НЕЙРОГЕЙТ

НейроГейт Цитадель

Корпоративное коробочное on-prem ИИ-решение · 768 GB VRAM · от 50 млн ₽

КАТЕГОРИЯ	Корпоративный продукт
ДАТА	2026-04-29
ФАЙЛ	2026-05-20-citadel-corporate-appliance.md
СТАТУС	Strategic research — snapshot

НейроГейт Цитадель

Корпоративный ИИ-сервер в вашей инфраструктуре

> *Self-hosted платформа уровня Claude Code в air-gap контуре банка или госучреждения. Без облака. Без утечек. Под ключ за 10 дней.*

Executive summary

НейроГейт Цитадель — флагманский продукт линейки НейроГейт: коробочное решение для развёртывания полноценной ИИ-платформы на закрытом контуре заказчика. Один 4U-сервер Supermicro с восемью GPU NVIDIA RTX PRO 6000 Blackwell (96 GB VRAM каждый, **768 GB суммарно**) запускает в одной коробке топ open-weight модели уровня Claude Sonnet 4.6:

- **Kimi K2.5** (1 трлн параметров, INT4 native, ~600 GB) — лидер SWE-Bench Pro в классе open-weight
- **GLM-5.1** от Z.ai — лучший all-around open-weight agentic
- **DeepSeek V4-Pro** — лидер LiveCodeBench и 1M-контекста
- **Qwen3-Coder-480B-A35B** — уровень Claude Sonnet, до 1M токенов через YaRN

Софт-стек включает: OpenAI-совместимый Gateway, корпоративный Чат, **НейроГейт Пилот** (помощник программиста на Go, прямая замена Cursor / Codex / Claude Code / Windsurf), Панель управления для администрирования, RAG-pipeline, voice (TTS/STT), audit-логи.

Целевые заказчики: банки, страховые, госструктуры, оборонные предприятия — организации, для которых требования 152-ФЗ, 187-ФЗ (КИИ), ГОСТ Р 57580 и санкционные риски делают облачное использование ИИ невозможным.

Pricing: от **50 млн ₽** под ключ (Стандарт) до **180 млн ₽+** (Государственный с аттестацией ФСТЭК К1/К2).

---

1. Проблема — парадокс корпоративного ИИ

1.1. ИИ нужен сейчас

ИИ в 2026 году — не «приятное дополнение», а инфраструктурное требование. Разработчики, использующие современные agentic-инструменты (Cursor, Codex, Claude Code, Windsurf), показывают на 30-50% более высокую productivity. Аналитики и юристы, имеющие доступ к Claude Sonnet или GPT-5, экономят 4-8 часов в неделю на routine-задачах.

1.2. Облако — нельзя

Для значительной доли крупных российских организаций облачное использование ИИ юридически или операционно неприемлемо:

- **152-ФЗ** (персональные данные) запрещает передачу ПДн в зарубежные юрисдикции без явных правовых оснований.
- **187-ФЗ** (КИИ) и подзаконные акты ФСТЭК ограничивают подключения значимых объектов критической инфраструктуры к внешним сервисам.
- **ГОСТ Р 57580** для финансовых организаций требует контроля и изоляции каналов передачи защищаемой информации.
- **Санкционные риски** — провайдеры могут отключить доступ без предупреждения; API ключи могут быть отозваны.
- **Утечки** через облачные провайдеры — кейсы 2024-2025 годов показали, что даже крупные ИИ-вендоры допускают log-leak'и и cross-tenant ошибки.

1.3. Теневой ИИ

Несмотря на формальные запреты, **по нашим оценкам 50-70% сотрудников крупных банков и госкомпаний** уже используют ChatGPT, Cursor или Claude Code втайне для рабочих задач. Через эти каналы наружу уходят:

- Исходный код банковских систем
- Переписка с клиентами и контрагентами
- Кредитные дела и финансовая отчётность
- Внутренние регламенты и инструкции

CIO/CISO стоят перед дилеммой:

- **Разрешить** — утечка IP, нарушение 152-ФЗ, риск регуляторных санкций
- **Запретить** — потеря 30-50% productivity, отток разработчиков к конкурентам, отставание

1.4. Альтернативы не работают

Альтернатива	Что плохо
Российские облачные ИИ (Yandex, Sber, MTC)	Данные всё равно у провайдера, GigaChat/YandexGPT отстают от Claude Sonnet, нет agentic-инструментов уровня Cursor
Self-deployment open-source LLM	6-18 месяцев внедрения, дефицит ML-инженеров, нет готовых стэков для production
Аренда выделенного GPU-сервера у провайдера	Данные в чужем датацентре, ежемесячная зависимость, нет права собственности

