



ОТЧЁТ ДЛЯ СОВЕТНИКОВ · АВГУСТ 2026

Национальная модель — это не вопрос. Это расписание.

Ответы на 6 вопросов, бизнес-экспертов

🔒 Документ для внутреннего круга: команда + советники

📖 Источники: ФЗ №243 от 26.07.2026 · Аналитика рынка on-prem AI · Конкурентная карта РФ

Вопросы: Темка или Бизнес?

Кратко: слишком много знаков вопроса и переменных. Кто-то точно заработает. Но это скорее «темка» и казино на данном этапе.

Моя работа на следующих 17 слайдах — превратить казино в расписание.

6 ВОПРОСОВ ОТ ЭКСПЕРТОВ

- 01 Нет гарантий, что сертификация будет скоро
- 02 Не понятно, кого обяжут переходить на нац. модели
- 03 Не задушат ли сертификацию BigTech'ом
- 04 Нужно каждую модель сертифицировать?
- 05 С чем приходить на рынок, как получить продажи
- 06 Где гарантии, что это кому-то надо в первый день

Автономная фабрика законченной работы, а не ещё одна AI-платформа

«Мы не строим «ещё одну российскую on-prem AI-платформу». Это уже есть у [МТС](#), [Яндекса](#), [Just AI](#), [Cloud.ru](#). Мы строим автономную фабрику: вы ставите бизнес-задачу вечером — к утру получаете проверенный результат и очередь решений, требующих человеческой подписи. **Не AI, который отвечает сотруднику. AI, который к утру сделал работу сотрудника.»**

— тезис идеи · для внутреннего круга советников

ЦИКЛ РАБОТЫ

1 НОЧЬ

Постановка бизнес-задачи вечером → готовый проверенный результат к утру.

КОНКУРЕНТЫ

4 / 0

4 on-prem платформы в РФ (МТС, Яндекс, Just AI, Cloud.ru) — но 0 делают фабрику законченной работы.

РЫНОК

71%

Крупных компаний РФ уже используют GenAI (Yakov & Partners / Яндекс, 2025) — спрос готов.

01

РАЗДЕЛ 01 ИЗ 03

6 вопросов от экспертов

И ответ на каждый — с цифрами, не с риторикой.

Q1 Сроки сертификации

Q4 Частота сертификации

Q2 Кого обяжут

Q5 Стратегия входа

Q3 Давление BigTech

Q6 Реальный спрос

01

ВОПРОС 1

Нет гарантий, что процедура сертификации будет скоро

ФОРМУЛИРОВКА ЭКСПЕРТА

«Нет гарантий, что процедура сертификации (чтобы стать нац. моделью) будет скоро. Может, всё это растянется на годы, и окно возможностей захлопнется раньше, чем мы успеем.»

” — из протокола обсуждения

ОТВЕТ · НА ОСНОВАНИИ ФЗ №243 И АНАЛИТИКИ

Расписание уже в законе. Процедуру примут в ноябре.

- 01 **ФЗ №243 подписан 26.07.2026.** Основная часть вступает в силу 1 сентября 2026. Минцифры поручено подготовить 5 постановлений Правительства до 1 ноября 2026 — включая порядок подтверждения соответствия.
- 02 Нормы о статусах и проверке соответствия вступают в силу **1 марта 2027**. Не «когда-нибудь» — конкретная дата. С этого момента статус можно получить.
- 03 Для систем, уже использующих иностранные модели на дату введения требования, **переходное исключение — до 1 сентября 2032**, при условии хранения данных в РФ.

🕒 Процедуру ждём в ноябре. Сертификацию начинаем весной. Это 6 месяцев, не 6 лет.

