режим с ограниченной задержкой; для производственной камеры — детерминированный периферийный конвейер. Управление доступом, хранение, мониторинг и аудит могут быть общими, но сами процессы формирования ответа лучше разделять.

Роли моделей вместо «одной самой умной»

Роль	Предпочтительный класс	Причина
Маршрутизация и классификация	3–8 млрд. параметров	быстро и экономно
Основной ассистент	8–32 млрд. параметров	баланс качества и параллельной нагрузки
Сложный анализ	70 млрд. параметров или разрешённое облако	редкие запросы и очередь
Векторный поиск	отдельная многоязычная модель	стабильный индекс
Повторное ранжирование	модель совместного кодирования	повышение релевантности
Распознавание текста и речи	специализированная модель	выше предсказуемость, чем у универсальной модели
Анализ текста и изображений	отдельная мультимодальная модель	документы, фотографии и схемы

Сначала небольшая модель определяет класс запроса, права и нужные инструменты. Основная модель решает типовую задачу. Более тяжёлая локальная модель или разрешённое облако подключается только при превышении порога сложности. Ошибки маршрутизации измеряются отдельно, а чувствительные категории всегда остаются в локальном контуре.

22. Модель совокупной стоимости владения и чувствительность

Что включено

В оборудование входят вычислительный узел, накопитель или сетевое хранилище, ИБП и базовая сеть. Внедрение включает установку, загрузку документов, настройку поиска, прав и испытаний. Эксплуатация включает обновления, мониторинг, резервное копирование, инциденты и поддержку пользователей. Электроэнергия рассчитывается отдельно, потому что её доля сильно различается между мини-компьютером и станцией с мощным ускорителем.

Что не включено

Не учтены ремонт помещения, выделенная линия электропитания, лицензии коммерческого ПО, внешняя сертификация, зарплата владельца продукта, очистка данных и потери при остановке. Для защищённого контура эти статьи добавляются отдельным бюджетом.

Чувствительность решения

Фактор	Если ниже базового	Если выше базового	Управленческий вывод
Загрузка	облако дешевле	локальный контур окупается быстрее	измерять запросы и токены
Стоимость специалиста	затраты близки к цене оборудования	эксплуатация доминирует	автоматизировать обновления и резервное копирование
Требования безопасности	достаточно простого пилота	изоляция увеличивает объём работ	не смешивать проверку концепции и защищённый проект
Качество локальной модели	можно отказаться от облака	нужен резервный маршрут или модель на 70 млрд. параметров	считать риск ошибок, а не только цену
Доступность оборудования	можно начать с одного узла	нужен запасной узел или второй канал поставки	фиксировать время восстановления и гарантию
Рост пользователей	достаточно второй копии сервиса	нужен серверный класс	масштабировать по 95-му процентилю и очереди

Как считать эффект

Для каждой задачи фиксируются число операций в месяц, текущее время сотрудника, время с ИИ, доля принятых результатов и стоимость ошибки. Экономия учитывается только для принятых и проверенных результатов. Выручку нельзя приписывать ИИ целиком: считается лишь измеримая доля, подтверждённая контрольной группой или сравнением до и после.

Пример: 20 сотрудников обрабатывают по 30 документов в месяц. Если проверенное сокращение времени составляет 15 минут на документ, система высвобождает 150 часов в месяц. Это не обязательно означает сокращение штата; эффект может выражаться в пропускной способности, сроке ответа и уменьшении переработок.

23. Контроль качества и приёмка

Качество ответа

- ответ опирается на доступные пользователю источники;
- ключевые утверждения имеют ссылку на документ и фрагмент;
- конфликтующие версии явно показаны;
- при отсутствии данных модель говорит об этом;
- формат ответа соответствует рабочему процессу;
- итог можно проверить без доступа к внутренним рассуждениям модели.

Качество сервиса

- целевое время измеряется по 95-му процентилю для полного цикла запроса;
- нехватка памяти и перегрев не возникают в длительном прогоне;
- очередь ограничена и имеет приоритеты;
- ошибки не приводят к повторной записи или двойному действию;
- новая модель проходит регрессионный набор до публикации;
- возврат к предыдущей версии выполняется по документированной процедуре.

Безопасность агента

- инструмент получает минимальные права и отдельную учётную запись;
- аргументы проверяются по схеме;
- запись, платёж, удаление и внешняя отправка требуют подтверждения;
- найденный в документе текст не становится системной инструкцией;
- секреты не попадают в запросы и журналы;
- каждый вызов инструмента связан с пользователем, целью и результатом.

Комплект эксплуатационной документации

1. Архитектурная схема и перечень данных.
2. Реестр моделей, лицензий и контрольных сумм.
3. Инструкция запуска, остановки и обновления.
4. Проверенная инструкция резервного копирования и восстановления.
5. Матрица ролей и источников документов.
6. Набор регрессионных вопросов и атак.
7. Журнал изменений конфигурации.
8. Контакты ответственных и порядок работы с инцидентами.

24. Глоссарий

Активные пользователи за день — уникальные пользователи, обратившиеся к сервису в течение суток; этот показатель не равен числу одновременных сессий.

Векторные представления — числовое описание фрагментов текста для смыслового поиска.

Видеопамять — выделенная память графического ускорителя; равный объём общей и выделенной памяти не означает равную скорость.

Внедрение вредоносной инструкции — попытка изменить правила системы через пользовательский запрос или найденный документ.

Квантизация — снижение точности весов модели для уменьшения объёма памяти и ускорения работы, иногда ценой качества.

Локальный маршрут по умолчанию — правило, при котором запрос обрабатывается внутри компании, а внешнее обращение допускается только для разрешённых категорий.

Многошаговая агентная задача — процесс, в котором один запрос пользователя вызывает несколько обращений к модели, поиску и инструментам.