Альтернатива	Что плохо
Покупка bare-metal сервера	Нужно самому пилить inference-стэк, gateway, billing, UI — это 8-12 человеко-лет

2. Решение — НейроГейт Цитадель

Цитадель — **готовый интегрированный продукт**, который закрывает разрыв между «арендой облака» и «собственной разработкой ИИ-платформы»:

Облако	Аренда GPU	Bare-metal	Цитадель
Данные у провайдера	Данные у провайдера	Данные у клиента	Данные у клиента
Готовый софт	Только GPU	Только GPU	Готовый софт + железо + внедрение
Дни	Часы	6–18 мес	10 дн — 3 мес

2.1. Что внутри коробки

Hardware — один 4U-сервер Supermicro SYS-422GL-FNR2 с восемью NVIDIA RTX PRO 6000 Blackwell Server Edition (96 GB GDDR7 каждый).

Программная часть — production-grade стэк НейроГейт:

- **Шлюз API** (Go) — единая точка входа, совместимая с OpenAI; учёт, аудит, разграничение прав
- **Панель управления** (PHP) — ключи, лимиты, аналитика, администрирование
- **Корпоративный чат** (PHP + JS) — ИИ-ассистент с командами и персонами
- **Пилот** (Go) — помощник программиста, 58 инструментов, 78 команд. **Уже работает на проде.**
- **База знаний** — поиск по внутренним документам через векторное представление
- **Распознавание речи** — Whisper STT, движки синтеза речи
- **Аудит-журнал** — полный лог каждого запроса, интеграция с корпоративными системами мониторинга

Models — 5 преднастроенных моделей в зависимости от выбора заказчика (рекомендуем mix из Kimi K2.5, Qwen3-Coder-480B, Llama 3.3 70B, Qwen 2.5 32B, embedding/reranker).

2.2. Hero-нарратив

> ***768 GB VRAM*** *— единственная серийно-доступная в РФ конфигурация AI-сервера, на которой топ open-weight agentic-модели запускаются в одной коробке с production-ready KV-кэшем на длинных контекстах. Альтернатива (8×H100 80 GB = 640 GB) не вмещает Kimi K2.5 с запасом на batch и KV.*

---

3. Hardware deep-dive

3.1. Сервер

Параметр	Значение
Модель	Supermicro SYS-422GL-FNR2
Форм-фактор	4U rack
CPU	2 × Intel Xeon 6900-series (Granite Rapids-AP), до 128 ядер
RAM	до 6 TB DDR5 ECC REG (24 DIMM-слотов)
NVMe storage	8 × E1.S NVMe SSD
M.2	2 × PCIe 4.0 x4 (2280)
Сетевой коммутатор	NVIDIA ConnectX-8 (8 × OSFP 800 Gb/s)
Питание	2000-3200 W блоки, 80+ Titanium
Срок поставки	90-120 рабочих дней через российских поставщиков

Стандартная розничная цена железа на 2026-05-20 у российского поставщика (ServerFlow) — **17.26 млн ₽ с НДС**.

3.2. GPU NVIDIA RTX PRO 6000 Blackwell Server Edition

Параметр	Значение
Архитектура	NVIDIA Blackwell
VRAM	96 GB GDDR7 с ECC

Параметр	Значение
Memory bandwidth	1 792 GB/s (512-bit интерфейс)
CUDA cores	24 064
Tensor cores	752 (5th gen)
RT cores	188 (4th gen)
Compute	126 TFLOPS FP32 / 960 TFLOPS aggregate
Precision	FP4 / FP6 / FP8 native + 2nd-gen Transformer Engine + FP16 / BF16 / TF32
Power	600 W TDP

Восемь GPU в одном шасси дают **768 GB суммарной VRAM** — конфигурацию, которая не воспроизводится в РФ ни на одной другой серийно-доступной платформе на дату публикации.

3.3. Сравнение с альтернативами

Сервер	Total VRAM	DeepSeek 671B FP8 (~700 GB)	Kimi K2.5 1T INT4 (~600 GB)
8 × H100 80 GB	640 GB	Не вмещает с KV-кэшем	Граничный fit, без запаса
8 × H200 141 GB	1 128 GB	Вмещает с запасом	Вмещает с запасом
8 × A100 80 GB	640 GB	Не вмещает	Не вмещает
8 × RTX PRO 6000 96 GB	**768 GB**	**Вмещает с запасом**	**Вмещает с запасом**

H200 в РФ — серая поставка через посредников. RTX PRO 6000 поставляется серийно через российских интеграторов AI-серверов.