Кого обяжут? Госы переходят медленно, культуры использования нет

ФОРМУЛИРОВКА ЭКСПЕРТА

«Не понятно, кого обяжут использовать конкретно нац. модели. Кажется, что процесс интеграции ИИ в гос. учреждения идет очень медленно, пока нет сформировавшейся культуры использования даже не запрещенных (пока) чатов GPT среди сотрудников, нет внутренних систем почти ни у кого, а тут надо перейти на что-то новое. Что, когда и как — непонятно.»

— из протокола обсуждения

ОТВЕТ · ЗАКОН ОПРЕДЕЛЯЕТ И КАТЕГОРИИ, И СРОКИ

Закон прямо даёт Правительству право указать случаи. И есть переходная вилка.

- 01 **Статья 5 ФЗ №243:** Правительство определяет случаи, когда допускается использование исключительно национальных или суверенных моделей. Это не абстракция — это конкретный правовой механизм.
- 02 Распространение идёт волной: **ноябрь 2026** — процедура готова → **март 2027** — первый контур «исключительного применения» (гос. и критическая инфраструктура) → далее массовый госсектор → далее крупный бизнес (банки, ОКИИ, госкомпании) → далее средний бизнес через стандарты и закупочные требования.
- 03 Уже сегодня: **71% проникновения GenAI** среди крупных компаний РФ (Yakov & Partners / Яндекс, 2025). Те, кто перейдёт на нац. модели в 2027, будут не «пионерами» — они будут догонять тех, кто уже встроил AI в процессы.

📅 **Закон не спрашивает «когда». Закон отвечает «вот когда».** Мы не ждём спроса — мы подстраиваемся под даты.

Не задушат ли сертификацию так, что её реально будет невозможно получить? BigTech всем займётся

ФОРМУЛИРОВКА ЭКСПЕРТА

«Не усложнят и не задушат ли сертификацию так, что её реально будет невозможно получить? BigTech будут всем заниматься, даже не делая передовые модели — отожмут себе всё.»

ОТВЕТ · BIGTECH ДЕЛАЕТ ДРУГУЮ ИГРУ

BigTech оптимизирует под стоимость и массовость. Мы — под качество длинной задачи.

- BigTech не может позволить себе поддерживать больше одной модели нормально. Сейчас они отстают на 2 поколения. Они оптимизируют бизнес под массовость, и делают продукты для мгновенного чата: GPU, TTFT, токены/сек.
- Длинная автономная задача — другая метрика: не «как быстро ответил», а «выполнил ли за ночь корпус из 40 000 документов и выдал 137 решений на подпись». Для этого latency не важен, важна стоимость завершённой работы. BigTech не оптимизирует под эту функцию — у них в P&L её нет.

Термин в индустрии — **agent long-running task horizon**: длительность автономной задачи, на которой модель не деградирует. BigTech держит горизонт коротким (массовый чат), мы растягиваем на ночь.



BIGTECH ПУТЬ

Массовый продукт → дешёвая модель → массовая деградация качества

degradation cycle: одна тупая итерация портит следующую



НАШ ПУТЬ

Вертикальная фабрика → качество одной принятой работы

Model pyramid: маленькие модели для фильтрации + специализ. 30–70B + frontier для 1–5% сложных



АРГУМЕНТ BPERF

Сертификация модели — это не сертификация платформы

BigTech-продукты (платформы) остаются без статуса БФМ

Нужно будет каждую модель сертифицировать? Каждые 2 месяца новая — а если сертификация 6 месяцев?

ФОРМУЛИРОВКА ЭКСПЕРТА

«Нужно будет каждую модель сертифицировать? Каждые 2 месяца новая — а если сертификация 6 месяцев? Получается, что модель устареет ещё до того, как мы её сертифицируем.»

💡 принцип

Не «каждая новая модель», а «одна сертифицированная БФМ + семья продуктов на ней». BigTech двигает массу. Мы двигаем ядро.

ОТВЕТ · НЕ НАДО ГНАТЬСЯ ЗА НОВОЙ МОДЕЛЮ

Достаточно обгонять BigTech. Они отстают от фронта на 2 поколения.

