Коммерческое обоснование внедрения корпоративного ИИ Comindware

Методология, экономика и комплаенс внедрения корпоративного GenAI

Апрель 2026

Оглавление

1. Введение

Отчёт документов регламентирует жизненный цикл корпоративного ИИ в резидентном контуре РФ:

- **организация внедрения и эксплуатации** — фазы, роли, KPI;
- **диапазоны CapEx/OpEx/TCO** — ориентиры для бюджетирования;
- **передача кода и ИС** — комплект KT/IP и приёмка;
- **архитектурный профиль Comindware** — границы и состав поставки;
- **риски и контроль** — критические узлы до промышленного запуска.

Методологическая основа: верифицированные рыночные котировки, международные стандарты (NIST, ISO) и операционная практика **Comindware**.

Материалы предназначены для формирования офферов и защиты инвестиционных решений.

1.1 Глоссарий

Глоссарий обеспечивает единство интерпретации архитектурных и экономических параметров проекта.

Корпоративный RAG-контур, **сервер инференса на базе vLLM/MOSEC** и **агентный слой Comindware Platform** — условные названия компонентов иллюстративного референс-стека **Comindware**, а не коммерческие SKU.

Договорные формулировки и точные границы — в приложениях:

- *Приложение A «Отчуждение ИС и кода: KT, IP, лицензии, критерии приёмки передачи*
- *Приложение B «Корпоративный ИИ Comindware: состав стека, границы, артефакты»*.

Термин	Определение
ADR (Architecture Decision Record)	Документально закрепленное архитектурное решение с обоснованием компромиссов и ограничений.
AgentOps (Agent Operations)	Методология промышленной эксплуатации ИИ-агентов в непрерывном замкнутом цикле: мониторинг, оценка, совершенствование.
AI TRiSM (AI Trust, Risk and Security Management)	Комплексная модель управления доверием, рисками и безопасностью ИИ-систем.
Arize Phoenix	Открытый инструмент наблюдаемости и экспериментов для LLM/RAG: трассировка, дашборды, оценка качества при self-hosted размещении. В референс-стеке **Comindware** — слой рядом с OpenTelemetry/OpenInference; не заменяет инфраструктурный мониторинг.
ASR (Automatic Speech Recognition)	Распознавание речи: преобразование речевого сигнала в текст (голосовой ввод, колл-центры, мультимодальные сценарии).
BOT (Build–Operate–Transfer)	Модель передачи актива: интегратор строит ИИ-контур, вводит в эксплуатацию, затем передаёт заказчику.
CapEx	Капитальные затраты: оборудование, лицензии и ввод в эксплуатацию.
CLI (Command Line Interface)	Интерфейс командной строки для администрирования серверов инференса: запуск, остановка, проверка статуса, тестирование моделей.
DPA (Data Processing Agreement)	Договор об обработке данных между сторонами, где одна передаёт данные другой для обработки (типичный ориентир в практике GDPR). В РФ переносится на роли оператора, поручения обработки ПДн и цепочку субобработчиков по 152-ФЗ и договору.
DSPy	Открытый фреймворк декларативной сборки и настройки LLM-конвейеров (модули, сигнатуры, оптимизация промптов и обучающих примеров). К тому же классу относятся другие библиотеки программной сборки промптов и контрактов вывода; в отчёте DSPy служит ориентиром из открытых учебных материалов, а не фиксированным стеком поставки.
ETL (Extract, Transform, Load)	Процесс извлечения данных из источников, их преобразования и загрузки в целевую систему. Включает подготовку и обогащение, очистку, разметку и структурирование данных.
FinOps (Financial Operations)	Подход к управлению облачными и ИИ-затратами через прозрачные метрики потребления и аллокацию затрат.
GenAI (Generative AI)	Генеративный ИИ — модели, которые создают текст, код и иные артефакты.
IP (Intellectual Property)	Интеллектуальная собственность: код, артефакты, модели, документация, права использования и условия передачи.

Термин	Определение
KT (Knowledge Transfer)	Передача знаний, эксплуатационных регламентов и обучения команде заказчика.
KV-кэш	Кэш пар «ключ–значение» для промежуточных состояний внимания при автогрессивной генерации (англ. KV cache); вместе с весами модели и батчем задаёт основную нагрузку на VRAM при длинном контексте.
LLM	Большая языковая модель.
LLM-судья (LLM-as-a-judge)	Подход, при котором отдельная модель используется как судья по заранее заданной рубрике.
LLMOps	Практики эксплуатации LLM-контуров: релизы, мониторинг, стоимость, качество и инциденты.
LoRA (Low-Rank Adaptation)	Адаптация модели небольшим числом параметров низкого ранга без полного дообучения; дешевле по памяти GPU и хранению. В отчёте упоминается в исследовательских и продуктовых контекстах (Doc-to-LoRA, подходы к «забыванию» весов).
MCP (Model Context Protocol)	Протокол подключения инструментов и внешних ресурсов к агенту через явные серверы и контракты вызова.
MERA / RAGAS / DeepEval	Контуры и фреймворки оценки качества RAG/LLM.
ML	Классическое машинное обучение: модели классификации, прогнозирования и ранжирования без генерации контента. Отличается от GenAI.
ModelOps	Практики управления жизненным циклом моделей как производственных компонентов: версии, выкаты, контроль качества и сопровождение.
MOSEC	Связка фреймворка Mosec и служебного слоя **Comindware** — HTTP-сервис вспомогательных моделей: эмбеддеров, ранжировщика и защитных моделей.
NIST AI RMF (AI Risk Management Framework)	Методология оценки рисков ИИ от NIST — используется как методологический ориентир, а не как замена нормам РФ.
On-prem (On-premise)	Размещение в собственном или выделенном контуре заказчика, а не в публичном управляемом API.
OpEx	Операционные затраты: эксплуатация, сопровождение, мониторинг и сервисы.
OpenInference	Открытый набор соглашений и плагинов для OpenTelemetry-совместимого инструментирования GenAI-приложений; поддерживается в Arize Phoenix.
PoC (Proof of Concept)	Короткий этап проверки гипотезы до пилота или масштабирования.

Термин	Определение
RAG (Retrieval Augmented Generation)	Генерация ответа с опорой на предварительный поиск по документам, данным или базе знаний.
SGLang	Открытый фреймворк высокопроизводительного инференса LLM; в отчёте — вариант движка наряду с vLLM.
SGR (Schema-Guided Reasoning)	Структурированное рассуждение по схеме — техника принудительного структурирования рассуждений LLM через предопределённые схемы. По отраслевым бенчмаркам даёт 5–10% прирост точности против неструктурированных промптов; обеспечивает воспроизводимое рассуждение и пошаговый аудит (*Schema-Guided Reasoning (SGR)*). В Comindware применяется в нескольких точках конвейера: анализ запросов, критика ответов агентов, планирование после фазы защиты и детерминированное управление выводом.
SLA (Service Level Agreement)	Обещанный уровень сервиса для заказчика.
SLM (Small Language Model)	Малая языковая модель для дешёвых или быстрых сценариев, работает аналогично LLM.
SLO (Service Level Objective)	Внутренняя целевая метрика качества сервиса.
SCQA (Situation-Complication-Question-Answer)	Метод структурирования управленческих решений: ситуация → вызов → задача → решение.
TCO (Total Cost of Ownership)	Совокупная стоимость владения решением на горизонте нескольких лет.
TOM (Target Operating Model)	Целевая операционная модель: роли, процессы, метрики и контуры ответственности.
TOON	Компактный формат структурированных данных, применяемый для снижения токеновых затрат относительно JSON.
TTS (Text-to-Speech)	Синтез речи: преобразование текста в речь (голосовой ответ ассистента, озвучивание, IVR).
vLLM	Движок для промышленного инференса LLM через OpenAI-совместимый API; включает конфигурацию, проверки доступности, эксплуатационный регламент.
Агентный RAG	Вариант RAG, где модель планирует шаги, вызывает инструменты и делает несколько итераций поиска и проверки.
Агентный слой Comindware Platform	Сценарии, где модель не только отвечает на вопросы, а инициирует действия в платформе через разрешённые инструменты и API.
Агенты для программирования (coding agents)	ИИ-агенты и среды (IDE, CLI, песочницы), автоматизирующие цикл разработки: правки кода, тесты, ревью, интеграция с PR/CI.

Термин	Определение
Батч (Batch)	Группа запросов, обрабатываемая за один цикл; размер батча влияет на расход VRAM и задержку.
Выборка (Sampling)	Отбор части журналов и трасс для хранения, чтобы ограничить объём и стоимость телеметрии без потери значимых сигналов.
Вызов инструментов (Tool calling / Function calling)	Способ работы модели, при котором она по контракту инициирует вызов внешнего инструмента, API или функции и использует результат в ответе.
Глубокое исследование (Deep research)	Многошаговый поиск в различных источниках со сверкой результатов и аналитической сборкой выводов.
Допустимая агентность	Границы действий и полномочий агента; выход за границы требует эскалации или блокируется политикой.
Доля автономного завершения	Доля запросов, выполненных без вмешательства человека. Ключевая метрика автономности агента.
Временный привилегированный доступ (Just-in-Time access)	Предоставление агенту или пользователю прав исключительно на период выполнения конкретной задачи с автоматическим отзывом по завершении. Устраняет постоянные привилегии (standing privileges), сужая окно компрометации до минимума.
Защитные механизмы (Guardrails)	Политики, фильтры, валидации и ограничения вокруг модели и инструментов, снижающие риск небезопасных действий.
Извлечение контекста (Retrieval)	Этап RAG-контура, в котором система находит и отбирает релевантные фрагменты документов, записей или иных данных по запросу пользователя для последующей генерации ответа.
Комплаенс (Compliance)	Соответствие нормам, договорам и политикам (регуляторика, ИБ, персональные данные).
Контур оценки качества	Наборы тестов, критерии, метрики и регрессионные проверки, которые позволяют отслеживать деградацию после изменений модели, индекса или промпта. Включает офлайн- и онлайн-оценку качества (см. отдельные строки ниже).
Корпоративный RAG-контур	Референсный контур ассистента **Comindware** с поиском по корпоративным данным и генерацией ответа.
Межагентная задержка	Задержка при передаче задачи между агентами в мультиагентных системах. Ключевая метрика для оценки производительности рабочего процесса.
Наблюдаемость (observability)	Возможность разбирать контур по трассам, метрикам, журналам и событиям.
Онлайн-оценка качества	Оценка ответов и траекторий на живом трафике; требует политики телеметрии, выборки и ПДн.
Открытые веса модели (Open weights)	Опубликованные веса модели, которые можно развернуть в своём контуре; это не тождественно отсутствию лицензионных ограничений.

Термин	Определение
Офлайн-оценка качества	Проверки на фиксированных наборах и рубриках до вывода в продакшн.
Пакетная обработка (Batching)	Совместная подача нескольких запросов в инференс (очередь, групповой проход) для загрузки GPU и пропускной способности.
Ретенция данных (Retention)	Установленный срок и правила хранения журналов, трасс, метрик и других артефактов наблюдаемости, после которого данные удаляются, архивируются или переводятся в более дешёвое хранилище.
Red Teaming	Моделирование атак специалистами по безопасности для выявления уязвимостей ИИ-контура: промпт-инъекции, обход защит, проверка границ агентности.
Скорость улучшений	Частота внедрения оптимизаций в неделю (промпты, извлечение, корректировка потока). Индикатор зрелости процесса непрерывного совершенствования.
Тензорный параллелизм (Tensor Parallelism, TP)	Распределение вычислений и фрагментов весов модели по нескольким GPU. Снижает объём памяти на одно устройство посредством обмена между картами.
Фактор автобуса (Bus factor)	Степень зависимости проекта или контура от ограниченного числа носителей критичных знаний; чем он ниже, тем выше риск остановки или деградации после выбытия ключевых сотрудников.
Человек в контуре (Human-in-the-loop, HITL)	Подход, при котором критические решения и спорные ответы модели проходят проверку силами человека.
Эксплуатационный регламент (Runbook)	Описание штатной эксплуатации, инцидентов, проверок и действий сопровождения.
Эффективность токенов промпта	Соотношение качества вывода к числу входных токенов. Показывает возможность получения того же качества с меньшим числом токенов (экономия до 39%).