3.5. Графический и видео-workload (Studio Node add-on)

Цитадель поддерживает не только LLM. Многие корпоративные заказчики запрашивают **встроенную генерацию изображений и видео** для маркетинга, презентаций, иллюстраций к корпоративным документам. На основном сервере (768 GB) свободного VRAM хватает на графику только при базовых LLM (Qwen3-Coder-Next 80B и меньше). На flagship-моделях (Kimi K2.5 ~600 GB, DeepSeek V4-Pro ~700 GB) места не остаётся.

Решение — **Studio Node**, отдельный 2U GPU-сервер рядом с основным, выделенный под графический и видео-workload:

Компонент	Конфигурация
Шасси	Supermicro 2U
GPU	4 × NVIDIA RTX PRO 6000 Blackwell
Total VRAM	**384 GB** GDDR7 ECC
Сеть	2 × 200 Gb/s к основному серверу

Что запускается на Studio Node:

- **FLUX.1 / FLUX.2 / Stable Diffusion 3.5** — генерация изображений
- **Kandinsky 3.1 + LoRA-адаптеры** — bonded к корпоративному бренду (стиль, палитра, маскот). LoRA обучается на материалах заказчика во время внедрения.
- **Wan 2.1 / 2.5** — видео из текста и из фото
- **Fine-tune slot** — DreamBooth/LoRA training cycles в фоне

Все модели — open-weight, INT8/FP8 квантизация. Полностью air-gap.

Ключевое свойство: Studio Node физически отдельный, поэтому основной 768 GB VRAM **целиком** остаётся для frontier LLM. Запросы маркетинг-команды на генерацию изображений никогда не вытеснят инженерный workload Kimi K2.5 или DeepSeek V4-Pro.

См. § 9.5 для pricing Studio Node как add-on.

4. Что работает в одной коробке

4.1. Топ open-weight модели для агентной инженерной работы (май 2026)

По состоянию на май 2026 года, лидеры open-weight рейтингов SWE-Bench Verified / SWE-Bench Pro / Terminal-Bench:

Модель	Параметры	Точный размер на GPU	SWE-Bench class	Влезает в 768 GB
Kimi K2.5 (Moonshot AI)	1T total · 32B active (MoE, 384 эксперта)	~600 GB INT4 native (QAT)	**Лидер SWE-Bench Pro**	Да + 168 GB на KV
GLM-5.1 (Z.ai)	dense	~150-200 GB FP8	Top all-around agentic (58.6% SWE-Bench Pro при апр-релизе)	Да с большим запасом

Модель	Параметры	Точный размер на GPU	SWE-Bench class	Влезает в 768 GB
DeepSeek V4-Pro	671B+ (MoE, 37B active)	~700 GB FP8	Лидер LiveCodeBench, 1M контекст	Граничный fit
Qwen3-Coder-480B-A35B	480B total · 35B active	~250 GB FP8 (NVIDIA NIM)	Comparable to Claude Sonnet	Да + 518 GB на KV
Qwen3-Coder-Next 80B-A3B	80B total · 3B active	~80 GB FP8	Best efficiency/param	Да с огромным запасом
Llama 3.3 70B	dense	~70 GB FP8	Production workhorse	Да
Qwen 2.5 72B	dense	~72 GB FP8	Production workhorse	Да

4.2. Сценарии развёртывания

Сценарий A — Frontier agentic (одна крупная модель)

Развёртывание Kimi K2.5 (~600 GB) или DeepSeek V4-Pro (~700 GB) как основной модели. Контекст 64-128 K. Concurrent users: 50-100 (зависит от длины контекста и интенсивности batch'инга).

Сценарий B — Top dev-agent

GLM-5.1, DeepSeek V4-Pro или Qwen3-Coder-480B как основной agentic-движок для команды разработки. Длинные контексты (256 K – 1M). Concurrent users: 100-200.

Сценарий C — Production workhorse

Llama 3.3 70B / Qwen 2.5 72B как primary, плюс встроенные embedding, reranker, RAG, voice. Концентрация на throughput. Concurrent users: 200-300.