- По нашим данным, российский BigTech сидит на 2 поколения позади фронта: Сбер/GigaChat, МТС/Cotype, Яндекс, Wildberries, Т-Банк — все используют зарубежные open-source модели. Закон прямо разрешает national-BFM быть fine-tune'ом иностранной открытой модели — существенные характеристики определяет российский разработчик.
- Нам не нужно выпускать новую модель каждый раз когда новая выходит. Нам нужно выпускать модель так же часто как BigTech в РФ
- Сертификация = сертификация одной БФМ. Платформа поверх неё, fine-tune'ы под вертикали, model adapters — всё это не требует отдельной сертификации. Один статус → много продуктовых обёрток.

С чем приходить на рынок, как получить продажи и конкретное преимущество?

ФОРМУЛИРОВКА ЭКСПЕРТА

«С чем приходить на рынок, как получить продажи и конкретное преимущество? Что мы реально продаём, и кому?»

Вопрос не про «как продавать». Вопрос про **что именно продаём** — это и формирует продажу.

ОТВЕТ · ПЕРВАЯ ВЕРТИКАЛЬ = REGULATORY CHANGE FACTORY

Не платформу. Один workflow с измеримым KPI.

- Beachhead: Regulatory Change Factory для юр. функции крупной компании. Вход = новый нормативный акт → корпус 1 000–100 000 внутренних документов → список затронутых положений → draft новой редакции → cross-check вторым агентом → очередь юристу → audit trail. Один процесс — один пилот — одна цена.
- Единица продажи: не «AI-платформа за 10 млн ₽», а «автоматизируем impact-analysis нового нормативного акта за фиксированный пилот с заранее определёнными KPI». 0,5–3 млн ₽ внедрение + 50–300 тыс. ₽/мес за процесс. Первые 2–3 пилота уйдут на design partners с customer-owned outcome.
- Конкретное преимущество: при расторжении у клиента остаются исходный код, веса моделей, dataset manifest, eval suite, infrastructure manifests, hardware BOM и audit history. Нас можно уволить — система продолжит работать. Это невозможно у BigTech, потому что их модель = lock-in в экосистему.



ЕДИНИЦА ПРОДУКТА

Одна законченная бизнес-работа, выполненная автономно внутри периметра за гарантированный бюджет, с доказательствами и последней подписью человека.

Где гарантии, что это кому-то надо в первый день? Может, обязательный переход в 2030 или позже

ФОРМУЛИРОВКА ЭКСПЕРТА

«Даже если станем компанией с национальной моделью, то где гарантии, что это реально кому-то надо будет в первый же день? Может, обязательный переход на национальные будет в 2030 или позже.»

Спрос в этом рынке создаётся не маркетингом. Его создаёт закон.

ОТВЕТ · СПРОС ОПРЕДЕЛЁН РЕГУЛЯТОРИКОЙ, НЕ МАРКЕТИНГОМ

Закон уже утвердил кто, когда и как переходит. Это не маркетинг, это расписание.

- Статья 5 ФЗ №243: Правительство определяет случаи, когда допускается использование исключительно национальных или суверенных моделей. Для финансового рынка эти ограничения согласуются с Банком России. Это значит, что у ЦБ, Минцифры и Правительства уже есть перечень применений, которые нельзя закрыть иностранными LLM после 1 марта 2027.
- Распространение: 2026 — принятие подзаконных актов → 2027 — первый контур «исключительного применения» → далее гос. и крупный бизнес → далее массовый рынок. Спрос создаётся законом, а не воспитанием рынка.
- Уже сегодня: рынок российского ИИ-ПО ≈ 25 млрд ₽ (TAdviser, 2025, +27% год-к-году). Расходы бизнеса на внедрение ИИ — 257 млрд ₽. Проникновение GenAI среди крупных компаний РФ — 71% (Yakov & Partners / Яндекс, 2025). Деньги уже в рынке. Не хватает именно certified national-BFM.