1.2 Навигация: вопрос → документ

1.2.1 Коммерция и экономика

Вопрос	Документ
Быстрый старт: создание ассистента с доступом к знаниям и интеграцией с Comindware Platform	*Быстрый старт*
Коммерческий обзор для руководителей: типовые пакеты, что остаётся у заказчика, матрица аргументов по ЛПР	*Стратегическое резюме*
KPI, числовые пороги go/no-go, политика интерпретации	*Стратегическое резюме: числовые пороги*; *Методология: процессы и KPI*
CapEx/OpEx/TCO — цифры и диапазоны для клиента	*Сайзинг и экономика*
Расчёт расхода токенов (портал поддержки)	*Сайзинг: токены*
Китайские альтернативные GPU для инференса (Ascend, Moore Threads, Cambricon, MetaX)	*Приложение E*

1.2.2 Методология внедрения

Вопрос	Документ
Внедрение в пром контуре: роли, фазы, контрольные точки качества	*Методология*
Где формируется преимущество в корпоративном ИИ (данные, семантика, агенты)	*Методология*
Глобальные бенчмарки OpenAI 2025 и оговорки по выборке	*Методология: эмпирика*
Стратегия внедрения, организационная зрелость, пилот vs scale, обучение (СКОЛКОВО)	*Методология: стратегия*
Бизнес-процессы для КТ (BPMN 2.0, LLM-генерация)	*Методология: BPMN*
SGR в практике Comindware	*Методология: SGR*
GenAI в маркетинге крупных брендов РФ (опрос CMO, red_mad_robot, 2025)	*Сайзинг: зрелость рынка*; *Методология: GenAI в маркетинге*
Российский рынок GenAI: сегменты, прогноз до 2030	*Методология: карта рынка*; *Сайзинг: статистика*

1.2.3 Передача и текущий стек

Вопрос	Документ
Комплект отчуждения ИС/кода (KT/IP)	*Приложение A*
Бизнес-процессы для KT (минимальный комплект)	*Приложение A: отчуждение*
Состав стека Comindware («что есть» vs «методология»)	*Приложение B*
Возможности агентов (RAG, MCP, SGR, индексация)	*Приложение B: арсенал*
Ассистент аналитика (49 инструментов, 6 провайдеров)	*Приложение B: аналитик*
Фреймворки инференса (MOSEC, vLLM, Infinity)	*Приложение B: инференс*

1.2.4 Безопасность и наблюдаемость

Вопрос	Документ
Безопасность, комплаенс, наблюдаемость	*Приложение C*
Наблюдаемость GenAI в РФ: локализация, self-hosted телеметрия	*Приложение C: наблюдаемость LLM*
Изоляция и сеть для агентского исполнения (граница доверия, egress)	*Приложение C: граница доверия*
Паттерны среды для агента, модель риска, минимальный состав платформы	*Приложение C: модель риска*
Сравнение песочниц E2B / Modal / Daytona	*Приложение C: песочницы*; *Методология: бенчмарки*
Безопасный MVP контура агента за ~30 дней	*Приложение C: MVP*; *Методология: MVP*
Поведенческие риски	*Приложение C: риски*
AI TRiSM: управление доверием и рисками	*Приложение C: AI TRiSM*

1.2.5 Справочные документы

Вопрос	Документ
Сжатый обзор для руководителей	Стратегическое резюме
Бюджетный риск и организационная зрелость	*Сайзинг: риски*
Shadow GenAI и маршрутизация моделей в маркетинге	*Приложение A: Shadow GenAI*
Артефакты PR-веток для агентного контура	*Приложение A: PR-артефакты*

2. Быстрый старт: создание агента с доступом к знаниям и Comindware Platform

2.1 Резюме

- **Ситуация:** заказчику требуется переход от чат-ботов к интеллектуальным ассистентам, глубоко интегрированным с **Comindware Platform** и способным работать с любыми источниками знаний — от внутренних регламентов до внешних реестров (ФНС, 1С, Консультант+).
- **Вызов:** RAG по своей сути — это не только поиск по базам знаний, но и интеграция с любыми корпусами данных. Отсутствие чёткой методологии внедрения RAG-ассистентов и понимания затрат тормозит масштабирование.
- **Задача:** собрать и запустить промышленного ассистента, интегрированного с платформой **Comindware**, обеспечив при этом наблюдаемость, безопасность и экономическую эффективность.
- **Решение:** использовать модульную архитектуру на базе референс-стека **Comindware**. Начать с подготовки данных и бизнес-анализа, последовательно внедрить компоненты поиска и извлечения данных, инференса и интеграции с платформой.

2.2 Необходимые компоненты

Ассистент с доступом к знаниям и интеграцией с **Comindware Platform** строится на трёх слоях:

- **Слой данных:** подготовка, очистка и обогащение корпусов знаний (внутренние регламенты, Wiki, Jira, внешние реестры). Включает ETL-конвейеры, инкрементную индексацию и векторизацию. Получение данных из внешних источников (ФНС, Консультант+, 1С) через API-коннекторы и преобразование в формат агента.
- **Слой агентской логики:** агенты, работающие с **Comindware Platform** через API и MCP. Агентный слой взаимодействует с данными и моделями платформы — читает и записывает сущности, создаёт записи, формирует резолюции, строит модели данных. Компоненты **агентного слоя Comindware** можно использовать как навыки (skills) для агентов общего назначения (например, OpenCode) или подключать по протоколу MCP — это позволяет интегрировать агентный слой в любой ИИ-ассистент, поддерживающий вызов внешних инструментов. Слой отвечает за асинхронную интеграцию с **Comindware Platform** (агент самостоятельно возвращает результат в платформу без действий с её стороны).

- **Слой интеллектуальной обработки:** гибридный поиск (векторный + ключевые слова), работа с документами в файловых хранилищах, промышленный инференс, наблюдаемость и безопасность. Включает глубокий веб-поиск и генерацию кода в песочнице для обработки данных (в том числе автогенерацию SQL-запросов к источникам).

2.2.1 Инвестиции в разработку (трудозатраты)

Фактически затраченные ресурсы на создание ядра стека:

Компонент	Затрачено времени (человеко-час.)
Агентный слой Comindware Platform	**320+**
RAG-движок	**900+**
Итого инвестировано в стек	**1220+**

При внедрении у заказчика эти инвестиции не оплачиваются повторно: заказчик оплачивает только адаптацию готового стека под конкретные сценарии и бизнес-процессы (см. *«Ресурсы и трудозатраты на сценарий»*).

Возможности агентного слоя Comindware

Агентный слой умеет читать почти все сущности платформы, включая формы и схемы, но не модифицирует и не строит визуальные компоненты (формы, таблицы, BPMN-схемы) — он работает с данными и моделями данных, редактирует записи, формирует резолюции и выполняет глубокий веб-поиск. Реализовано около 50 инструментов из необходимых 100 для полного покрытия сущностей платформы.

2.3 Архитектура решения

Референс-стек **Comindware** включает:

- **RAG-движок:** ядро системы для многошагового извлечения контекста, приоритизации данных и стабильной индексации с защитой от дублей.
- **Серверы инференса:** высокопроизводительный слой для основных языковых моделей и вспомогательных моделей (эмбеддеры, ранжировщики, классификаторы ПДн, анонимизаторы), защитных моделей (фильтрация вредоносных запросов, спама, джейлбрейка).
- **Наблюдаемость:** система трассировки запросов, оценки качества ответов и поиска «провалов» в базе знаний:

- **Arize Phoenix** — трассировка, дашборды и оценка качества LLM/RAG: детализированная трассировка по всей длине агентского цикла, эксперименты и оценки — из коробки, без написания собственного кода мониторинга.
- **HiveTrace** — защита: обнаружение промпт-инъекций, блокировка джейлбрейка и утечки данных в реальном времени.

Phoenix и HiveTrace дополняют друг друга: Phoenix отвечает за качество и отладку, HiveTrace — за безопасность. Оба слоя трассируют токены, цены и метаданные и передают данные в платформу или внешние системы наблюдаемости.

Подробнее — в *Приложении С «Безопасность, комплаенс, наблюдаемость»*.

- **Безопасность:** контур модерации на базе защитных моделей. Классифицирует запросы и ответы на вредоносность, спам, джейлбрейк, утечки ПДн и соответствие политикам компании.

Текущее состояние — PoC/MVP

Компоненты **агентного слоя Comindware** и **серверы локального инференса** развёрнуты в среде разработки без продуктовых артефактов и выпуска версий.

Создание Docker-контейнеров и Kubernetes-манифестов для промышленной эксплуатации — в планах.

2.4 RAG — универсальный ETL-конвейер данных

Бизнес-данные — главный источник добавленной стоимости ИИ-решений.

Агент без доступа к актуальным данным превращается в чат-пустышку с данными, полученными при обучении модели годичной и более давности.

Качество ответов агента на 80% зависит не от выбора языковой модели, а от качества подготовки данных, контекст-инжиниринга, архитектуры и бизнес-логики агента.

Поэтому этап ETL (Extract, Transform, Load — извлечение, преобразование, загрузка) и интеграции с внешними системами занимает львиную долю бюджета и трудозатрат проекта и определяет его успех.

⚠ Стройте архитектуру ETL в первую очередь

Начинайте ETL задолго до создания любых ИИ-агентов — фокус на глубоком бизнес-анализе состава, характера и потоков данных.

Только после формирования архитектуры данных и правил извлечения, очистки, обогащения и обновления (инкрементные конвейеры) переходите к технической реализации.

2.4.1 Анализ и планирование конвейеров данных

Перед индексацией или извлечением данных команда проводит:

- **Инвентаризация датасетов:** определение корпоративных наборов данных (Wiki, регламенты, прошлые решения поддержки, базы 1С) и их предварительная оценка — валидация достоверности и разметка критериев «хорошо/плохо» для каждого кейса и бизнеса заказчика. Эти наборы становятся основой для системных и промежуточных промптов и контекстного инжиниринга.
- **Инвентаризация динамических источников:** помимо статичных документов, RAG-контур получает динамические данные через API:
  - **ФНС (ЕГРЮЛ/ИП, РНП):** проверка статуса контрагентов, получение выписок и мониторинг изменений в реестре.
  - **1С (Предприятие 8.3+):** выгрузка данных для интеграции со складскими остатками, заказами и справочниками контрагентов.
  - **Консультант+/Гарант:** получение актуальных редакций документов и судебной практики из справочно-правовых систем.
- **Семантический аудит:** анализ терминологии, синонимов и специфического сленга компании.
- **Проектирование трансформаций:** правила очистки (удаление устаревших версий, очистка от мета-мусора) и обогащения данных (автоматическое добавление тегов категорий, связей с продуктами).
- **Выбор подхода к индексации:** файловый поиск (агент ищет документы в файловой системе), гибридный поиск (по векторам и ключевым словам), API (извлечение из внешних систем по запросу).
  - Файловый поиск — для часто меняющихся или добавляемых корпусов знаний в виде файлов.
  - Гибридный поиск — для корпусов, где новые знания добавляются, а старые остаются стабильными.

- Извлечение по API — для динамических источников (ФНС, 1С, Консультант+).

2.5 Интеграция с **Comindware Platform**

Ключевая особенность решения — **асинхронное взаимодействие** агента с **Comindware Platform**.

- **YAML-схемы:** логика взаимодействия и сопоставление сущностей агента и платформы — в декларативных конфигурационных файлах. Это позволяет менять структуру полей в приложениях без изменения кода агента.
- **Асинхронный вызов:** платформа инициирует задачу и продолжает работу. Агент в фоновом режиме собирает данные, рассуждает, формирует и проверяет выводы, затем записывает результат обратно в платформу, создаёт необходимые записи и выставляет статусы.

При построении архитектуры важно **определить границы:** что выполняет **Comindware Platform** по своим сценариям и схемам, а что делегируется агенту.

- **Comindware Platform** обрабатывает рутинные операции и предсказуемые маршруты (валидация данных, маршрутизация заявок, вычисления по формулам), снижая вычислительную нагрузку и трафик между агентом, внешними источниками и платформой.
- Агент вступает в работу там, где нужно рассуждение, недетерминированный или глубокий поиск и анализ данных, сложные непредсказуемые цепочки обращений к внешним источникам.

2.6 Этапы реализации

Фазы внедрения соответствуют *Методологии разработки и внедрения ИИ*.

От идеи к продакшн-агенту:

- **Подготовка данных и ETL:** сбор, очистка и обогащение данных из всех источников (внутренние регламенты, Wiki, Jira, внешние реестры). Качество ответов агента на 80% зависит от данных — без них даже мощная модель даёт нерелевантные результаты.
- **Архитектура и проектирование:** определение «границ агентности» — что агент делает сам, а где нужна эскалация к человеку. Балансировка вычислительной нагрузки между **Comindware Platform**, агентом и внешними источниками. Проектирование схемы данных, диаграмм последовательностей и правил взаимодействия с внешними системами (API ФНС, 1С, справочные системы).

- **Настройка поиска и ранжирования:** стратегии извлечения контекста и ранжирования — из сотни фрагментов оставить 5–10 релевантных. Точность ответов растёт на 15–20%, токены экономятся.
- **Безопасность и наблюдаемость:** контуры модерации (фильтрация вредоносных запросов, спам, джейлбрейк, утечки ПДн) и система наблюдаемости (трассировка, оценка качества). Закладываются с первого дня пилота.
- **Инференс и масштабирование:** выбор модели размещения (локально, облако, гибрид) и оптимизация серверов. Квантование позволяет запускать крупные модели на стандартном оборудовании.
- **Интеграция с платформой и оркестрация:** соединение компонентов в единый конвейер, маппинг полей ассистента в атрибуты **Comindware Platform**. Агент работает асинхронно: платформа инициирует задачу, агент выполняет в фоне и записывает результат.
- **Передача знаний и запуск:** эксплуатационные регламенты для команды заказчика — обновление индекса, чтение трассировочных данных, изменение промптов, схем и политик безопасности.

Критерии приёмки и артефакты

Каждый этап имеет критерии приёмки и артефакты.

Не переходите к следующему, пока не утверждены результаты текущего — это снижает риск переделок и перерасхода бюджета.

2.7 Практический сценарий: ассистент по закупкам с проверкой контрагентов

2.7.1 Описание кейса

Бизнес-задача: автоматизировать обработку заявок на закупку с проверкой благонадёжности поставщика по реестрам ФНС, ФАС и открытым источникам.

Алгоритм работы агента:

1. **Триггер:** в **Comindware Platform** создаётся заявка на закупку с ИНН поставщика.
2. **Асинхронный запуск:** платформа отправляет ID заявки агенту. Агент самостоятельно забирает ИНН, товарную категорию, название компании и все необходимые данные.
3. **Поиск в базе знаний:** агент ищет регламенты закупки для данной категории.

4. **Внешняя проверка (вызов инструмента):** агент вызывает инструмент *«Проверка в ФНС»*, передаёт ИНН и получает актуальный статус компании (действующая, в стадии ликвидации).
5. **Углублённый веб-поиск (вызов инструментов):** если статус спорный, агент ищет в интернете (новости, отзывы, арбитражные дела) и других подключённых источниках данных о контрагенте (например, наличие в реестре недобросовестных поставщиков по реестру ФАС, наличие арбитражных дел).
6. **Формирование вывода:** агент собирает полученные факты, например: *«Регламент допускает закупку, ФНС подтверждает статус „Действующая", арбитражных дел нет. Рекомендация: Одобрить»*.
7. **Запись результата:** агент записывает ответ в поле *«Резолюция ИИ»* в заявке и выставляет статус, например *«Проверено: Безопасно»*.

2.7.2 Ресурсы и трудозатраты

Для реализации с нуля до рабочего прототипа на базе готового стека **Comindware** требуются:

- **Бизнес-аналитик (1 FTE):** сбор требований, подготовка тестовых вопросов, разметка данных. **Трудозатраты:** 40–60 человеко-часов.
- **ИИ-инженер / Разработчик (1 FTE):** настройка ETL, тюнинг промптов, интеграция с API ФНС, отладка RAG-конвейера. **Трудозатраты:** 80–120 человеко-часов.
- **DevOps / Системный администратор (0,5 FTE):** развёртывание серверов инференса, систем наблюдаемости и безопасности, настройка GPU. **Трудозатраты:** 20–30 человеко-часов.