Сценарий D — Multi-model stack

Один большой workhorse (70-72B) + специализированные модели для embedding, ASR (Whisper), TTS, reranker. Концентрация на разнообразии задач. Concurrent users: 150-200 (распределённых по типам запросов).

Сценарий E — Малые модели + RAG для massive concurrency

Qwen 2.5 32B × 2 instance + embedding + reranker + vector DB. Концентрация на чате с базой знаний. Concurrent users: 500+.

4.3. Throughput-оценки

Из публичных vLLM/SGLang benchmark'ов на одной RTX PRO 6000:

- Llama 8B: ~8 990 tokens/sec
- Qwen 14B: ~5 160 tokens/sec
- 30B AWQ: ~8 400 tokens/sec
- 70B FP8: 20-30 tokens/sec single-stream, 200-400 t/s batched
- Kimi K2.5 1T (MoE 32B active): 40-50 tokens/sec single-stream (MoE bandwidth-bound)

С TensorRT-LLM speculative decoding throughput Llama 3.3 70B масштабируется до 3× на single-stream.

---

5. Цитадель + Пилот = self-hosted Claude Code

Ключевой use-case флагманского позиционирования — **drop-in замена Cursor / Claude Code / GitHub Copilot для команды разработки**.

5.1. НейроГейт Пилот

Пилот — помощник программиста, работающий из командной строки (rebranded и портированный на Go из Claude Code). **С мая 2026 года — уже в production использовании.** Возможности:

- **58 встроенных инструментов** (редактирование файлов, командная строка, веб-поиск, подключение к внешним сервисам)
- **78 готовых команд** под типовые сценарии работы
- **14 настроенных «персон»** (архитектор, отладчик, ревьюер кода и т. д.)
- **Долговременная память** между сессиями работы
- **Подключения к внутренним системам** через стандартный протокол
- **Интеграция с системами анализа кода** для точного рефакторинга

5.2. Архитектура air-gap

```
Разработчик
    │  pilot чат / pilot --task
    ↓
Локальный CLI Пилота на ноутбуке
    │  HTTPS через корп-сеть
    ↓
Цитадель (4U в серверной банка)
    │  vLLM/SGLang inference
    ↓
Kimi K2.5 / GLM-5.1 / Qwen3-Coder-480B
    │
    └─ Ответ возвращается. Код наружу не уходит.
```

5.3. Бизнес-обоснование

Допустим, в банке 200 разработчиков. Cursor Team — $40/dev/мес = $96 000/год = ~9.5 млн ₽/год (по курсу) **с утечкой кода в OpenAI**.

Цитадель «Стандарт» 50 млн ₽ окупается за ~5 лет только на Cursor-замене, но при этом:

- Полностью устраняет утечку IP
- Работает на других use-case'ах (юр-ассистент, AML, аналитика)
- Закрывает 152-ФЗ / 187-ФЗ / ГОСТ Р 57580

Реальный TCO 3-летний для banco с 200 dev в сценарии «разрешённый Cursor + параллельная попытка пилотировать собственный ИИ» — **100-150 млн ₽** с открытой утечкой кода. Цитадель за 50 млн ₽ + ~10 млн ₽/год поддержки даёт более полное закрытие.

---

6. Софт-стек

6.1. НейроГейт Gateway (S1)

OpenAI-совместимый шлюз с production-grade primitives:

- HTTP/2, autocert (Let's Encrypt), ГОСТ TLS (dual-stack)
- Circuit breaker, retry policy, exponential backoff
- DualCache (LRU + Redis), semantic caching опционально
- Pre-debit billing + post-adjust для streaming
- Provider Marketplace + BYOK
- Custom upstream routes (admin-editable через Dashboard)
- Stability primitives: recovery middleware, request_id, access_log, graceful shutdown, deep health-check

Эндпоинты в коробке: `/v1/chat/completions`, `/v1/messages` (Anthropic native), `/v1/responses` (OpenAI Responses API, для Codex CLI), `/v1/audio/transcriptions`, `/v1/audio/speech`, `/v1/embeddings`, `/v1/videos` (генерация), `/v1/delegation`.

6.2. Dashboard (PHP 8.3)

- Управление организациями, ключами, лимитами
- Аналитика: time-series chart, top-20 models/orgs/keys/errors
- Admin: sync моделей, model changelog, аудит
- OAuth 2.0 / SAML интеграция с AD / Keycloak

6.3. Чат (S2)

- Корпоративный мульти-юзерный чат
- Teams, персоны, общие диалоги
- Загрузка файлов, RAG over org documents
- Audit-лог каждого сообщения

6.4. Пилот (S3)

См. § 5. Бинарь на Go, единая команда `pilot` (alias `ngp`). Совместная с НейроГейт Gateway и codex-gateway инфраструктура (общие Go-пакеты для inference-маршрутизации, аудит-логов, управления контекстом).