АНАЛОГИЯ · ГЛОБАЛЬНЫЙ ТРЕНД

ЕС, Китай и США идут тем же путём — обязательная локализация чувствительных применений на национальные модели. Смена политического режима в РФ этот тренд не отменит.

02

РАЗДЕЛ 02 ИЗ 03

РЫНОК И ОКНО ВОЗМОЖНОСТЕЙ

Где сидят BigTech, где пустая ниша, и сколько это стоит.

 Таймлайн 243-ФЗ

 Конкурентная карта

 Экономика модели

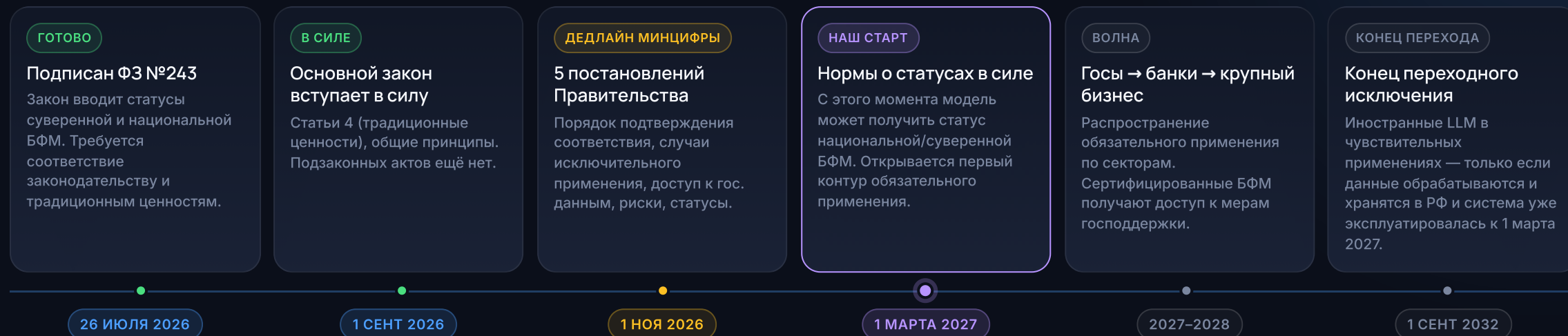
 Риски и митигация

 Go-to-market: beachhead

→ 5 слайдов раздела

Расписание: от подписания до конца переходного режима — 6 лет

Закон подписан 26 июля 2026. Дальше — конкретные даты.



ОКНО ВОЗМОЖНОСТЕЙ · 6,5 МЕСЯЦЕВ

Между 1 ноября 2026 (процедура готова) и 1 марта 2027 (можно подавать на статус) — наш спринт сертификации. За эти полгода: пройти сертификацию, поставить пилот у первого заказчика, доказать ТСО.

5 BigTech уже заняли «on-prem AI platform». Наша позиция — выше.

BigTech продаёт платформу, на которой клиент строит. Мы продаём готовую фабрику результата.

ПРЯМОЙ КОНКУРЕНТ #1 · МТС

MWS AI Agents Platform

6,9 млрд ₽ выручка 2025 (816 млн ₽ — Agents Platform)

Модели: Cotyре = Qwen 2.5. >95% внедрений — в закрытом контуре. Visual workflow, дообучение, multi-model. Поддержка Qwen, DeepSeek, Cotyре, ChatGPT, Claude.

ПРОДАЁТ ПЛАТФОРМУ, НЕ РЕЗУЛЬТАТ

ПРЯМОЙ КОНКУРЕНТ #2 · ЯНДЕКС

Yandex AI Studio

~2 млрд ₽ · 51 тыс. клиентов Yandex Cloud (2025)

On-premises с моделями по выбору, Workflows, MCP Gateway, Audit Trails, batch inference. Full-stack on-prem infrastructure вокруг Stackland.