Итого: 140–210 человеко-часов (2,5 FTE).

Срок реализации: 6—8 недель до запуска пилотной версии.

2.8 Состав и роли проектной команды

Для внедрения и эксплуатации корпоративного ИИ-агента сформируйте компактную кросс-функциональную команду:

Роль	Ключевые компетенции	Зона ответственности
Аналитик и инженер контекста (ИИ/RAG)	Бизнес-анализ, контекст- и промпт-инжиниринг, работа с данными	Бизнес-логика, тестовые вопросы, оценка качества ответов, разметка данных для RAG
Инженер-разработчик (LLM/Python)	Python, LangChain/ LlamaIndex, SQL, API-интеграции	ETL-конвейеры, инструменты агента, интеграция с **Comindware Platform**, тюнинг поиска
Инженер инфраструктуры (DevOps/LLMOps)	Docker, Kubernetes, мониторинг	Инференс-серверы, GPU-ресурсы, наблюдаемость
Эксперт предметной области	Знание предметной области (юриспруденция, ТП, др.)	Верификация знаний в базе, приёмка ответов на соответствие регламентам
Руководитель проекта	Управление ИТ-проектами, SDLC	Координация фаз (концепт → пилот → масштабирование), ожидания стейкхолдеров, SLA/SLO
Инженер по качеству (QA/Eval)	Тестирование ПО, метрики LLM (RAGAS, DeepEval)	Регрессионное тестирование, проверка на галлюцинации, контроль дрейфа качества

Минимальный состав для старта: 3–4 специалиста. На этапе концепта аналитик, разработчик и инженер могут совмещать роли. Ключевой элемент — **аналитик и инженер контекста**: качество данных и промптов определяет итоговую ценность для бизнеса.

2.9 Экономика и риски: три модели развёртывания

Подробный расчёт факторов стоимости и сценарный сайзинг приведены в разделе *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Выбор модели размещения (облачный API, аренда GPU, локальный On-prem) — это баланс между скоростью запуска, контролем и долгосрочными затратами.

2.9.1 Обзорное сравнение моделей

Ниже — усреднённые ориентиры для сценария «среднее предприятие» при устойчивой нагрузке.

Детальные расчёты — в *«Сайзинг и экономика»*.

Модель	CapEx	OpEx/ год	TCO 1 год	TCO 3 года	Когда выгодно
Облачный API	0	~2,7 млн руб.	~2,7 млн руб.	~8,1 млн руб.	PoC, пилот, переменная нагрузка
Аренда GPU	0	~1,5 млн руб.	~1,5 млн руб.	~4,5 млн руб.	Из пилота в продакшн, гибкость
On-prem	~1,5 млн руб.	~0,8 млн руб.	~2,3 млн руб.	~3,9 млн руб.	Устойчивая нагрузка >60%, суверенитет

Опорные цены

- **Аренда GPU:** 2×RTX 4090 48 ГБ — 100 000—150 000 руб./мес. (1dedic, Cloud.ru).
- **Покупка GPU:** NVIDIA RTX PRO 6000 Blackwell 96 ГБ — ~900 000 руб. (GPU) + сервер ~300 000 руб.
- **Токены API:** усреднённая оценка по тарифам YandexGPT, GigaChat, Cloud.ru (~300 руб./млн токенов).
- **Инфраструктура (on-prem):** электроэнергия, ЦОД, обслуживание — ~300 000 руб./год.
- **Поддержка (on-prem):** базовый LLMOps — ~425 000 руб./год (*Сайзинг и экономика*).

Ключевые выводы:

- **Аренда GPU** — минимальный TCO на горизонте 1–2 лет, оптимально для пилота.
- **On-prem** окупается относительно облачного API при устойчивой утилизации >60% и горизонте >2 лет.
- **Суверенитет данных** (152-ФЗ, КИИ) часто перевешивает чистую экономику — on-prem устраняет утечку запросов за периметр.

2.9.2 Риски внедрения

1. **Деградация качества:** база знаний со временем устаревает, модели дают некорректные ответы. **Меры:** постоянный мониторинг через систему

наблюдаемости и еженедельное регрессионное тестирование на эталонном наборе вопросов.

2. **Зависимость от ключевого специалиста:** риск остановки проекта при уходе единственного инженера. **Меры:** использование стандартизированного стека **Comindware** и ведение журнала архитектурных решений для обеспечения передаваемости знаний.
3. **Отказ инфраструктуры:** дефицит или выход из строя GPU-ресурсов. **Меры:** гибридный инференс с автоматическим переключением на облачного провайдера при сбое.
4. **Дрейф входных данных:** изменение формата входящих данных (новые типы заявок), к которым агент не адаптирован. **Меры:** мониторинг ошибок обработки с автоматизированными оповещениями и отслеживание аномального снижения достоверности ответов.

3. Стратегическое резюме: корпоративный ИИ — внедрение и отчуждение

Comindware обеспечивает системное внедрение, развёртывание и передачу технологий генеративного ИИ в контур заказчика.

Конечный результат — формирование **суверенного цифрового актива** с регламентированной моделью владения, независимого от внешних подписок и проприетарных ограничений.

Целевая траектория: PoC в облачной инфраструктуре РФ → Пилот → Промышленное масштабирование → BOT (Построение — Эксплуатация — Передача). Переход между этапами детерминирован достижением контрольных метрик качества, утилизации и операционной эффективности.

3.1 Решение для руководства за 60 секунд

- **Модель владения:** полная передача интеллектуальной собственности и компетенций (KT/IP/BOT). Заказчик получает автономный актив, а не сервисную зависимость.
- **Управленческий контроль:** каждый этап завершается верификацией KPI и фиксацией фактического экономического эффекта. Масштабирование без подтверждения целевых метрик исключено.
- **Финансовый фундамент:** оценки TCO базируются на актуальных тарифах российских облачных провайдеров и рыночной стоимости GPU-инфраструктуры.
- **Границы применимости:** глобальные корпоративные бенчмарки — сравнительный контекст; норму для резидентного контура РФ определяют 152-ФЗ, локальные тарифы и отдельная правовая оценка.
- **Резолюция:** утвердите план реализации 30/60/90 дней, назначьте ответственных за целевую операционную модель (TOM) и закрепите критерии приёмки активов.

> ⚠ **Единая граница применимости**
>
> В РФ приоритетны 152-ФЗ, резидентность данных и локальные тарифы; зарубежные и глобальные ориентиры используются только как сравнительный бенчмарк.

3.2 Концепция коммерческого решения

- **Ситуация:** рынок GenAI перешел от стадии интереса к фазе массового спроса. Ключевым дефицитом является не сама технология, а методология её промышленного внедрения с соблюдением требований комплаенса и финансовой эффективности.
- **Вызов:** без прозрачной экономики и формализованной передачи знаний (KT) пилотные проекты не трансформируются в устойчивые активы. По данным McKinsey (2025), при охвате ИИ в **88%** организаций, значимый эффект на EBIT отмечают лишь **39%**. Согласно BCG, разрыв между приоритетностью ИИ (**75%**) и реальной ценностью (**25%**) критичен.
- **Задача:** развертывание корпоративной ИИ-инфраструктуры с предсказуемым TCO и гарантированной передачей контроля заказчику.
- **Решение:** реализация поэтапной программы с едиными KPI, детерминированными критериями приёмки и комплексным пакетом отчуждения (KT/IP/BOT).

3.3 Бизнес-ценность Comindware

- **Внедрение:** переход от PoC к промышленному контуру с верифицированными контрольными точками качества и наблюдаемости.
- **Эксплуатация и рост:** целевая операционная модель (TOM), роли, FinOps, оценка качества, регрессионные проверки, управление рисками.
- **Передача владения:** код, конфигурации, эксплуатационный регламент, обучение, критерии приёмки, юридически чистый контур.
- **Суверенный контур РФ:** архитектура и данные проектируются под требования резидентности и комплаенса с первого дня.
- **Доказанная инженерная база:** агентный контур Comindware — работающие компоненты, задокументированные в открытых репозиториях, а не концепт-документы.
- **Платформа, а не чат-бот:** специализированные инструменты для работы с **Comindware Platform** исключают галлюцинации на уровне архитектуры.
- **Отказоустойчивость:** каскад провайдеров LLM со интеллектуальным переключением обеспечивает непрерывность сервиса.
- **Полная наблюдаемость:** сквозное отслеживание каждого шага агента через self-hosted инструменты (Langfuse, Arize Phoenix, OpenTelemetry) — без передачи данных во внешние облака.
- **Гибкость развёртывания:** от изолированного on-prem-контура без выхода в интернет до режима внешнего агентного шлюза (MCP).

3.4 Пакеты по этапам внедрения

3.4.1 Пакет 1. Управленческая диагностика + выбор 2–3 кейсов

Результат: согласованный набор приоритетных сценариев, KPI и ограничений, требования к данным и комплаенсу, согласованная модель принятия решений.

3.4.2 Пакет 2. PoC (2–4 недели)

Результат: рабочий прототип на 1–2 кейсах с первичным измерением эффекта, стоимости и рисков.

3.4.3 Пакет 3. Пилот (1–3 месяца)

Результат: пилот в производственной среде: интеграции, наблюдаемость, первые пользователи, базовые метрики.

3.4.4 Пакет 4. Масштабирование (3–12 месяцев)

Результат: промышленный контур с целевой операционной моделью (роли, процессы, KPI), план сопровождения, контроль качества и дорожная карта расширения.

3.4.5 Пакет 5. Передача / завершение BOT (6–12 месяцев)

Результат: полный комплект передачи и критерии приёмки, закрепляющие способность заказчика эксплуатировать и развивать контур самостоятельно.

3.5 Матрица аргументов по ролям ЛПР заказчика

Роль	Что важно	Фокус аргумента	Аргумент из комплекта
CEO	P&L, капитализация, независимость	Go/no-go на этапах внедрения	ИИ-контур становится внутренним активом, а не внешней подпиской.
CFO	Бюджет, TCO, владение активами	Границы CapEx/OpEx и пороги окупаемости	При устойчивой высокой утилизации и горизонте владения в несколько лет собственный или гибридный контур может стать выгоднее SaaS-потребления.
CRO	Упаковка и переговоры	Этапные пакеты + бюджетные вилки	Передача и суверенный контур (резидентность данных) повышают ценность сделки для клиента.
CPO	Roadmap, качество, ROI	Приоритизация сценариев и критерии масштаба	Масштабирование опирается на измеримый эффект, а не демо-результаты.
COO	Операционная устойчивость	Готовность процесса после передачи	Роли, эскалации и эксплуатационный регламент обеспечивают устойчивую эксплуатацию.
CIO / CTO	Архитектура и управляемость	Облако РФ / on-prem / гибрид	Контроль данных, интеграций, наблюдаемости и жизненного цикла знаний.
CISO	Периметр, комплаенс, безопасность	Политики телеметрии и данных	152-ФЗ, минимизация данных, контроль обработки ПДн, OWASP/NIST-ориентиры.
Аналитик	Скорость настройки, аудит соответствия	Инструменты + естественный язык	Пакетная настройка сущностей по ТЗ, автоматический аудит приложений, реверс-инжиниринг ТЗ по стендам.
Конечный пользователь	Простота, скорость ответа	Чат на естественном языке без обучения	Поиск информации, создание записей, отчёты — без предварительного обучения.

3.6 Артефакты передачи

- Исходный код и манифест зависимостей (воспроизводимая сборка).
- Конфигурации без секретов + перечень env-переменных.
- Эксплуатационный регламент: старт/стоп/резервное копирование/масштабирование/инциденты.
- Наборы для оценки качества и регрессионных проверок с базовым уровнем и критериями деградации.
- Политика наблюдаемости: выборка, ретенция, маскирование ПДн, дашборды, оповещения.
- Контур наблюдаемости GenAI: self-hosted стек (OpenTelemetry/OpenInference + Arize Phoenix) — трассы, дашборды и оценка качества без передачи данных в облако.
- Политика данных и индексации RAG: ingestion, обновления, права доступа.
- Роли/эскалации после передачи.
- Программа обучения: бизнес, эксплуатация, разработка, ИБ.
- YAML-реестры моделей для воспроизводимой конфигурации инференса.
- OpenAPI-спецификации интеграции с Comindware Platform.

3.7 Экономическое обоснование сделки (CapEx / OpEx / TCO)

Финансовая база:

- Бюджетное обоснование стройте на тарифах облаков РФ, профилях GPU и сценарном сайзинге.
- Разделяйте затраты на разовые и повторяющиеся: инфраструктура, интеграции, сопровождение, безопасность/оценка качества, токены/API.
- В переговорах применяйте ориентиры, согласованные с единой тарифной и валютной политикой .

Выбор модели владения:

- Облачный PoC в РФ — стартовый сценарий для входа в проект.
- При устойчивой утилизации **порядка >60%** и горизонте владения **несколько лет** рассматривайте переход к гибриду или on-prem.
- Для решения CFO/CIO сопоставляйте сценарии по единым допущениям, одной валютной политике и полному TCO.

Применимость цифр:

- Вилки и бенчмарки задают порядок величин для переговоров и бюджетного коридора.
- Финальные значения для КП фиксируйте после сверки актуальных прайсов и стендовых замеров под целевой SLO.

3.8 Числовые пороги решения

- **Утилизация:** целевой ориентир >60% регулярного использования в целевой группе.
- **Эффективность:** целевой диапазон 30–40% сокращения времени выполнения типового сценария.
- **Качество:** целевой ориентир >95% по внутренней рубрике (LLM-as-judge) при стабильной методике оценки.
- **TCO break-even:** рассматривайте переход к on-prem/гибриду при устойчивой высокой утилизации **порядка >60%** и горизонте владения **несколько лет**.

Эти пороги — **внутренние операционные ориентиры** для go/no-go и масштабирования, а не универсальные рыночные нормативы. Базовую группу, окно измерения и методику оценки фиксируйте до старта пилота.

3.9 База расчётов: март 2026 г.

Если в тексте не оговорено иное, **цены, тарифы, рыночные метрики и количественные ориентиры** составляют единый справочный срез **на март 2026 года**.

Для **управленческих решений** используйте комплект как основу для сравнения сценариев, архитектур и диапазонов TCO.