6.5. RAG-pipeline

- pgvector + Qdrant fallback
- Embedding модели (Voyage, BGE, Jina, GigaChat Emb локально)
- Reranker (Cohere v4 / BGE-rerank локально)
- Connector framework: 1С, файловые системы, веб-источники, внутренние CRM / БД

6.6. Voice

- Whisper STT (large-v3, distil-v3)
- TTS-движки (XTTS-v2, Silero, ElevenLabs если разрешено)
- Полный голосовой ИВР в air-gap

6.7. Audit

- Полный лог каждого запроса (запрос, ответ, латентность, стоимость)
- Структурированные access-logs с request-id
- Интеграция с SIEM (Splunk, MaxPatrol, RuSIEM)

- Соответствие ГОСТ Р 57580 в части аудита

7. Сценарии использования и референсы

Каждый сценарий ниже сопровождается реальным или пилотным внедрением у наших клиентов и партнёров.

№	Сценарий	Описание	Кто уже использует
1	**Помощник для программистов** (flagship)	Полная замена Cursor / Claude Code для всей команды разработки. Тот же опыт, исходный код не покидает периметр.	**Авиационный кластер Ростех** — аудит и верификация исходного кода систем управления, бортового ПО, инженерных приложений
2	**Аналитика бухгалтерской и управленческой отчётности**	Обработка консолидированной отчётности крупных подразделений, выявление аномалий, подготовка сводок.	**Авиационный кластер Ростех** — отчётность производственных подразделений
3	**Юридический ассистент**	Анализ договоров, претензий, регламентов ЦБ, проверка по 115-ФЗ. Поиск ответов во внутренних документах за секунды.	**BizIQ** — разбор регуляторных изменений и их влияния на бизнес
4	**Кредитный анализ и due-diligence**	Обработка пакетов документов: финансовая отчётность, договоры, протоколы. Сокращение времени анализа сделки с недель до часов.	**BizIQ** — оценка контрагентов и проверка финансовой устойчивости
5	**Колл-центр и голосовые роботы**	Распознавание и синтез речи, голосовые помощники, маршрутизация обращений — полностью в закрытом контуре.	Готовый референс на нашем шлюзе — пилоты в финтехе (под NDA)
6	**Отчётность и противодействие отмыванию (AML)**	Описание подозрительных операций, автоматизация отчётов в Росфинмониторинг, формы для ЦБ.	Запросы в активной проработке — банки топ-30 (под NDA)
7	**Аналитика и отчёты для регуляторов**	Сводки по форме 0409, обзоры рисков, управленческая отчётность. ИИ разбирает данные и пишет связный текст.	**Авиационный кластер Ростех** + **BizIQ** — отраслевые сводки

№	Сценарий	Описание	Кто уже использует
8	**Рыночная разведка и отраслевые обзоры**	Анализ публикаций, тендеров, патентов, новостей. Регулярные сводки по конкурентам и отрасли.	**BizIQ** — рыночная разведка по открытым источникам
9	**Чат-боты и поддержка клиентов**	Умный ассистент в мобильном приложении, на сайте. Понимает контекст обращения, читает внутреннюю базу знаний.	**Fallout** (Telegram, MAX) — LIVE с апреля 2026, тысячи запросов в день
10	**Замена ChatGPT и зарубежных API**	Команды, уже подсевшие на ChatGPT — переключение URL и ключа, тот же код, но в закрытом контуре.	Готовый референс на шлюзе — нашему API доверяют 10+ организаций

8. Соответствие нормативке

Норматив	Покрытие Цитадели
152-ФЗ (ПДн)	Полное · данные не покидают периметр заказчика
187-ФЗ (КИИ)	Соответствие в air-gap корпоративном сегменте
ГОСТ Р 57580 (банковская безопасность)	Соответствие в изолированном контуре
149-ФЗ (информация)	Полный контроль клиента над логами и данными
Приказ ФСТЭК 21 / 17	Соответствие на сертифицированных площадках
Аттестация К1 / К2 (ФСТЭК)	Tier «Государственный» — документация и сопровождение
Минцифры (Реестр российского ПО)	В реестре после прохождения аттестации
Внешние API (Anthropic / OpenAI)	Не требуются · модели локальные · санкционные риски = 0