ГОРИЗОНТАЛЬНАЯ ПЛАТФОРМА, НЕ ВЕРТИКАЛЬНАЯ

НЕДООЦЕНЁННЫЙ КОНКУРЕНТ

Just AI Agent Platform Enterprise

от 100 тыс. ₽/мес Enterprise Cloud

On-prem, 152-ФЗ, локальные + внешние LLM, MCP, RAG, action/chat/voice agents. Self-hosted дистрибутив (апрель 2026) — без обязательной связи с облаком Just AI.

«СКАЧАЙ И ПОСТАВЬ САМ» УЖЕ НЕ УНИКАЛЬНО

ИНФРА + ЭКОСИСТЕМА

Cloud.ru AI Bundle · Сбер GigaChain

5,3 млн ₽/мес тариф H6X 8xH100 (Cloud.ru)

Cloud.ru: Evolution AI Factory + AI Bundle (полный цикл в контуре). Сбер: GigaChain SDK + MCP, центральный участник Альянса в сфере ИИ и MERA.

ИНФРАСТРУКТУРА, НЕ ПРОДУКТ

НЕВИДИМЫЙ КОНКУРЕНТ

Dify · n8n · Open WebUI · LangGraph · vLLM · llama.cpp

60-70% горизонтального стека за несколько недель

Внутренний AI-инженер клиента собирает большую часть продукта на open-source. Главная угроза — не BigTech, а competent customer.

МЫ · НАША НИША

Автономная фабрика законченной работы

₽ / принятая работа

Не мгновенный чат, а длительные фоновые процессы, где стоимость вычисления важнее latency. Customer-owned open stack: код + веса + evals + pipeline + BOM. Model pyramid: маленькие модели для фильтрации + 30-70B

10-100x

ЭКОНОМИКА

Reference BOM против тарифов BigTech: 10–100× меньше дорогой работы

Архитектура делает в 10–100 раз меньше дорогой работы — это гипотеза, которую докажет августовский benchmark.

БЛОК 1 · REFERENCE BOM

BUDGET

5-10 млн ₹

CPU + минимальная GPU. Базовый tier для пилота.

STANDARD

30 млн ₹

CPU + средняя GPU + KV cache. Production tier.

BURST

60 млн ₹

Высоконагруженная система со множеством задач

БЛОК 2 · СРАВНЕНИЕ С BIGTECH ТАРИФАМИ

ИГРОК	ТАРИФ	ЧТО ЗА ДЕНЬГИ
Cloud.ru (HGX 8×H100)	5,3 млн ₹/мес	голая инфра
Selectel (Qwen3 235B / 4×H100)	1,3 млн ₹/мес	inference на чужой модели
Just AI Enterprise Cloud	от 100 тыс. ₹/мес	платформа + lock-in (не on-prem)
Мы (Reference BOM)	20-30 млн ₹ единоразово	owned стек + 0 ₹/мес за inference

Сервер за 50–150 млн ₹ рынком не подтверждается. Наши цифры — на benchmark Selectel.

БЛОК 3 · ЮНИТ-ЭКОНОМИКА ПИЛОТА

ВНЕДРЕНИЕ

РЫНОК ЭКОНОМИКА

1 млн ₹

ПОДПИСКА

420 тыс. ₹/мес

КЕЙС

18 мин ручной работы на заказ → 4 мин; 7 сотрудников → 4;
12% ручной перепроверки → 3%

Beachhead: Regulatory Change Factory для юр. функции крупной компании

Один процесс. Один пилот. Одна цена. Один подписанный результат наутро.