Для **сметы, КП и договора** финальные значения верифицируйте по актуальным первичным источникам и условиям сделки на момент расчёта.

3.10 Валюта и правила для коммерческих предложений

1 USD = 85 руб. — единый справочный ориентир для сопоставления международных прайс-листов и рублевых оценок в материалах на март 2026 г.

- В сметах и КП применяйте курс ЦБ РФ на момент расчёта или курс, закреплённый в договоре.
- Для оценки волатильности бюджета рекомендуется закладывать чувствительность **±10%** к базовому курсу для валютозависимых статей (импортное оборудование, зарубежные облачные сервисы).

3.11 Минимальные условия передачи экспертизы (go/no-go перед BOT)

- Согласован целостный пакет отчуждения и распределение прав.
- Выполнены критерии приёмки передачи: воспроизводимая сборка, верифицированная оценка качества, эксплуатационный регламент, владельцы компонентов и срок интенсивного сопровождения после передачи (hypercare).
- Зафиксированы организационные условия после передачи и программа обучения по ролям.

3.12 Минимальные условия безопасности и наблюдаемости (pre-scale gate)

- Подтверждены требования CISO/CIO к доверию и безопасности системы перед промышленным запуском.
- Утверждены правила телеметрии и ПДн: минимизация, ретенция, доступ, периметр до LLM и состав наблюдаемости в пакете передачи.
- Проект ФЗ об ИИ используйте как ориентир дорожной карты, а не как действующую норму.
- Для сделок с EU-составляющей отдельно проверяйте роль по регламенту, требования прозрачности и штрафные риски.

3.13 Типовые возражения и ответы

Сложно обосновать ROI

Применяйте диапазоны CapEx/OpEx/TCO для раннего управленческого решения. Точный ROI фиксируйте после замеров на стенде заказчика и базового цикла пилота.

Будет vendor lock-in

Передача по BOT включает код, конфигурации, эксплуатационный регламент, контур оценки качества и обучение. Заказчик получает полную операционную автономию в своём контуре.

Риски по ИБ и 152-ФЗ слишком высокие

Архитектура строится под требования заказчика: периметр до LLM, минимизация данных, маскирование и политика журналирования. Для управляемых API — проверьте договорной контур обработки данных отдельно.

3.14 Рыночные сигналы для переговоров (контекст, не норма КП)

- По данным *«Yakov & Partners, 2025»*, GenAI используется хотя бы в одной функции у **71%** российских компаний; по *«McKinsey, 2025»* — лишь **21%** организаций фундаментально переработали хотя бы часть рабочих процессов. Рынок не испытывает дефицита интереса — дефицит в промышленном внедрении и масштабировании под требования комплаенса и экономики.
- Предложение управляемых LLM-платформ и enterprise-инструментов в РФ расширяется, снижая барьер входа. Требования к комплаенсу, TCO и модели передачи при этом не изменяются.
- Применяйте международные отчёты (OpenAI, McKinsey, Stanford) как сравнительный контекст, а не как договорную норму в КП.

3.15 Коммерческий план 30/60/90

3.15.1 0–30 дней

- Проведите рабочую сессию и выберите 2–5 приоритетных сценариев.
- Проведите аудит готовности данных и требований 152-ФЗ.
- Зафиксируйте исходный уровень: время выполнения сценария, эскалации, текущую стоимость.
- Согласуйте KPI, состав пилота и критерии перехода между этапами.

3.15.2 30–60 дней

- Разверните PoC/пилотный контур (RAG + инференс + наблюдаемость) в выбранной среде.
- Проведите техническую и бизнес-валидацию сценариев.
- Подготовьте первичную оценку ROI и управленческое решение: переход к пилоту или остановка проекта.

3.15.3 60–90 дней

- Запустите промышленный пилот на реальной нагрузке.
- Подтвердите целевые метрики качества, экономики и утилизации.
- Формализуйте комплект передачи, график отчуждения и модель сопровождения.
- Утвердите дорожную карту масштабирования в контуре заказчика.

3.16 Материалы для подготовки переговоров и КП

- В переговорах опирайтесь на шесть блоков: методология внедрения, экономика и TCO, KT/IP и приёмка, границы референс-стека, безопасность/комплаенс/наблюдаемость и единые правила KPI/FX.
- Для КП и брифа используйте и адаптируйте числа и критерии из этого документа: TCO-диапазоны, пороги перехода и критерии приёмки.

4. Методология разработки и внедрения ИИ

4.1 Обзор

Операционная модель, этапы развертывания и производственные стандарты для корпоративных ИИ-систем на базе RAG и мультиагентных архитектур в резидентном контуре РФ.

Практический смысл: формирование единого регламента внедрения, оценка организационной готовности и фиксация протоколов передачи активов (КТ) без потери операционной управляемости.

Для бюджета и ТСО: *«Сайзинг и экономика»*.

При составлении КП всегда адаптируйте данные под профиль нагрузки заказчика и проводите правовую оценку.

4.2 Концепция внедрения

- **Ситуация:** в 2026 году GenAI оценивается по P&L, а в РФ добавляются требования суверенитета данных и регуляторные инициативы по ИИ.
- **Вызов:** без явного **периметра до LLM** (минимизация и обезличивание входа, разделение вспомогательных и основной модели, политика телеметрии) растут риски по 152-ФЗ и стоимость инцидентов. Без **офлайн- и онлайн-оценки качества** невозможно доказуемо связать смену модели или индекса с качеством и бюджетом.
- **Задача:** как внедрять и масштабировать ассистентов на стеке **Comindware** (**корпоративный RAG-контур** / **MOSEC/vLLM** / **агентный слой Comindware Platform**) и передавать экспертизу и артефакты клиенту без потери управляемости?
- **Решение:**
  - Опирайтесь на следующую модель внедрения корпоративного ИИ: PoC → Пилот → Масштабирование → BOT. Предусмотрите формальные контрольные точки для контроля качества и экономики на каждом переходе.
  - Целевая операционная модель (TOM), комплект отчуждения (код, конфигурации, регламент, обучение) и блок комплаенса (152-ФЗ, Приказ № 140 Роскомнадзора, NIST AI RMF) образуют единый контур управляемого внедрения.
  - Для бюджета и TCO используйте данные раздела *«Сайзинг и экономика»*.
  - Глобальные бенчмарки задают сравнительный контекст; норму для резидентного КП определяют отдельная правовая и тарифная оценка.

4.2.1 Ключевые регуляторные вехи (2025–2027)

При планировании ИИ-проектов учитывайте следующие вехи:

Дата	Событие	Что изменилось
Июль 2025	Поправки к 152-ФЗ	Ужесточение локализации: запрет на обработку ПДн российских граждан через зарубежные базы данных
Сентябрь 2025	Отдельное согласие на ПДн	Согласие должно быть получено как отдельный документ
Сентябрь 2027	Ожидаемый закон об ИИ	Категории AI-моделей; требования к локализации для сервисов с >500К пользователей

Подробнее — см. *Приложение C «Безопасность, комплаенс, наблюдаемость»*.

4.2.2 Реестр согласованных опор для решений

Опора	Что это значит для решения	Источник	Статус
Этапность внедрения PoC → Пилот → Масштабирование	Формирует единые точки go/no-go и управляемый переход между фазами внедрения	Методология внедрения	Проверено 2026-03-31
Порог утилизации для оценки on-prem	Запускает пересмотр модели владения при устойчивой нагрузке	Логика решения (SCQA) в отчёте по экономике	Проверено 2026-03-31
Критерии передачи (КТ/ВОТ)	Подтверждает готовность заказчика к самостоятельной эксплуатации после передачи	Критерии приёмки передачи (Приложение А)	Проверено 2026-03-31

Решения по модели внедрения и передаче контура принимаются на основании этих опор. Отклонения допустимы только при явной фиксации причины и даты пересмотра.

4.3 Источник преимущества в корпоративном ИИ

Конкурентное превосходство в эпоху GenAI обеспечивается не доступом к вычислительным мощностям или моделям, а **качеством проприетарного контекста** организации.

В условиях коммодитизации LLM ключевым активом становятся структурированные датасеты, семантические графы и формализованная операционная логика компании.

Решающим фактором успеха является **AI-готовность данных**: способность систем переходить от простого извлечения информации к автономному исполнению бизнес-процессов в распределенной среде.

Для формирования устойчивого рабочего слоя ИИ необходимо внедрение следующих компонентов:

4.3.1 Семантическая связность

Фрагментированные данные не позволяют реализовать потенциал автоматизации. Необходим граф знаний, устанавливающий связи между объектами, статусами и бизнес-регламентами. Система должна четко идентифицировать сущности (клиент, договор, заказ) и понимать логику переходов между ними.

Отсутствие этого слоя ограничивает ИИ-агентов функцией поиска (Q&A), исключая возможность реального исполнения процессов.

Gartner относит тему **данных, готовых для AI**, к числу быстрорастущих в повестке по ИИ (*«Gartner — пресс-релиз: нехватка данных, готовых для AI, подрывает ИИ-проекты (26.02.2025)»*).

Инженерная проработка баз знаний и онтологий в целевой операционной модели — у роли **Knowledge Engineer** ниже.

4.3.2 Архитектура доступа

Для рабочих сценариев важно, чтобы в момент действия система получала **полный и согласованный** контекст.

Если сведения распределены по разным системам, разнятся и подтягиваются с задержкой, точность снижается уже на уровне базовых операций.

Архитектура доступа влияет на стоимость исполнения, длину сценария, количество проверок и устойчивость процесса при росте нагрузки.

Важна способность собрать **единый рабочий слой** для конкретного действия.

Связь с юнит-экономикой и TCO при многошаговых агентских цепочках — см. *«FinOps и юнит-экономика нагрузки»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

Для **пилотов линии поддержки** используйте *примерные расчёты расхода токенов* из публичного корпуса заявок.

4.3.3 Исполняемые правила

По мере роста автономности правила доступа, ограничения и маршруты согласования должны работать **автоматически**.

Для промышленного использования нужны исполняемые правила, которые применяются на уровне каждого запроса, перехода и действия: это снижает операционный риск и делает результат воспроизводимым.

Практический контур политик, защитных механизмов и комплаенса — в *Приложении C «Безопасность, комплаенс и наблюдаемость»*.

4.3.4 Внутренний контур данных

Основная прикладная ценность смещается во **внутренний контур данных** компании: бизнес-правила, историю операций, предметную логику и накопленные связи между сущностями.

Этот слой задаёт качество решения в конкретной отрасли, функции и операционной модели.

Чем точнее компания умеет формализовать и поддерживать такой контекст, тем выше качество исполнения, устойчивость сценариев и потенциал тиражирования на соседние процессы.

4.3.5 Что это значит для бизнеса

Качество внутреннего контекста определяет, способна ли система работать в реальных процессах, соблюдать логику действий и давать воспроизводимый результат.

Это подтверждает внешняя статистика: Gartner указывает, что **63%** организаций либо не имеют, либо не уверены в корректности своих практик управления данными для ИИ, и прогнозирует отказ от **60%** ИИ-проектов без **AI-ready data** (*«Gartner — lack of AI-ready data puts AI projects at risk»*).

Преимущество получают компании, создавшие **связный и управляемый слой** для работы ИИ — позволяющий встраивать систему в реальные процессы и масштабировать её за пределы локальных сценариев.

4.4 Структурированное рассуждение по схеме (SGR) в практике Comindware

Концепция: Schema-Guided Reasoning (структурированное рассуждение по схеме) — техника принудительного структурирования рассуждений LLM через предопределённые схемы. По отраслевым бенчмаркам — 5–10 % улучшение точности по сравнению с неструктурированными промптами (*«Schema-Guided Reasoning (SGR)»*).

Применение в Comindware: SGR используется в нескольких точках конвейеров — для анализа запросов, критики ответов агентов, планирования после фазы зазиты и детерминированного управления любым мышлением. Генерирует оценку спама, уверенность намерения, подзапросы для поиска, план действий.

Бизнес-смысл: предсказуемость ответов, аудит каждого шага, возможность отклонить спам до ресурсоёмкого поиска.

4.5 План разрешения инцидента

Концепция: после генерации ответа отдельный вызов LLM формирует план для инженеров — нужна ли эскалация, краткое содержание проблемы, рекомендации.

Реализация в Comindware: встроенный инструмент формирования плана анализирует ответ и контекст, возвращает структурированный план.

Бизнес-смысл: автоматическая триажа обращений, снижение нагрузки на инженеров, документирование каждого случая.

4.6 Стратегия внедрения ИИ и организационная зрелость

Предыдущий раздел фиксирует **данные и рабочий слой** как источник преимущества. Ниже — **организационная** сторона: без неё доступ к моделям и пилоты редко конвертируются в устойчивый эффект в P&L.

BCG подтверждает это правилом **10–20–70**: около **10%** результата определяют алгоритмы, **20%** — данные и технологии, **70%** — люди, процессы и организационные изменения (*«BCG — Closing the AI Impact Gap»*). Центр трансформации — способность компании **переобучать команды, пересобирать роли и закреплять новые практики** в ежедневной работе.

Разрыв между внедрением и экономическим эффектом подтверждают независимые данные: McKinsey фиксирует, что ИИ регулярно используется хотя бы в одной функции у **88%** организаций, однако эффект на enterprise-level EBIT отмечают лишь **39%**.

BCG указывает, что ИИ входит в top-3 приоритетов у **75%** руководителей, тогда как значимую ценность видят только **25%** (*«McKinsey — The state of AI»*, *«BCG — Closing the AI Impact Gap»*).

Разрыв зрелости: интерес к ИИ и инструменты доступны, однако **системные процессы внедрения, обучения и поддержки команд** не выстроены. Этот разрыв закрывают фазы **PoC** → **Пилот** → **Масштабирование**, комплект отчуждения и программа обучения — см. *Приложение А «Отчуждение ИС и кода: KT, IP, лицензии, критерии приёмки передачи»*.

Барьеры внедрения: - недоверие к результатам при отсутствии контура оценки качества и исходного уровня; - дефицит компетенций; - сопротивление команд; - отсутствие ясной модели использования; - ожидание гарантированного эффекта без измеримых KPI; - слабые публичные примеры со стороны топ-менеджмента; - страх ошибки и потери контроля.