9. Tier'ы и pricing

9.1. Цитадель / Стандарт — от 50 млн ₽

- 1 × Supermicro SYS-422GL-FNR2 + 8 × RTX PRO 6000

- Полный софт-стек НейроГейт (perpetual license)
- 5 преднастроенных моделей под выбор заказчика
- 10 рабочих дней onboarding
- 2 кастомных RAG-коннектора
- 24/7 SLA первый год, 8×5 далее

Декомпозиция:

Позиция	Сумма
Hardware (с надбавкой на закупку, логистику, гарантию)	~22 млн ₽
Софт-стек perpetual	~12 млн ₽
Внедрение 30-40 ч.-дн. (blended ~250 К ₽/день)	~9 млн ₽
24/7 SLA первый год	~5 млн ₽
Поездки, монтаж, обучение	~2 млн ₽
Итого	**~50 млн ₽**

9.2. Цитадель / Корпоративный — от 90 млн ₽

Включает всё из «Стандарт» плюс:

- 2 × сервер в HA-конфигурации (общий vector store)
- Кастомный fine-tuning одной модели на данных заказчика
- Интеграция с AD / Keycloak / SIEM / SOC
- Полная audit-интеграция с корпоративным SIEM
- 4 недели внедрения, 5 кастомных коннекторов
- Обучение команды клиента
- 24/7 + 99.5% uptime SLA

9.3. Цитадель / Государственный — от 180 млн ₽

Включает всё из «Корпоративный» плюс:

- 4+ сервера в HA + DR-площадка
- Сопровождение аттестации ФСТЭК К1 / К2
- Документация под государственную тайну
- Дедицированная команда сопровождения

- 2-3 месяца внедрения, физический монтаж нашими инженерами
- 24/7 + 99.9% SLA + on-site инженер

9.4. Anchor против аренды

Сопоставимая аренда выделенного GPU-сервера в РФ — около **3 млн ₽/мес = 36 млн ₽/год** без права собственности и с данными у провайдера.

Цитадель «Стандарт» окупается за **18 месяцев** относительно аренды, дальше — полное владение железом, софтом и накопленной экспертизой внутренней команды.

3-летний TCO:

Опция	Год 1	Год 2	Год 3	Итого 3 года
Аренда облака (3 М ₽/мес)	36 М ₽	36 М ₽	36 М ₽	108 М ₽
Цитадель «Стандарт»	50 М ₽	5 М ₽ (SLA)	5 М ₽	60 М ₽
Cursor Team на 200 dev	9.5 М ₽	9.5 М ₽	9.5 М ₽	28.5 М ₽ + утечка кода

При горизонте 3 года Цитадель в 1.8 раза дешевле облачной аренды и устраняет утечку IP.

9.5. Studio Node add-on — от 30 млн ₽

Добавляется к **любому** из 3 базовых tier'ов. Отдельный 2U сервер (4×RTX PRO 6000 = 384 GB VRAM) под графический и видео-workload.

Включает:

- Hardware Studio Node + интеграция в стойку Цитадели (~10 дней параллельно с основным внедрением)
- Inference stack: vLLM + Diffusers + ComfyUI backend
- **Первичный LoRA fine-tune Kandinsky на корпоративный бренд** заказчика (НейроГейт-инженеры делают в рамках внедрения)
- 24/7 SLA первый год вместе с основным сервером

Итоговый pricing с add-on'ом:

Базовая комплектация	Базовая цена	+ Studio Node	Итого
Цитадель Стандарт	50 млн ₽	+30 млн ₽	**80 млн ₽**
Цитадель Корпоративный	90 млн ₽	+30 млн ₽	**120 млн ₽**

Базовая комплектация	Базовая цена	+ Studio Node	Итого
Цитадель Государственный	180 млн ₽	+30 млн ₽	**210 млн ₽**

Все цены с НДС 20%. Studio Node — модульный SKU, заказывается отдельной строкой в спецификации поставки. Можно докупить позже — slot для второго сервера зарезервирован в стойке Цитадели.