ПЕРВЫЙ ПРОДУКТ BEACHHEAD

Regulatory Change Factory

- **вход:** новый нормативный акт от регулятора
- **ПРОЦЕСС:** корпус 1 000–100 000 внутренних документов → индексация → триаж → поиск затронутых положений → глубокая проверка только спорных случаев → draft новой редакции → перекрёстная AI-проверка вторым агентом → очередь юристу
- **выход:** 137 решений на подпись вместо 40 000 документов к ручному разбору
- **МЕТРИКА:** recall затронутых документов, false negatives, процент принятых юристом изменений, время юриста на 1 000 документов, ₽ за corpus-run

«Сколько человеко-дней после нового требования у вас уходит на impact analysis? Дайте нам закрытый исторический кейс. Мы прогоняем его на нашем сервере и показываем recall, false negatives, часы юриста и цену одного утверждённого изменения.»

ШАГ 1 · АВГ 2026

Экономический benchmark

1 нормативный документ × 1–10 тыс. внутренних документов × 3 класса моделей × CPU/RAM + дешёвая GPU + арендованный сильный GPU. Заморозить первый reference BOM.

ШАГ 2 · СЕН–ОКТ 2026

Законченный продукт на чужой модели

Pipeline upload/new-law → corpus scan → candidates → deep reasoning → proposed changes → validation → human review → export/commit → immutable audit log. Переживает reboot, очередь, паузу 12 часов.

ШАГ 3 · НОЯ 2026

Первый настоящий заказчик

данные клиента → его контур → его конкретный процесс → measurable acceptance criteria. Installer: OS → inference → model → orchestration → identity → MCP → storage → monitoring → audit.

ШАГ 4 · ДЕК 2026

Переносимость

второй заказчик ставится воспроизводимой установкой, не «руками две недели». 80% системы одинаково у клиента А и В.

ШАГ 5 · ДО 1 МАРТА 2027

Доказать TCO

2 контура · 1 vertical product · 1 reference hardware BOM · 1 reproducible model · 1 eval suite · 1 compliance dossier · 1 installer · 1 documented DR · 1 measurable TCO case.

Что может пойти не так — и что мы делаем сейчас

Я признаю каждый риск. У каждого — конкретная митигация.

⚠️ РИСКИ · TOP 5

РИСК 01 · ЭКОНОМИКА

Экономический тезис о дешёвом железе пока гипотеза

Архитектура model-pyramid обещает 10–100× меньше дорогой работы. Без benchmark это

РИСК 02 · СЕРТИФИКАЦИЯ

Процедура может стать непроходимой

Дискреция Правительства огромна. Минцифры может в ноябре выпустить «тест традиционных

РИСК 03 · BIGTECH

BigTech может скопировать подход

MWS/Yandex/Just AI могут за 6 месяцев повторить автономную фабрику. Их distro + enterprise

РИСК 04 · SALES CYCLE

Длинный enterprise sales в РФ

Госсектор и enterprise покупают 6–18 месяцев. Можно умереть от кассового разрыва до первого

РИСК 05 · КОМАНДА

Критический дефицит Inference & Hardware Lead

Без человека, способного положить на стол «сервер X за 6,8 млн ₽ даёт N completed cases/

🟢 МИТИГАЦИЯ · 1:1

МИТИГАЦИЯ · R1

Августовский benchmark

1 нормативный документ × 1–10 тыс. документов × 3 класса моделей × CPU/RAM +

МИТИГАЦИЯ · R2

Dossier 243-ФЗ + диалог с Минцифры

Юрист параллельно ведёт dossier под №243-ФЗ. Национальная модель — это fine-tune открытой

МИТИГАЦИЯ · R3

Vertical workflow + customer ownership

BigTech оптимизирует под массу. Мы — под vertical-specific outcome. Код, веса, evals, BOM

МИТИГАЦИЯ · R4

Продавай результат, не платформу

Пилот за 0,5–3 млн ₽ с заранее определёнными KPI. 2–3 design partners вместо широкого

МИТИГАЦИЯ · R5

Критический найм в первые 4 недели

Профиль: CPU/GPU / quantization / NUMA / vLLM / llama.cpp / SGLang / server BOM / performance.