Архитектурный и процессный ответ: человек в контуре, матрица ролей, обучение и политики; поведенческий слой детально разобран в *«Организационные и поведенческие факторы риска»* в Приложении С.

Стратегический горизонт (ориентир 3–5 лет): на таких горизонтах ИИ рационально рассматривать не как отдельный инструмент, а как часть **системы стратегической аналитики**: мониторинг технологий, отраслей, конкурентов и регулирования; внутренние бизнес-метрики; карта целевых клиентов; портфель гипотез по новым направлениям.

Управленческая рамка после привязки ИИ к **стратегическим целям** и к **измеримости влияния** на продукт и бизнес:

- интеграция ИИ в бизнес-процессы;
- распределение ответственности за внедрение;
- закрепление новых практик в **операционной модели**.

Люди и менеджмент: особенно ценны специалисты на **стыке функций**, способные быстро переводить технологию в рабочие сценарии и доводить их до измеримого эффекта; для управления командами усиливаются требования к **обучению**, **внутренней мобильности**, **поддержке экспериментов** и роли руководителей в **масштабировании** практик.

Ориентир по управленческому переобучению — программа *«Переход в ИИ: трансформация бизнес-процессов — Школа управления СКОЛКОВО»*.

Рыночные сигналы — в *«Публичные рыночные сигналы и коммерческие выводы»*.

4.7 Целевая операционная модель (Target Operating Model)

Масштабирование ИИ-решений требует перехода от централизованного AI CoE к **федеративной модели** с сильным центром компетенций.

Сводная матрица приоритетов по ролям ЛПР — *«Стратегическое резюме: матрица аргументов»*.

Методологический контекст — какие фазы и артефакты закрывают ключевые опасения каждой роли — *«Роли и ответственности»*.

4.7.1 Роли и ответственность

- **Владелец ИИ-продукта:** ответственность за бизнес-эффект, приоритизацию гипотез и road-map продукта.
- **LLMOps / Архитектор ИИ-систем:** проектирование инфраструктуры (vLLM/MOSEC), мониторинг качества (RAGAS/DeepEval — фреймворки оценки RAG), целевая архитектура **телеметрии** (трассировки, метрики токенов и латентности, политика выборки и ретенции) и согласование с ИБ при контурах с ПДн; совместно с владельцами разработки — **среда для агентов** (инструменты, линтеры, CI, контуры офлайн-оценки качества и мультиагентные циклы разработки).
- **Специалист по ИИ-безопасности:** комплаенс с 152-ФЗ и NIST AI RMF, аудит безопасности (Red Teaming).
- **Инженер знаний:** подготовка и актуализация базы знаний (Qdrant, Chroma DB, PostgreSQL+pgvector), управление онтологиями.

4.7.2 Процессы и KPI

Методология **AgentOps** задаёт три взаимозависимых слоя контроля — их порядок принципиален: нельзя улучшать то, что не измеряется, и нельзя измерять то, что не видно.

Слой 1. Наблюдаемость

Полная траектория каждого решения: вызовы инструментов, обращения к LLM, межагентные взаимодействия. Отсутствие сквозной видимости делает разбор инцидентов реактивным, а оптимизацию задержек — невозможной.

- **Сквозная длительность трассы:** время от запроса пользователя до финального ответа — ключевой показатель производительности.
- **Межагентная задержка:** время передачи задачи от одного агента к другому (цель: менее 500 мс).

- **Стоимость запроса:** совокупные затраты на API-вызовы для одного взаимодействия.

Слой 2. Оценка качества

Доказательная база результативности агента — без неё наблюдаемость остаётся мониторингом ради мониторинга.

- **Доля автономного завершения:** доля запросов, выполненных без вмешательства человека (цель: более 90%).
- **Частота нарушений защитных механизмов:** частота попыток агента нарушить политики безопасности — утечка данных, неавторизованные действия (цель: менее 1%).
- **Точность фактов:** доля корректных фактов — диагностические коды, номера полисов, дозировки. Критично для регулируемых отраслей.

Слой 3. Оптимизация

Управляемое снижение затрат и повышение качества — на данных слоёв 1 и 2, не на интуиции.

- **Эффективность токенов промпта:** экономия токенов при сохранении качества вывода (до 39% в кейсах оптимизации промптов).
- **Точность извлечения:** доля релевантных документов в top-K результатов поиска (цель: более 0,8).
- **Успешность межагентной передачи:** доля успешных передач задач между агентами (цель: более 98%).
- **Скорость улучшений:** частота внедрения оптимизаций в неделю — индикатор зрелости процесса непрерывного совершенствования.

Операционные пороги

Ориентиры задают **общий язык** для резюме, переговоров и сопоставления — применяйте их как **операционные пороги**. Они **не** заменяют юридические критерии соответствия и **не** фиксируют строки сметы без адаптации под заказчика.

- **Охват пользователей:** % сотрудников, использующих ИИ ежедневно (цель: >60%).
- **Эффективность:** сокращение времени на решение тикета/задачи (цель: 30–40%).
- **Качество по внутренней рубрике (LLM-оценщик):** оценка по **зафиксированной** рубрике и **регрессионному** набору сценариев (цель: более 95%).

Бизнес-интерпретация порога качества >95%

Порог >95% — это **внутренний операционный барьер** для релиза/масштабирования, **не** эквивалент внешней разметки человеком, **не** юридическая гарантия отсутствия ошибок и **не** «точность» в смысле научного бенчмарка без оговорки методики.

- **Юнит-экономика:** стоимость одного успешного ответа (P&L вклад).

Пример KPI доверия и эскалаций (для линии поддержки)

- **Доля ответов с проверяемой цитатой на источник** (бизнес-KPI доверия): измеряется по политике заказчика; рост доли снижает спорные обращения и эскалации.
- **Доля обращений, ушедших на эскалацию** (или снижение их числа при росте объёма): прямой сигнал для P&L линии поддержки.

Независимые бенчмарки и воспроизводимость оценки

- **Внешний эталон для русскоязычных моделей:** экосистема **MERA** (mera.a-ai.ru) на площадке **Альянса в сфере искусственного интеллекта** (a-ai.ru) даёт открытый контур сравнения фундаментальных моделей и снижает риск «оценки в вакууме» только внутренними метриками.
- **Аргумент для продаж и закупки:** участие **MTS AI** и других игроков показывает, что отрасль движется к стандартизации оценки качества; это усиливает обоснование выбора моделей для заказчика в цикле PoC → Пилот → Масштабирование.
- **Как использовать в проекте:** внешние ориентиры (MERA) должны дополнять, а не заменять внутренний контур оценки заказчика (RAGAS, DeepEval, LLM-evaluator) при фиксации KPI качества в проектной документации.
- **Требование к отчуждению и воспроизводимости:** разбор цикла улучшения **Cotype** с опорой на LLM-судей (*«Хабр, MTS AI»*) используем как методологический референс; в пакет передачи необходимо включать промпты судей, эталоны и регрессионные наборы.

4.7.3 Публичные рыночные сигналы и коммерческие выводы

Материалы канала лаборатории **red_mad_robot** **@Redmadnews** служат внешним индикатором рыночного спроса для формирования коммерческой позиции на уровне C-Suite и приоритизации сценариев внедрения и отчуждения ИИ.

> ⚠ **Границы применения**
>
> Публикации в каналах и подкастах используйте как рыночный сигнал для приоритизации гипотез и форматов сделки.
>
> Для утверждения бюджета, SLA, состава поставки и ответственности сторон опирайтесь только на первичные документы: договор, приложения, тарифы и юридические заключения на дату согласования.

- **Сигнал 1 — «фабрика ИИ-агентов» и СП (Билайн/ВымпелКом):** бизнес подтверждает спрос на масштабируемые, промышленно управляемые ИИ-контуры.

  Вывод для Comindware: усиливать позиционирование продажи через модель **PoC → Пилот → Масштабирование** и заранее предлагать формат **совместной разработки/поэтапной передачи**.

  Источник: «*СП и фабрика агентов (пост канала)*».

- **Сигнал 2 — ожидания ведущей роли R&D и измеримости эффекта:** рынок требует не «витринного ИИ», а управляемого центра компетенций с доказуемым бизнес-эффектом.

  Вывод для Comindware: в коммерческих диалогах фиксировать KPI перехода **PoC → Пилот → Масштабирование** и связывать их с SLA/экономикой проекта.

  Контекст публикации: **Т-Банк**, **Авито**, **MWS AI**, **ВкусВилл**, **red_mad_robot**.

- **Сигнал 3 — постоянный поток инженерных новинок и R&D-дайджестов:** технологический горизонт меняется быстрее типового цикла закупки.

  Вывод для Comindware: использовать такие материалы как вход в методический радар, но не как замену продуктовым требованиям, архитектурным ограничениям и составу поставки по КП.

  Источник: «*R&D в AI в 2026 (пост канала)*».

- **Сигнал 4 — AI-first организационная модель:** зрелые игроки перестраивают продуктовые и корпоративные процессы вокруг ИИ.

  Вывод для Comindware: усиливать направление передачи экспертизы (KT, role-based enablement, операционные регламенты) как часть коммерческого предложения, а не как опциональную активность.

  Источники: «*AI-first стратегия: подкаст (пост канала)*», «*Подкаст «Ноосфера» #129 (YouTube)*».

4.8 Методология внедрения (этапы и качество)

Внедрение агентных ИИ-систем требует смены управленческой парадигмы. Детерминированная логика («одинаковые входы — одинаковый выход») уступает вероятностным системам. Среда перестаёт быть статичной и адаптируется на основе обратной связи. Приоритет смещается от реализации кода к измерению результатов — принцип **«оценка (eval) первична»**: успех определяется не тем, что написано, а тем, что измеримо достигнуто.

DevSecOps встраивает безопасность в каждый этап — проектирование, разработку, тестирование, эксплуатацию и мониторинг — не как контроль после запуска, а как неотъемлемое свойство контура.

Реализуйте 4-фазный подход, основанный на практиках **red_mad_robot** и **Just AI**, с интегрированной безопасностью на каждом этапе:

4.8.1 Фаза 1. PoC (2–4 недели)

- **Цель:** проверка технической осуществимости.
- **Парадигма:** зафиксируйте границы **допустимой агентности** — что агент вправе делать самостоятельно, а что требует подтверждения человеком; установите базовые eval-метрики, по которым PoC считается пройденным.
- **Инструментарий:** базовый корпоративный RAG-контур, базовый инференс; при платформенных сценариях — **агентный слой Comindware Platform**.
- **Наблюдаемость:** базовые трассировки и учёт токенов (ориентир OpenTelemetry GenAI; при необходимости **Phoenix/OpenInference** в песочнице) — см. *«Промышленная наблюдаемость LLM, RAG и агентов»*.
- **Артефакты:** прототип, базовые метрики (латентность, качество, стоимость), перечень ограничений для перехода в пилот.
- **Контроль:** успешное выполнение 10 критических сценариев.

4.8.2 Фаза 2. Пилот (1–3 месяца)

- **Цель:** валидация в промышленном окружении на ограниченной группе пользователей.
- **Инструментарий:** оптимизированный инференс, внедрение защитных механизмов, согласование нагрузки со стороны корпоративного RAG-контура и **агентного слоя Comindware Platform**.
- **Наблюдаемость:** продакшн-телеметрия с политикой выборки и ретенции; связка трасс с офлайн-метриками качества — см. *«Связь с контуром оценки качества»*.
- **Артефакты:** пилотный контур в промышленной среде, отчёт по ROI и качеству, backlog доработок перед масштабированием.

- **Контроль:** замер ROI, сбор обратной связи (Human-in-the-loop).

4.8.3 Фаза 3. Масштабирование (3–12 месяцев)

- **Цель:** Enterprise-wide внедрение.
- **Инструментарий:** масштабирование инференса, развитие корпоративного RAG-контура под нагрузкой; для операций с сущностями платформы — **агентный слой Comindware Platform**; при необходимости рой агентов (координатор/воркер).
- **Наблюдаемость:** единый контур FinOps (токены, задержки, ошибки) и регрессии после смены модели или индекса по связке трасс с **офлайн- и онлайн-оценкой качества** — см. *«Рынок РФ, наблюдаемость LLM и референс-стек Comindware»*.
- **Артефакты:** тиражируемый продакшн-контур, регламенты эксплуатации и мониторинга, утверждённая модель масштабирования по доменам.
- **Контроль:** стабильность под нагрузкой (SLA 99,9%), соответствие бюджету (FinOps).

4.8.4 Фаза 4. Оптимизация (Постоянно)

- **Цель:** снижение TCO и повышение качества.
- **Инструментарий:** DSPy для оптимизации промптов, квантование моделей, кэширование (LMCache); при зрелости команды — регламент мультиагентных циклов (план → реализация → независимая проверка) и меры против энтропии документации относительно кода (периодическая синхронизация, «сборка мусора» артефактов) в духе отраслевой инженерии обвязки (OpenAI — Harness engineering, Хабр).
- **Скорость улучшений:** метрика частоты внедрения оптимизаций — типично 3 оптимизации в неделю (промпты, извлечение, корректировка потока). Показывает зрелость процесса непрерывного совершенствования.
- **Артефакты:** контур непрерывного улучшения качества/стоимости, обновляемый реестр оптимизаций и проверенный цикл сопровождения.
- **Контроль:** устойчивое снижение TCO, отсутствие регрессий качества на контрольных наборах, соблюдение целевых SLA/FinOps.

4.9 Детальная архитектура внедрения

Область применимости архитектуры

Приведённый ниже **универсальный** архитектурный паттерн применяется к корпоративным сценариям (поддержка, сервис-деск, внутренние ассистенты, обработка регламентов) с использованием корпоративных данных.

Примеры метрик и процессов основаны на сценарии линии поддержки и иллюстративны — они не ограничивают область применения.

4.9.1 Основные компоненты

Компонент	Проект	Роль	Технология
RAG-движок	**Корпоративный RAG-контур**	Оркестрация поиска, генерации и логики агентов, получающих корпоративные данные	Python, LangChain, Gradio
Сервер инференса (универсальный)	**Сервер инференса MOSEC**	Развёртывание специализированных моделей: эмбеддера, ранжировщика и защитника	MOSEC, PyTorch
Сервер инференса (высокопроизводительный)	**Инференс на базе vLLM**	Развёртывание LLM	vLLM, CUDA
Векторное хранилище	**Корпоративный RAG-контур**	Постоянное хранение эмбеддингов документов	Qdrant, Chroma DB, PostgreSQL+pgvector (HTTP)

4.9.2 Поток данных и конвейер

1. **Загрузка данных:**
   - Документы (Markdown, MkDocs) обрабатываются модулем обработки документов RAG-движка.
   - Разбиваются на чанки через токен-зависимый чанкер.
   - Векторизуются через компонент эмбеддера (FRIDA/Qwen3).