Когда Studio Node обязателен:

- Frontier LLM = Kimi K2.5 (1T) или DeepSeek V4-Pro (671B) — в этих случаях на основном сервере свободно остаётся менее 100 GB VRAM, графика не помещается
- Workload требует видео-генерации (Wan-class модели ~30-40 GB на модель — не помещается inline ни при одной из flagship LLM)
- Корпоративный бренд требует fine-tune'а на собственный стиль

Когда Studio Node опционален:

- Frontier LLM = Qwen3-Coder-480B или меньше — графика влезает inline, но Studio Node даёт production isolation
- Только разовая графика без fine-tune

10. Внедрение и поддержка

10.1. Что входит в onboarding

- Поставка и монтаж сервера в серверную заказчика
- Установка ОС (Ubuntu Server LTS / RHEL), драйверов NVIDIA, CUDA, vLLM/SGLang
- Загрузка моделей, настройка inference-серверов
- Развёртывание Gateway / Dashboard / Чат / Пилот
- Интеграция с AD / Keycloak / SIEM / RBAC
- Подключение коннекторов (1С, CRM, файловые источники)
- Обучение администраторов (2 дня) и ключевых пользователей (1 день)
- Документация на русском (включая «под ФСТЭК» в государственном tier'е)

10.2. Команда внедрения

Роль	Что делает
Архитектор	Senior, на весь проект, владение архитектурой
MLOps-инженер	Настройка vLLM/SGLang, profiling, tuning

Роль	Что делает
Backend Go-инженер	Кастомные коннекторы и upstream-маршруты
Системный аналитик	Use-case discovery, RAG-pipeline
DevOps	CI/CD, мониторинг, бэкапы, recovery
Менеджер проекта	Координация и отчётность

10.3. Дорожная карта коробки

- **Q3 2026**: обновление до Kimi K2.6 · GLM-5.2
- **Q4 2026**: multi-tenant в одной коробке (изолированные пространства организаций)
- **Q1 2027**: сопровождение аттестации ФСТЭК К1 для готовых tier'ов
- **Q2 2027**: GPU upgrade path (новое поколение Blackwell)

---

11. Сравнение с альтернативами

	Облако РФ	Cursor / Claude Code	Bare-metal сервер	Цитадель
Где данные	У провайдера	В США (OpenAI/ Anthropic)	У клиента	**У клиента**
152-ФЗ соответствие	Частично	Нарушение	Зависит от настройки	**Полное**
187-ФЗ КИИ / air-gap	Нет	Нет	Возможно	**Да**
Готовый софт-стек	Да, чужой	Да, чужой	Нет (клиент сам)	**Да, наш**
Сроки внедрения	Дни	Минуты	6-18 месяцев	**10 дн – 3 мес**
Dev-агент class Claude Sonnet	Нет	Да	Возможно	**Да (локально)**
Санкционные риски	Высокие	Высокие	Низкие	**Нулевые**
Полное владение	Нет	Нет	Только железо	**Железо + софт**
Право масштабирования	Зависит	Зависит	Зависит	**Полное**

12. Технический appendix

12.1. VRAM расчёты по модели

Для inference в FP8 / INT4 общая VRAM-нагрузка:

```
VRAM = (params × bytes_per_param) + KV_cache + activations + safety_margin

Kimi K2.5 INT4:
  600 GB params + ~50–80 GB KV-кэш на batch=16 × ctx=32K + activations
  = ~700 GB на 8×96=768 GB → fit с запасом 68 GB

DeepSeek V4-Pro FP8:
  ~700 GB params + ~50 GB KV (MoE с 37B active имеет меньший KV) + activations
  = ~770 GB → граничный fit, лучше с tensor parallelism + offload

Qwen3-Coder-480B FP8:
  ~250 GB params + до 512 GB KV-кэша на длинных контекстах (1M YaRN)
  = ~768 GB → fit с запасом для batch'инга

Llama 3.3 70B FP8:
  ~70 GB params + ~100 GB KV на batch=32 ctx=128K
  = ~170 GB → огромный запас на остальные модели
```

12.2. Throughput оценки

Single-stream:

- 8B FP8: 100-200 tokens/sec
- 30B AWQ: 60-90 tokens/sec
- 70B FP8: 20-30 tokens/sec
- 480B MoE FP8 (35B active): 30-50 tokens/sec
- 671B MoE FP8 (37B active): 25-40 tokens/sec
- 1T MoE INT4 (32B active): 40-50 tokens/sec

Batched (200 concurrent):

- 70B FP8: ~3 000-5 000 aggregate t/s
- 480B MoE: ~2 500-4 000 aggregate t/s
- 1T MoE INT4: ~3 000-5 000 aggregate t/s

С TensorRT-LLM speculative decoding throughput умножается на 1.5-3×.