Объединённая память — общий пул памяти процессора и графического ускорителя.

Пакетная обработка запросов — объединение нескольких запросов для более эффективной загрузки ускорителя.

Параллельная нагрузка — число запросов, которые система обслуживает одновременно.

Полностью изолированный контур — среда без прямого сетевого обмена с внешней инфраструктурой.

Повторное ранжирование — дополнительная сортировка найденных фрагментов по релевантности.

Резервный облачный маршрут — управляемое обращение к внешней модели только для разрешённых данных и сценариев.

Совокупная стоимость владения — цена оборудования, внедрения, эксплуатации, поддержки и энергии за выбранный срок.

Формирование ответа с поиском по документам — подход, при котором модель получает найденные корпоративные материалы и формирует ответ со ссылками на них.

25. Источники

1. Apple. Mac mini — Technical Specifications. https://www.apple.com/mac-mini/specs/
2. Apple Newsroom. Mac mini with M4 and M4 Pro. https://www.apple.com/newsroom/2024/10/apples-new-mac-mini-is-more-mighty-more-mini-and-built-for-apple-intelligence/
3. Framework. Introducing the Framework Desktop. https://frame.work/blog/introducing-the-framework-desktop
4. Minisforum. MS-S1 Max specifications and price. https://store.minisforum.com/products/minisforum-ms-s1-max-mini-pc
5. Minisforum Russia. MS-S1 Max configurations and availability. https://minisforum.ru/products/minisforum-ms-s1-max
6. NVIDIA. DGX Spark Hardware Overview. https://docs.nvidia.com/dgx/dgx-spark/hardware.html
7. NVIDIA Developer Forums. DGX Spark price change, 23 February 2026. https://forums.developer.nvidia.com/t/2-23-2026-price-change-announcement/361713
8. NVIDIA. DGX Spark Release Notes. https://docs.nvidia.com/dgx/dgx-spark/release-notes.html
9. Apple Support. Mac Studio (2025) technical specifications. https://support.apple.com/en-mide/122211
10. NVIDIA. RTX PRO 6000 Blackwell specifications. https://www.nvidia.com/en-us/products/workstations/professional-desktop-gpus/rtx-pro-6000/
11. NVIDIA Marketplace. RTX PRO 6000 Blackwell price. https://marketplace.nvidia.com/en-us/enterprise/laptops-workstations/nvidia-rtx-pro-6000-blackwell-workstation-edition/
12. CompDay. RTX 5090 / 96 GB workstation, Russian price. https://www.compday.ru/rabochie-stantsii/570688.htm
13. Shopozz. DGX Spark import listing. https://shopozz.ru/de/items/397240396620-nvidia-dgx-spark-4-tb-nvme-128-gb-ddr5x
14. Выберу.ру со ссылкой на данные Банка России. Курс на 13 июля 2026 года. https://www.vbr.ru/novosti/kursy-valyut/2026/07/13/kyrs-valut-13-iulya-2026/
15. AMD ROCm. Linux support matrix for Ryzen AI Max+ 395. https://rocm.docs.amd.com/projects/radeon-ryzen/en/latest/docs/compatibility/compatibilityryz/native_linux/native_linux_compatibility.html
16. ggml-org. llama.cpp supported backends. https://github.com/ggml-org/llama.cpp
17. КонсультантПлюс. Статья 18 и локализация ПДн. https://www.consultant.ru/document/cons_doc_LAW_61801/
18. Банк России. Рекомендации по безопасному использованию ИИ в финансовой сфере, 16 июня 2026 года. https://www.cbr.ru/press/event/?id=32627
19. YADRO. VEGMAN R220 G3. https://yadro.com/ru/vegman/r220g3/
20. NVIDIA Customer Story. Onetera. https://www.nvidia.com/en-us/case-studies/ai-powered-smarter-municipal-operations/
21. NVIDIA Customer Story. Bouygues Construction. https://www.nvidia.com/en-us/case-studies/digital-twins-sustainable-construction/
22. NVIDIA Customer Story. Lowe's. https://www.nvidia.com/en-gb/case-studies/lowes/
23. NVIDIA Blog. AiFi retail case. https://blogs.nvidia.com/blog/ai-fi-retail/

24. NVIDIA Technical Blog. Edge AI quality inspection on beverage packaging. https://developer.nvidia.com/blog/nvidia-gtc-industrial-at-the-edge/
25. Red Hat. Banco Galicia on-premise NLP platform. https://www.redhat.com/en/success-stories/banco-galicia-NLP
26. NVIDIA Customer Story. ThinkDeep sovereign AI agents. https://www.nvidia.com/en-us/case-studies/thinkdeep-sovereign-ai-agents-automate-public-services/
27. Федеральный закон от 30.11.2024 № 420-ФЗ. https://publication.pravo.gov.ru/document/0001202411300011
28. КонсультантПлюс. Статья 18.1 152-ФЗ: публикация политики оператора. https://www.consultant.ru/document/cons_doc_LAW_61801/eeeebe22bf738fd65bb66b95cc278911ae2525ee/
29. NVIDIA. GeForce RTX 5090 — specifications. https://www.nvidia.com/en-us/geforce/graphics-cards/50-series/rtx-5090/
30. iQmac. Apple Mac mini M4 Pro, 64 ГБ / 1 ТБ — российская карточка цены. https://www.iqmac.ru/catalog/mac/mac_mini/24893/
31. Price.ru. Apple Mac mini M4 Pro, 64 ГБ / 1 ТБ — срез предложений российского рынка. https://price.ru/sistemnye-bloki/apple-mac-mini-m4-1tb/

© 2026 Алексей Торицын. Независимое исследование.

www.toritsyn.com · Telegram: @ant_online_channel