- Векторы и метаданные сохраняются в векторной БД (Qdrant, Chroma DB, PostgreSQL+pgvector).

2. **Поиск (RAG):**
   - Пользовательский запрос поступает в поисковый конвейер.
   - **Векторный поиск:** векторная БД извлекает top-k чанков.
   - **Реранкинг:** кросс-энкодер или LLM-ранжировщик уточняет результаты.
   - **Сборка контекста:** статьи восстанавливаются, при необходимости суммируются (модуль суммаризации).
3. **Генерация:**
   - **Режим агента (Рекомендуется):** агент LangChain анализирует запрос, принудительно вызывает инструмент извлечения контекста и генерирует ответ с цитатами.
   - **Прямой режим:** менеджер LLM генерирует ответ напрямую из найденного контекста.
4. **Доставка:**
   - **Веб-интерфейс:** Gradio ChatInterface для работы с RAG-движком в чате.
   - **API:** REST-эндпоинт `/api/query_rag`.
   - **Виджет:** встраиваемый HTML/JS виджет для внедрения на любые сайты.

4.9.3 Конфигурация сервера инференса

4.9.4 MOSEC, vLLM и наработки **Comindware**

Базовые технологии (апстрим) — открытые проекты для развёртывания больших языковых и специализированных моделей:

- **MOSEC** — фреймворк с Rust-вебслоем и Python-воркерами: динамическая пакетная обработка, поэтапные пайплайны, облачные практики (прогрев, graceful shutdown, метрики). **Меньший расход памяти** — комбинация эмбеддинг+ранжировщик+защитник укладывается в ~5 ГБ (против ~10–15 ГБ у vLLM), что критично для ограниченных GPU. Подробные замеры — в *«Перерасход памяти vLLM»*.
  - (Репозиторий)
  - Документация.
- **vLLM** — движок инференса с OpenAI-совместимым API: PagedAttention, непрерывная пакетная обработка, выгрузка KV-кэша. Ориентир — **максимальная производительность LLM** ценой повышенного расхода памяти (KV-кэш, батчинг в памяти).

- (Репозиторий)
- Документация.

Наработки Comindware — прикладные комплекты обвязки вокруг MOSEC и vLLM:

- **сервер инференса MOSEC** — управление процессом, YAML-реестр с проверенными конфигурациями, воркеры эмбеддера/ранжировщика/защитника на **одном HTTP-порту** с OpenAI-совместимыми маршрутами. **Одна сетевая точка** — проще политики безопасности и сопровождение. CLI: установка, запуск, статус, остановка, тестирование.
- **инференс на базе vLLM** — жизненный цикл процессов (загрузка, проверки здоровья), pooling-режимы для эмбеддеров/скоринга, YAML-реестр, гибкий выбор чекпоинтов под нагрузку. Выгрузка KV-кэша через LMCache (vLLM v1).

4.9.5 Одна HTTP-точка и несколько серверных процессов

В **сервер инференса MOSEC** на **одном HTTP-порту** сосуществуют **разные роли** (эмбеддинг, ранжирование, модерация) в рамках **одного MOSEC-сервиса** с разными воркерами — это **не** размещение нескольких независимых процессов vLLM за одним портом. У **vLLM** распространённый паттерн — **отдельный серверный процесс на модель/конфигурацию**; несколько моделей обычно означает **несколько инстансов** (часто на разных портах) и маршрутизацию на стороне клиента, API-шлюза или балансировщика. Исключения и тонкости multi-GPU/репликации одной модели — по документации vLLM для выбранной версии.

4.9.6 Вариант A: унифицированный сервер (сервер инференса MOSEC)

- **Эксплуатация:** запуск объединённого сервиса через CLI комплекта **сервер инференса MOSEC** (порт и активные модели задаются конфигурацией; типичный порт по умолчанию — 8001, см. поставляемую документацию).
- **Модели:** эмбеддер, ранжировщик и защитник могут подключаться динамически в рамках поддержанного набора.
- **Выгоды для внедрения:** меньше сетевых конечных точек, проще обучение эксплуатации и отчуждение эксплуатационного регламента клиенту; хороший старт для пилотов **корпоративный RAG-контур**.
- **Сайзинг:** VRAM делится между фактически загруженными моделями на узле; детальные оценки памяти публикуются вместе с комплектом **сервер инференса MOSEC** (артефакты замеров и методика — в документации репозитория).
- **Ограничения:** расширение модельного ряда упирается в то, что команда интегрировала в MOSEC-воркеры (меньше «произвольного зоопарка», чем у голого vLLM).

4.9.7 Вариант Б: распределённые инстансы vLLM (инференс на базе vLLM)

- **Эксплуатация:** отдельный процесс vLLM на выбранную модель и порт через CLI **инференс на базе vLLM** (точные флаги и примеры — в поставляемой документации **инференса на базе vLLM**).
- **Типичная схема сети:** отдельные порты для LLM, эмбеддера, ранжировщика, защитника, если все роли вынесены на vLLM (например, 8100, 8101, 8105 — иллюстративно; фактические значения задаются политикой развёртывания).
- **Выгоды для внедрения:** зрелые GPU-оптимизации vLLM (в т.ч. KV-кэш, непрерывная пакетная обработка), удобное горизонтальное масштабирование реплик под SLA по задержке и пропускной способности.
- **Сайзинг:** выше суммарный перерасход памяти VRAM и число процессов; зато предсказуемее поведение под пиковые нагрузки и длинный контекст при правильном шардировании и профиле **корпоративный RAG-контур** / **агентный слой Comindware Platform**.
- **Ограничения:** сложнее операционная картина (несколько сервисов); смена модели чаще требует перезапуска процесса по сравнению с динамической загрузкой в **сервер инференса MOSEC**.

Команды CLI, примеры портов и переменные окружения — в поставляемой документации **сервера инференса MOSEC** и **инференса на базе vLLM**; в этом документе — архитектурный выбор, экономика и риски.

4.9.8 Ассистент аналитика как проверенный агентный паттерн

Ассистент аналитика Comindware — проверенный агент для прямого взаимодействия с Comindware Platform на естественном языке.

- Инструменты для BPM-платформы
- Мультипровайдер LLM
- Сессионная изоляция: каждый пользователь получает отдельный экземпляр агента и LLM
- Наблюдаемость: LangSmith (трассировка) + Langfuse (наблюдение) + Arize Phoenix (мониторинг) + учёт токенов и стоимости
- Варианты развёртывания: замкнутый контур, VPN, облачная LLM, MCP-server mode
- Обработчик ошибок с классификацией TF-IDF (частотно-обратная индексная частота): адаптация на ходу

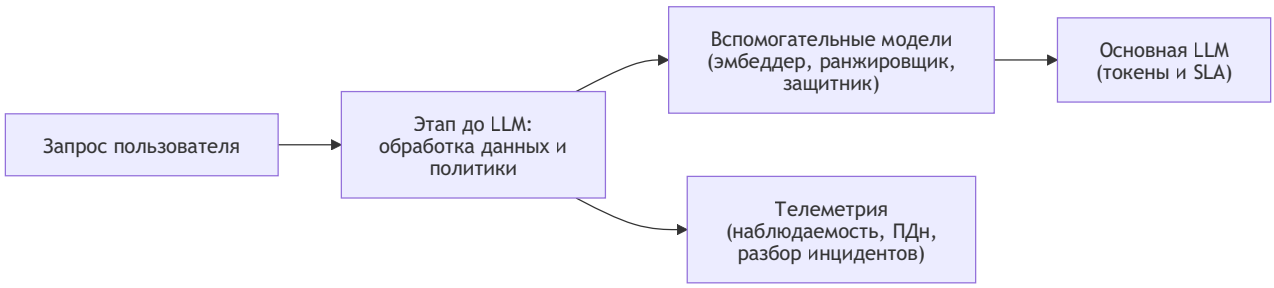
4.9.9 Три измерения гибридного размещения и выбор бэкенда по типу модели

Три **независимых решения** при проектировании гибрида: где живут данные, где крутятся вспомогательные модели и где вызывается основная LLM. Это **не** последовательные шаги одного запроса, а **разрезы для сметы и договора**.

1. **Данные и ПДн:** где хранятся и обрабатываются исходные сообщения, индекс RAG, журналы; соответствует требованиям локализации и согласий.
2. **Вспомогательные модели:** эмбеддер, ранжировщик, защитник, при необходимости слой маскирования/NER до LLM; часто совмещаются на одном унифицированном HTTP-сервисе (**MOSEC**) или распределяются по отдельным процессам (**vLLM** и др.) в зависимости от нагрузки и поддерживаемых форматов.
3. **Основная LLM:** управляемый API в РФ или self-hosted; здесь концентрируется основной счётчик токенов и требования к задержке.

Для **вспомогательных моделей** (**второе измерение**) инженерные замеры на референс-стеке показали, что **разные классы моделей** не всегда совместимы с конкретным серверным движком без потери корректности (например, корректный pooling для эмбеддеров и ограничения для генеративного ранжировщика). Это влияет на **число процессов, фрагментацию GPU и регрессионное тестирование** при обновлениях — количественные ориентиры и строки TCO — см. *«Слой перед LLM и режимы нагрузки (ориентиры для модели затрат)»* в «Сайзинг и экономика (CapEx / OpEx / TCO)».

Цепочка для переговоров и сметы: в **потоке запроса** запрос сначала проходит **этап с низким расходом токенов** — предобработку и **вспомогательные модели** (эмбеддер, ранжировщик, защитник), затем попадает в **основную LLM**, где концентрируется **основной счётчик токенов**; **наблюдаемость и следы для ПДн** закладывают **до** дорогого вызова, чтобы инциденты и аудит не догоняли эксплуатацию.



4.9.10 Российские облачные провайдеры ИИ

Раздел для руководителей, формирующих предложение **Comindware** по внедрению и отчуждению ИИ-экспертизы в контуре заказчика.

Приоритеты: РФ-резидентность данных, управляемый комплаенс-риск и предсказуемая экономика владения.

- **Управленческая цель:** быстро выбрать модель поставки (локальный managed API, гибрид, закрытый контур) под требования заказчика к данным, SLA и передаче компетенций.
- **Коммерческая цель:** использовать единый набор проверяемых аргументов в пресейле и в переговорах о передаче (KT/IP), согласовывая тарифные цифры с *«Тарифы российских облачных провайдеров ИИ»* в «Сайзинг и экономика (CapEx / OpEx / TCO)» (без расхождений между материалами комплекта).
- **Принцип чтения блока:** здесь фиксируются роли провайдеров, составы модельных линеек и правила сверки SKU; расчётные значения и сценарный сайзинг — в *«Тарифы российских облачных провайдеров ИИ»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

Тарифы и сценарный сайзинг для переговоров

Для расчётов и ценовых сравнений опирайтесь на *«Тарифы российских облачных провайдеров ИИ»* в «Сайзинг и экономика (CapEx / OpEx / TCO)»: количественные тарифы (руб. за токены, комплекты, руб./час GPU), дерево факторов стоимости и сценарный сайзинг.

Дополнительные ориентиры по аренде GPU (IaaS РФ) для поставщиков вне основной сводной таблицы — в *«Цены на GPU-оборудование (покупка и аренда)»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

Cloud.ru (Evolution Foundation Models) · продукт · тарифы

- **API:** OpenAI-совместимый доступ к моделям в российских ЦОД.
- **Каталог (на странице продукта перечислены позиции с идентификаторами Hugging Face `org/repo`):**
  - **GigaChat:** коммерческие SKU `GigaChat`, `GigaChat Lite`, `GigaChat Pro`, `GigaChat-2-Max` и ветка `ai-sage/GigaChat3-10B-A1.8B` (для SKU-to-Hub мэппинга проверять соответствие по официальному каталогу и карточке модели).
  - **GLM (Zhipu, org `zai-org`):** `GLM-4.6`, `GLM-4.7`, `GLM-4.7-Flash` (пример карточки); крупное семейство `GLM-5` — на HF.
  - **Qwen (Alibaba, org `Qwen`):** `Qwen3-235B-A22B-Instruct-2507`, семейства `Qwen3-Coder-*`, `Qwen3-Next-80B-A3B-Instruct`; линейка `Qwen3.5-*` (в т.ч. MoE) — сверять наличие в каталоге и в прайсе на дату.
  - **T-Tech:** линейки `t-tech/T-lite-it-*`, `T-pro-it-*`.

  - **Прочие текстовые LLM:** `openai/gpt-oss-120b`, `MiniMaxAI/MiniMax-M2`.
  - **Эмбеддеры и ранжировщики:** `BAAI/bge-m3`, `BAAI/bge-reranker-v2-m3`, `Qwen/Qwen3-Embedding-0.6B`, `Qwen/Qwen3-Reranker-0.6B`.
  - **Речь и документы:** `openai/whisper-large-v3`, `deepseek-ai/DeepSeek-OCR-2`.
- **Тарификация:** оплата **по токенам** (входные и генерируемые — отдельно, см. официальный прайс). **Все ₽/млн и расшифровка по строкам** (в т.ч. GigaChat3-10B-A1.8B, Qwen3-235B, GigaChat-2-Max, GLM-4.6, MiniMax-M2) — в *«Тарифы российских облачных провайдеров ИИ»* в «Сайзинг и экономика (CapEx / OpEx / TCO)»; маркетинговый перечень на сайте может быть **шире** прайса.
- **SKU vs Hub:** имя в биллинге **не** гарантирует ту же ревизию весов, что на Hugging Face, без явной проверки.