Кто в команде, кого критически не хватает, что в следующие 90 дней

Текущая конфигурация + 3 критические вакансии + конкретные milestones.

Текущая конфигурация

ORG CHART

ЕСТЬ · 2 / 5

- Никита — Product Architecture / Agents / Runtime / Distribution
- Григорий — AI Engineering / Product Vertical / Model Pipeline / Evals

КРИТИЧЕСКИЙ НАЙМ · 3 / 5

- Inference & Hardware Lead (CPU/GPU / Quantization / NUMA / vLLM / llama.cpp / SGLang / Server BOM / Performance)САМЫЙ СРОЧНЫЙ
- Enterprise Founder/Sales (Design Partners / Procurement / Paid Deployment)
- Legal & AI Compliance (243-ФЗ / Licenses / Data / National-BFM Dossier)

«Самый срочный дефицит — не ещё один универсальный AI engineer. Это человек, который способен через неделю положить на стол: «сервер X за 6,8 млн ₽ даёт N completed cases/сутки». Без него центральный тезис о дешёвом железе нельзя превратить из идеи в moat.»

30 ДНЕЙ · ДО 15 СЕН

Экономический benchmark + Reference BOM

- Августовский benchmark: 1 нормативный документ × 1–10 тыс. документов × 3 класса моделей
- Заморозить Reference BOM Budget ≤5 млн / Standard ≤10 млн
- Первый установленный контакт с design partner (юр. функция крупной компании)
- Найм Inference & Hardware Lead (offer sent)

60 ДНЕЙ · ДО 15 ОКТ

Законченный продукт на чужой модели

- Pipeline end-to-end: upload/new-law → corpus scan → candidates → deep reasoning → proposed changes → validation → human review → export/commit → immutable audit log
- Pipeline переживает reboot, недоступный tool, очередь, паузу 12 часов
- Юрист начинает вести dossier под №243-ФЗ

90 ДНЕЙ · ДО 15 НОЯ

Первый настоящий заказчик + модель

- Installer: OS → inference → model → orchestration → identity → MCP → storage → monitoring → backups → audit → application
- Первый настоящий deployment в контуре клиента с measurable acceptance criteria
- Собственный model pipeline: 2–3 открытые модели → dataset versioning → training → merge →

→

ЧТО МЫ ИЩЕМ

Мы не просим денег. Мы ищем 4 конкретные вещи.

Этот слайд — обращение к советникам. Не к инвесторам.



Контакты в Минцифры

Нам нужен диалог о будущем порядке подтверждения соответствия (243-ФЗ, постановление №3 из 5). Не лоббизм — обмен аналитикой. Это сократит неопределённость по процедуре сертификации в ноябре 2026.

ASK 01 · R2 → MITIGATION



2–3 пилотных клиента

Юр. функция крупной компании, где impact-analysis новых нормативных актов — реальная боль. Банки, госкомпании, large enterprise. Закрытый исторический кейс → августовский benchmark → LOI к 1 октября.

ASK 02 · R4 → MITIGATION



Inference & Hardware Lead

Профиль: vLLM/llama.cpp/SGLang/quantization/server BOM/performance. Не «ещё один AI engineer». Сильный советник может порекомендовать из своей сети за 2 недели. Без этого найма тезис о дешёвом железе — обещание, не moat.

ASK 03 · R5 → MITIGATION



Дверь в Selectel / Cloud.ru как reference hardware partner

Мы не хотим конкурировать с инфра-провайдерами. Мы хотим быть их vertical product на их железе. Это даёт нам стандартный BOM, им — новый use case для GPU-тарифов.

ASK 04 · R1 → MITIGATION



Итог: казино становится расписанием, когда у тебя есть дата, конкурент и метрика.

Дата — 1 марта 2027. Конкурент — BigTech с горизонтальной платформой. Метрика — ₽ за принятую работу. Дальше — исполнение.