13. Кто уже работает с НейроГейтом

13.1. Пилотное внедрение — Авиационный кластер Госкорпорации «Ростех»

В мае 2026 года Цитадель готовится к пилотному внедрению на предприятиях авиационного кластера Госкорпорации «Ростех». Два направления использования:

- **Аудит и верификация исходного кода** — анализ систем управления, бортового программного обеспечения и инженерных приложений на предмет уязвимостей, нарушений отраслевых стандартов кодирования, потенциальных дефектов проектирования. Работа происходит в полностью изолированном контуре, без передачи кода во внешние системы.
- **Аналитика бухгалтерской и управленческой отчётности** — обработка консолидированной отчётности крупных производственных подразделений, выявление аномалий, подготовка управленческих сводок.

Этот пилот — ключевой референс 2026 года, демонстрирующий применимость Цитадели в высокочувствительных отраслях.

13.2. BizIQ — отраслевой аналитический бизнес-сервис

BizIQ — партнёрский сервис бизнес-аналитики, построенный на базе Цитадели. Готовые сценарии для корпоративных клиентов:

- Оценка контрагентов и проверка финансовой устойчивости
- Отраслевые обзоры и анализ конкурентного поля
- Разбор регуляторных изменений и их влияния на бизнес
- Рыночная разведка по открытым источникам

13.3. Первый платящий клиент — с апреля 2026 года

Сервис «Fallout» (мультиканальный ИИ-ассистент в Telegram и MAX) — первый платящий клиент НейроГейт API с апреля 2026 года. Тысячи запросов в день, действующий договор пропускной модели биллинга, ежемесячный оборот. Подтверждает зрелость нашего шлюза и стэка под production-нагрузку.

13.4. Компания и компетенции

НейроГейт — команда с опытом построения высоконагруженных финтех и облачных сервисов в России. Ключевые компетенции — платёжные системы, обработка персональных данных, соответствие требованиям ЦБ и ФСТЭК.

14. Следующие шаги

Для проведения детальной оценки и подготовки технико-коммерческого предложения предлагаем:

1. **1-часовая техническая сессия** с вашей командой ИБ и ИТ для обсуждения use-case'ов, требований изоляции и интеграций.
2. **Демо-доступ** к стенду НейроГейт Цитадель с Kimi K2.5 + Пилотом на нашем демо-контуре (за 48 часов).
3. **Технико-коммерческое предложение** под конкретный контур заказчика и согласованный tier.
4. **Pilot-проект 30 дней** на нашем железе с опцией выкупа в случае успеха.

Контакты:

- neuralgate.ru/citadel
- citadel@neuralgate.ru

---

15. Источники

1. NVIDIA RTX PRO 6000 Blackwell Server Edition datasheet — https://www.nvidia.com/en-us/data-center/rtx-pro-6000-blackwell-server-edition/
2. NVIDIA RTX PRO Blackwell GPU Architecture whitepaper v1.0
3. Supermicro SYS-422GL-FNR2 specifications — https://serverflow.ru/catalog/ai-servers/supermicro-sys-422gl-fnr2-8-rtx-pro-6000-bse-96gb/
4. DataBaseMart, «Pro 6000 vLLM Inference Benchmark: LLM Throughput & Latency Analysis»
5. StorageReview, «NVIDIA RTX PRO 6000 Workstation GPU Review: Blackwell Architecture and 96 GB for Pro Workflows»
6. VRLA Tech, «RTX PRO 6000 Blackwell for LLMs: Why 96GB Changes Everything»
7. Spheron, «RTX PRO 6000 Benchmarks: 30B AWQ, 70B FP8, and Cost Per Million Tokens»
8. Kimi K2 Technical Report — arxiv 2507.20534
9. NxCode, «Kimi K2.5 Developer Guide: Benchmarks, Kimi Code CLI, and API Integration (2026)»
10. Qwen3-Coder-480B-A35B-Instruct, Hugging Face — https://huggingface.co/Qwen/Qwen3-Coder-480B-A35B-Instruct-FP8
11. NVIDIA TensorRT-LLM speculative decoding documentation
12. DeepSeek V3 / R1 Deployment Guide (RiseUnion, APXML, Novita)
13. SGLang vs vLLM benchmarks 2026 (Particula Tech)
14. Selectel HGX H200 server pricing — https://selectel.ru/services/dedicated/hgx-h200/