Yandex Cloud (Yandex AI Studio / YandexGPT) · модели · тарификация

- **Модели (текст, базовый инстанс):** в обзорах и переговорах часто выделяют **YandexGPT Pro 5.1** и **Alice AI LLM**; полный перечень — доступные генеративные модели: Alice AI LLM; YandexGPT Pro 5.1 и Pro 5; YandexGPT Lite 5; DeepSeek V3.2; Qwen3 235B; gpt-oss-120b и gpt-oss-20b; Gemma 3 27B (условия Gemma); дообученная YandexGPT Lite; YandexART и Realtime.
- **Тарифы:** первоисточник — правила тарификации AI Studio: таблица Model Gallery, ₽ **с НДС** за **1000** токенов (входящие, кеш, инструменты, исходящие); для агентов — отдельно токены инструментов. Эквиваленты **₽/млн** и строки по моделям — в *«Тарифы российских облачных провайдеров ИИ»*, «Сайзинг и экономика (CapEx / OpEx / TCO)». Публикации СМИ (например, ориентиры порядка ~0,5 ₽ за 1000 токенов) — для справки; для КП использовать только официальный прайс.
- **Особенности: OpenAI-совместимый** доступ к ряду моделей; **интеграция с экосистемой Yandex Cloud** (данные, идентичность, смежные сервисы — по политике заказчика и документации Яндекса); линейка **YandexGPT / Alice** ориентирована в том числе на **русскоязычные** сценарии наряду с мультиязычными моделями в галерее.

SberCloud (GigaChat API) · портал · юридические тарифы

- **Модели:** GigaChat-2 Lite, Pro, Max.
- **Тарифы:** комплекты токенов по юридическим тарифам; эквиваленты **₽/млн** и размеры комплектов — в *«Тарифы российских облачных провайдеров ИИ»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

Selectel (Foundation Models Catalog) источник

- Каталог с выделенным endpoint, API **совместим с OpenAI**; оплата за **CPU, GPU, RAM, диски**, не за токены. **Private Preview**, список моделей в панели (ссылки на HF). Свои веса **не** заявлены (FAQ на сайте).

MWS GPT (MTC Web Services) · продукт · тарифы

- OpenAI-совместимый API, SLA **99,95%** (для части моделей), режимы **SaaS / hybrid / on-prem**. Прайс **без НДС** за 1000 токенов под внутренними именами; сопоставление с публичными названиями — у поставщика. **Цифры** (лендинг, таблица «Модель N», НДС) — в *«Тарифы российских облачных провайдеров ИИ»* в «Сайзинг и экономика (CapEx / OpEx / TCO)» (блок **MWS GPT**).

VK Cloud (ML) документация

- **Cloud ML Platform**, Spark, Cloud Voice, Vision — **без** публичного каталога готовых LLM в формате Evolution FM / AI Studio; типичный путь — **своя** модель и MLOps.

4.9.11 Матрица: управляемый API в РФ и открытые веса

Контур	API в РФ	Self-host / HF	Примеры семейств
Cloud.ru Evolution FM	Да	Часто те же `org/repo`, что в каталоге FM	GigaChat, GLM-4.6–4.7-Flash, Qwen3-235B / Coder / Next, gpt-oss, MiniMax-M2, T-tech
Yandex AI Studio	Да	Отдельные модели на HF (в т.ч. кастомные лицензии)	YandexGPT, Alice, DeepSeek V3.2, Qwen3 235B, gpt-oss, Gemma 3
Sber GigaChat API	Да	**GigaChat 3.1** MIT на HF (ai-sage)	Коммерческий API и открытые веса — разный TCO
Selectel FMC	Да (Private Preview)	Каталог → HF; свои веса не заявлены	Оплата **инфраструктура**, не токены
MWS GPT	Да	Публичный каталог HF не сведён	Прайс по кодам «Модель N»
VK Cloud ML	Нет LLM-каталога в документации	BYO на ML Platform	Инфраструктура под **инференс на базе vLLM/ MOSEC**

Типично только открытые веса (доставка в РФ — GPU-облако или on-prem): в следующей таблице — **родственные чекпойнты** на Hugging Face по группам; многие те же `org/repo`, что в каталоге **Cloud.ru Evolution FM**.

Группа	Репозитории на Hugging Face (родственные модели)	Заметка для заказчика
GLM (Zhipu, `zai-org`)	GLM-4.6 · GLM-4.7 · **GLM-4.7-Flash** (более компактная ветка) · GLM-5 (флагман MoE)	Линейка **4.6–4.7** и **GLM-5** — разный масштаб VRAM; **4.7-Flash** — типичный кандидат, когда нужен меньший след по железу при том же бренде
gpt-oss (OpenAI)	openai/gpt-oss-20b · openai/gpt-oss-120b; варианты с фильтрацией: gpt-oss-safeguard-20b · gpt-oss-safeguard-120b	**Apache-2.0**; те же публичные имена, что у **Yandex AI Studio** и **Cloud.ru** FM, но хостинг и комплаенс — на стороне заказчика
Qwen3 / Qwen3.5 (`Qwen`)	org Qwen: MoE Qwen3-235B-A22B-Instruct-2507, Qwen3-Next-80B-A3B-Instruct; код: Qwen3-Coder-30B-A3B-Instruct, Qwen3-Coder-480B-A35B-Instruct; **Qwen3.5:** например Qwen3.5-35B-A3B и др. на Hub	Семейство шире перечисления; сверять **лицензию**, **gated** и поддержку **vLLM/SGLang** по карточке
GigaChat (открытые веса Сбера, `ai-sage`)	GigaChat3-10B-A1.8B (3.0) · GigaChat3.1-10B-A1.8B; крупный чекпойнт: GigaChat3.1-702B-A36B	**MIT** на публичных весах; **GigaChat API** (SberCloud) и self-host — разный TCO (см. абзац ниже)
MiniMax M2	MiniMaxAI/MiniMax-M2	На HF — **modified MIT** / особая лицензия в карточке; дублируется как SKU **Cloud.ru** FM — сверять прайс и условия
DeepSeek R1 distill	DeepSeek-R1-Distill-Qwen-32B · DeepSeek-R1-Distill-Llama-70B и др. на `deepseek-ai`	Плотные модели разного размера под локальный инференс; рядом на Hub — полные ветки **DeepSeek-V3 / R1** (другой сайзинг)
NVIDIA Nemotron 3	NVIDIA-Nemotron-3-Nano-30B-A3B-FP8 и др. в org nvidia	MoE, заявленный контекст до **1M** токенов (обзор); **не** готовый **API РФ** без своего контура
Kimi (Moonshot)	moonshotai/Kimi-K2-Base; линейка K2.5 — в org moonshotai	Часто в IDE и агрегаторах; для КП требуется явный контур и лицензия

Все **числовые** ориентиры по управляемым API — в *«Тарифы российских облачных провайдеров ИИ»*, «Сайзинг и экономика (CapEx / OpEx / TCO)». Отдельно

Сбер публикует **открытые веса** GigaChat-3.1-Ultra и Lightning под **MIT** (Хабр): экономика смещается в **CapEx/OpEx GPU** — см. *«Открытые веса и API: влияние на TCO»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

Паттерн «чекпойнт на Hugging Face + отдельная лицензия» (не эквивалент permissive open source вроде MIT) меняет комплект отчуждения и учёт: у публичной ветки **YandexGPT-5-Lite-8B** применяется **кастомное лицензионное соглашение**, где при коммерческом использовании при достижении **10 миллионов выходных токенов в месяц** лицензиат в течение **30 календарных дней** после такого месяца обязан связаться с правообладателем для согласования дальнейшего использования, иначе лицензии прекращаются (полный текст). В том же тексте зафиксированы **применимое право РФ** и требования к **указанию авторства** при распространении — это входит в юридический контур передачи и в **мониторинг объёма генерации**, параллельно со сдвигом TCO в сторону **GPU и эксплуатации**, как у любого self-hosted чекпойнта.

Исследовательские публикации лабораторий перечисляют направления вроде **эффективных LLM** и оптимизации (пример — дайджест за 2025 год); как **инженерный ориентир** для PoC по памяти при длинном контексте полезен класс работ по **сжатию KV-кэша** (arXiv:2501.19392, среди принятых к ICML 2025).

4.10 Рекомендации по производственной эксплуатации (2026)

На основе исследования «Продвинутые подходы к RAG»:

1. **Гибридный поиск:** реализуйте BM25 + плотный поиск для достижения качества уровня enterprise (прирост 4–7,5%).
2. **Адаптивная маршрутизация:** анализируйте сложность запроса и маршрутизируйте простые запросы напрямую в LLM, избегая лишнего поиска.
3. **Самокоррекция:** реализуйте механизмы критики для сложных запросов, снижающие галлюцинации.
4. **Мониторинг и наблюдаемость:** отслеживайте точность поиска, релевантность контекста и частоту галлюцинаций. Закрепите **трассировки по этапам RAG и агента** и **метрики токенов и задержек** согласно *«Промышленная наблюдаемость LLM, RAG и агентов»* в Приложении C и *«OpenTelemetry GenAI»*, учитывая статус **Development** GenAI semconv и требования к политике данных по ПДн.
5. **Длинные ответы и зацикливание:** измеряйте устойчивость генерации (повторы, «хвостовые» циклы). Сбер публично описывает борьбу с зацикливанием в постобучении MoE-моделей GigaChat 3.1 (Хабр). Воспроизводите эти показатели на **своих** сценариях оценки качества — не принимайте как гарантию без замеров.

4.11 Общие рекомендации

1. **Для новых внедрений:**
   - Начинайте с **сервера инференса MOSEC/vLLM** для упрощения операций (единая точка входа).
   - Применяйте режим агента в **корпоративном RAG-контуре** для динамического вызова инструментов.
   - Для сценариев управления через **Comindware Platform** подключайте **агентный слой Comindware Platform** и совместно планируйте нагрузку на LLM/API с **корпоративным RAG-контуром**.
   - Внедрите гибридный поиск (BM25 + вектор) для оптимального качества.
2. **Для масштабирования:**
   - Переходите на **инференс на базе vLLM** для LLM-инференса (оптимальная производительность).
   - Масштабируйте **корпоративный RAG-контур** и **агентный слой Comindware Platform** независимо по профилям трафика (RAG vs операции платформы).
   - Используйте отдельные инстансы vLLM для каждой роли (эмбеддер, ранжировщик, защитник, анонимизатор, основная LLM) и вспомогательных сервисов для балансировки нагрузки.

- Применяйте Kubernetes для оркестрации при масштабировании на несколько узлов.

3. **Для отчуждения:**

- Архивируйте исходные документы перед удалением векторных данных.
- Перед остановкой контура выполните диагностику векторного хранилища стандартными утилитами сопровождения.

4.12 Практики и архитектуры RAG: NeuralDeep и продвинутая поисковая инженерия

Корпоративный RAG-контур при отчуждении должен оставаться полностью воспроизводимым: получение данных, чанкинг, формирование эмбеддингов, вызовы LLM, ранжирование, выбор фреймворка, agentic-петли, контур оценки качества и защитные механизмы. Консолидированные разработки по NeuralDeep и паттернам @ai_archnadzor приведены в *«Практики и архитектуры RAG»*.

4.12.1 Извлекаемые уроки из публичных материалов OZON Tech (РФ)

Излагаемые ниже идеи — **не продвижение**, а переносимые управленческие и инженерные практики из материалов **Ozon Tech** (Хабр, анонсы митапов). **Классический ML в поиске и рекламе, равно как сценарные чат-боты с навыками, не эквивалентны GenAI/RAG**.

Применяйте материалы как **аналогии** для поискового слоя, платформенного внедрения и MLOps — но не как замену стандартам (NIST AI RMF), практикам FinOps и комплекту отчуждения, принятому в организации.

- **Платформа вместо разовых ботов:** переход от узкой команды сценаристов к **no-code**-конструктору, **масштабирование на организацию** и цель **запуска нового бота за сутки** (против «не менее недели» в прежней модели), плюс поэтапный **MVP на одном боте** с последующим переносом остальных — близко к идее **федеративной ТОМ** и **платформенного** внедрения множества ассистентов, а не только одного RAG-контура (Хабр, Ozon Tech).
- **Поисковый слой: не везде «только вектор»:** в задаче подсказок/текстового поиска обсуждаются компромиссы **ANN (эмбеддеры) vs обратный индекс**, фильтрация по бизнес-правилам в рантайме, **латентность и ресурсы**, интерпретируемость выдачи — по смыслу сонаправлено с **гибридным поиском** в RAG и с **FinOps-учётом стоимости и задержки** этапа извлечения (Хабр, Ozon Tech).
- **MLOps-ритм:** в программе публичного митапа описана **ML-инфраструктура**, позволяющая **регулярно тестировать новую функциональность, обучать**

модели и автоматически выкатывать их — перекликается с требованиями к **LLMOps, регрессиям и выкатке** в этом документе (Хабр, Ozon Tech).

- **Отчуждение vs открытая инженерия:** публикации статей и **открытые репозитории** на GitHub — пример **обмена практиками** с рынком; это **не эквивалент** полноценному комплекту передачи (код, данные, модели, эксплуатационный регламент, IP, обучение) из раздела *«Что передаётся клиенту при отчуждении знаний»* в этом документе (организация ozontech на GitHub).

4.12.2 NeuralDeep: данные, модельный ряд, agentic RAG и безопасность

4.12.3 ETL и подготовка данных

- **markitdown** — конвертация документов в Markdown (GitHub)
- **marker** — быстрое извлечение текста из PDF (GitHub)
- **docling** — продвинутое извлечение данных из документов (GitHub)

4.12.4 Чанкование (Chunking)

- **Chonkie** — быстрая и легковесная библиотека для чанкования (GitHub)
- LangChain text splitters (GitHub)

4.12.5 Векторные модели для русского языка

- **ai-forever/FRIDA** — российская модель, оптимизированная для русского
- **BAAI/bge-m3** — мультиязычная модель
- **intfloat/multilingual-e5-large** — мультиязычные эмбеддеры
- **Qwen3-Embedding-8B** — большая мультиязычная модель

4.12.6 Суверенный стек одного вендора (опционально)

Помимо LLM из коллекции GigaChat 3.1 на Hugging Face у организации ai-sage опубликованы коллекции GigaEmbeddings, GigaAM (модели для речи) и GigaChat Lite. Их можно рассматривать при цели **единого открытого контура** под одним вендором весов; это **не** обязательная замена рекомендованных для **корпоративный RAG-контур** эмбеддеров (FRIDA, Qwen3 и т.д.): решение фиксируется в **ADR**, с оценкой качества RAG и проверкой **лицензии** на каждой карточке модели.

4.12.7 LLM и vLLM модели для русского сегмента

Рекомендации сообщества по соотношению цена/качество:

- **t-tech/T-lite-it-1.0** — легкая модель для русского языка
- **t-tech/T-pro-it-2.0** — продвинутая модель для русского языка
- **Qwen3-30B-A3B-Instruct-2507** — рекомендуется для Agentic RAG (GitHub)
- **RefalMachine/RuadaptQwen2.5-14B-Instruct** — адаптированная для русского

4.12.8 ранжировщики

- **BAAI/bge-reranker-v2-m3** — мультиязычный кросс-энкодер
- **Qwen3-Reranker-8B** — большая модель для ранжирования

4.12.9 Фреймворки для RAG

Одобрено сообществом NeuralDeep: - **Dify** — Low-code платформа для AI-приложений (GitHub) - **AutoRAG** — автоматический RAG оптимизатор (GitHub) - **LlamaIndex** — структурированная работа с данными (GitHub) - **Mastra** — AI-фреймворк для продакшна (GitHub)

4.12.10 Архитектура агентного RAG

SGR (Schema-Guided Reasoning) — фреймворк для агентов от neuraldeep:

- SGR Agent Core (GitHub) — 1k+ stars
- Запуск и философия | SGR vs Tools | Бенчмарки
- Агентный RAG на локальных моделях (Qwen3-30B-A3B)

Рыночные RAG-цепочки интеграторов — в открытых обзорах и кейсах встречаются собственные конвейеры:

- декомпозиция запроса (query decomposition);
- гипотетические документы для улучшения извлечения контекста (HyDE);
- двойной вызов с порогом сходства (DCD);
- schema-guided рассуждения (SGR);
- извлечение структуры из PDF (Marker, Docling);
- хранение метаданных в PostgreSQL и векторный слой (Qdrant, Chroma DB, PostgreSQL+pgvector).

Это иллюстрация зрелости рынка интеграции, а не требование воспроизвести все приёмы; пересечение с референс-стеком **Comindware** оценивают по целевому threat model и ТОМ.

4.12.11 Кейс: RAG для ФСК (Строительная компания)

По *«Хабр — red_mad_robot: кейс RAG для ФСК»*:

- **Задача:** RAG-чат-бот для ФСК — B2B
- **Срок:** внедрение за **2 месяца**
- **Масштаб:** корпус **> 1 млн токенов знаний**
- **Результат:** снижение нагрузки на команду поддержки и коммерческий департамент на **30–40%**
- **Архитектура:** Router-компонент + два workflow AI-агента
- **Фокус:** предотвращение галлюцинаций для минимизации репутационных рисков

Кейс полезен как публичный ориентир по архитектуре и диапазону эффекта; целевые KPI для заказчика фиксируются после пилота и замеров в его контуре.

Для руководства в этом блоке важны три контура контроля: **оценка качества до запуска**, **наблюдаемость в проде** и **безопасность / защитные механизмы** для снижения репутационных и комплаенс-рисков. Примеры инструментов для этих ролей приведены в подразделах *«Контур оценки качества»*, *«Безопасность»* и *«Продвинутая индексация, качество ответа и экономика поискового слоя»*.

4.12.12 Контур оценки качества

- **RAGAS** — метрики для RAG (Документация)
- **ARES** — автоматическая оценка RAG (GitHub)

4.12.13 Безопасность

- **NVIDIA NeMo Guardrails** — удержание бота в рамках темы (GitHub)
- **Lakera / Rebuff** — детекторы инъекций (Платформа), (GitHub)
- **Garak** — сканер уязвимостей LLM (GitHub)

4.12.14 Продвинутая индексация, качество ответа и экономика поискового слоя (@ai_archnadzor)

Материалы канала **@ai_archnadzor** задают ориентиры по логике рассуждений, графам, задержке (TTFT) и стоимости индексации; конкретный выбор паттерна для **корпоративный RAG-контур** фиксируется в ADR и комплекте отчуждения.

4.12.15 Disco-RAG: логический анализ вместо «плоского супа» из фактов

Концепция: внедрение теории риторических структур (RST). Модель понимает, где аргумент, где противоречие, где условие.

Архитектура:

- **Intra-chunk RST Trees:** для каждого чанка строится дерево связей (Nucleus/Satellite)
- **Inter-chunk Rhetorical Graph:** анализ отношений между чанками (дополняет/противоречит)
- **Discourse-Aware Planning:** план ответа на основе графа связей перед генерацией

Результат: превращает RAG из «читателя фактов» в «аналитика логики»

4.12.16 REFRAG: ускорение RAG в 30 раз

Проблема: огромный контекст убивает TTFT и «съедает» KV-кэш

Решение: сжатие «сырых» чанков в компактные эмбеддинги через RoBERTa + селективное расширение через RL-политику

Для кого: Tier-1 системы с миллионами запросов, где важна скорость

4.12.17 Cog-RAG: гиперграфы и «тематическое» мышление

Концепция: двойные гиперграфы (темы и сущности) для имитации человеческого подхода «от общего к частному»

Результат: Win Rate выше на **84,5%** по сравнению с обычным RAG

Вердикт: мощно, но дорого по индексации. Идеально для медицины и науки

4.12.18 HippoRAG 2: Экономим на графах в 12 раз

Инновация: Dual-Node архитектура (узлы-сущности + узлы-пассажи)

Экономика: снижение затрат на токены при индексации в **12 раз** (9 млн токенов vs 115 млн)

Стек: `pip install hipporag`

4.12.19 Topo-RAG: победа над «табличной слепотой»

Проблема: линеаризация таблиц в один вектор превращает данные в «семантический шум»

Решение: мульти-векторный индекс (каждой ячейке — свой вектор) + умный роутер

Результат: снижение галлюцинаций в цифрах с **45% до 8%**. Маст-хэв для финтеха и логистики

4.12.20 DSPy 3 и GEPA: Промышленный промпт-инжиниринг

DSPy 3: LLM как вычислительное устройство. Архитектор описывает Signatures, система генерирует и оптимизирует код промпта

GEPA (Genetic-Pareto Prompt Optimizer):

- Генетические алгоритмы для «скрещивания» лучших промптов
- Языковая рефлексия — модель анализирует свои ошибки текстом
- **Результат:** в **35 раз быстрее** MIPROv2, промпты в **9 раз короче**, на **10% точнее**

4.12.21 Новый «старый» OCR: NEMOTRON-PARSE, Chandra, DOTS.OCR

Модель	Фокус	Выход	Для кого
NVIDIA Nemotron (885M)	Скорость и Enterprise RAG	Markdown / LaTeX	Высоконагруженные RAG-системы
Chandra (~1B)	Рукопись и точность	MD / JSON / HTML	Архивы, оцифровка
dots.ocr (1.7B, MIT)	Агенты и лицензия	MD / HTML	Коммерческие SaaS

4.12.22 BitNet: 1-битные LLM для CPU-инференса

Концепция: 1-бит веса для Attention/MLP слоев + 8/16 бит для активаций

Почему важно:

- **Edge AI:** огромные модели теперь могут жить локально
- **Снижение TCO:** CPU-инстансы на порядок дешевле GPU
- **Гибридные кластеры:** обучаем на GPU, деплоим на CPU

Вердикт: не «убийца GPU» для обучения, но подтачивает монополию GPU на инференс

4.12.23 Doc-to-LoRA: Конец «налога на контекст»

Проблема: KV-кэш поглощает гигабайты VRAM для длинных контекстов

Решение: гиперсеть генерирует LoRA-адаптер из документа за один проход

Результаты:

- Потребление VRAM: **12 ГБ** → **50 МБ** (99% экономия)
- Скорость усвоения: **<1 секунда** (vs 100+ секунд при дообучении)
- Требования: **<2 ГБ VRAM** (vs 40+ ГБ для градиентных методов)

4.13 Инженерия обвязки для агентов

Обвязка в смысле отраслевой практики — это не замена сильной модели, а **среда исполнения** агента: что он видит в контексте, какие инструменты доступны, какие **детерминированные** проверки и петли обратной связи окружают генерацию. Подход описан и развивается в публичных материалах OpenAI (инженерия обвязки), Anthropic (длительные агентские сессии разработки), Thoughtworks / Martin Fowler (интерпретация и пробелы) и обзорах на русском языке (например, Хабр).

Там, где агент может инициировать **исполнение кода**, широкие сетевые вызовы или доступ к чувствительным API, класс **изоляции среды** и политики **сети и удостоверений** задаются по **модели угроз**, а не только по привычному стеку разработки — ориентиры и опора на NIST по границам контейнеров, **паттерны** песочницы под **PR** и **долгоживущую dev-среду**, таблица «вопрос → класс сценария» и **минимальный состав** платформы задач — см. *«Граница доверия, сеть и среда исполнения агента»*, *«Модель риска, паттерны среды и минимальный состав платформы»*.

4.13.1 Логические роли: планирование, исполнение, контроль (модель-контролёр)

В качестве переносимого шаблона удобно различать три **логические** роли (не обязательно три отдельные команды): **планировщик** формирует или уточняет спецификацию и границы задачи; **исполнитель** вносит изменения в код и

конфигурацию; **модель-контролёр** (часто отдельный запуск той же или иной модели по отдельному промпту) оценивает результат **независимо** от исполнителя. Anthropic показывает, что такое разделение снижает типичную для одного агента **самопохвалу** и поверхностное тестирование; при этом **модель-контролёр**, которая только выставляет вердикт по промпту, остаётся **склонной к завышенной оценке**, поэтому критичны **жёсткие пороги по критериям**, **эталонные примеры в промпте** модели-контролёра и **проверка действиями** (клики в интерфейсе, вызовы программного интерфейса, сверка состояния данных), а не одна только «самооценка» модели (Anthropic — Harness design for long-running application development).

Сопоставление с ТОМ из настоящего документа: планирование — зона **владельца продукта с ИИ** и архитектуры; исполнение — разработка и агенты, которые пишут код, под регламентом; проверка — **контроль качества, приёмочные сценарии и информационная безопасность** плюс регрессионные и **сквозные** тесты. Блоки про MERA, RAGAS, DeepEval и **модель-контролёр** остаются в силе: вердикт **по промпту** дополняет, но **не заменяет** согласованные приёмочные критерии и тесты.

4.13.2 Контекст в репозитории и «карта», а не энциклопедия

OpenAI и независимые обзоры сходятся в том, что **монолитный** сверхдлинный файл правил для агента вытесняет из контекста код и задачу, быстро устаревает и плохо проверяется автоматически. Практичнее держать **короткий** верхнеуровневый регламент (оглавление, куда смотреть) и детали — в структурированном каталоге документации и ADR; правило «для агента не существует того, что не закреплено в репозитории» переносится на знания о продукте и решениях (OpenAI — Harness engineering, Хабр — обвязка для агентов).

4.13.3 Архитектурные ограничения и обратная связь

Детерминированная часть обвязки — **линтеры, структурные тесты, явные границы модулей**; сообщения об ошибках целесообразно формулировать так, чтобы агент (или человек) сразу видел **как исправить** нарушение. Это согласуется с акцентом на **снижение пространства решений** для устойчивого AI-generated кода (Martin Fowler — Harness Engineering).

4.13.4 Длительные задачи: handoff, сброс контекста и компакция

На длительных агентских прогонах актуальны **структурированные артефакты передачи** между шагами и сессиями. Anthropic различает **компакцию** истории (сжатие на месте) и **полный сброс** контекста с явным handoff: второй вариант дороже по оркестрации и токенам, но снимает эффект «тревоги по контексту», когда

модель преждевременно сворачивает работу; выбор политики — предмет настройки обвязки, а не замена политики **ретенции** телеметрии и ПДн (Anthropic — Harness design for long-running application development, Anthropic — Effective harnesses for long-running agents).

4.13.5 Поведение продукта и «разрыв верификации»

Fowler справедливо отмечает, что в публичных описаниях обвязки сильнее прозвучивают **поддерживаемость** и внутренняя качество кода, а **проверка функционального поведения** перед пользователем должна быть явно заложена в методологию: приёмочные тесты, **сквозные** сценарии, согласованные с заказчиком, — в дополнение к обвязке (Martin Fowler — Harness Engineering).

4.13.6 Российский контур и ПДн

Если **сценарий проверки** использует браузерную автоматизацию, снимки экрана или прогон против стендов с чувствительными данными, действуют те же принципы, что и для телеметрии генеративного ИИ: **минимизация**, сроки хранения артефактов, контур хранения и матрица доступа — см. *«Периметр до LLM»*, *«Персональные данные и содержимое в телеметрии»*. Новые нормативные тезисы здесь не вводятся.

4.13.7 Отчуждение обвязки

При передаче клиенту в комплект имеет смысл включать **версионируемые** skills и регламенты сценариев, конфигурацию **MCP**, **CI** и **CD** под согласованный контур, **рубрики и промпты** для **модели-контролёра** и регламент периодической синхронизации документации с кодом («сборка мусора» / садовник документации в духе публичных практик) (OpenAI — Harness engineering, Хабр — обвязка для агентов).

4.13.8 Справочно: формализация процессов (BPMN 2.0) и генерация с помощью LLM

Машиночитаемый **BPMN 2.0 XML** уместен в методологическом комплекте для передачи знаний, согласования с владельцами процесса и аудита — **рядом с** рубриками и промптами для **модели-контролёра**, а **не вместо** кода и регламентов разработки.

Стандарт BPMN 2.0 разделяет **семантику процесса** и слой диаграммы (**BPMNDI**); при генерации языковой моделью без явных правил в промпте и

последующей проверки типичны пропуск визуального слоя, расхождение атрибута `bpmnElement` с `id` узлов и некорректные `sequenceFlow`.

Практично опираться на **детальные шаблоны промптов** (иллюстрация — «*Генерация BPMN 2.0 XML (промпт-шаблон)*») и **обязательную валидацию** открытием в Camunda Modeler, bpmn.io или эквиваленте перед включением артефакта в комплект передачи.

В организациях с **Confluence** схемы процессов нередко остаются в **плагинах** вики; для КТ и согласования **вне** страницы вики переносимым артефактом остаётся отдельный **файл BPMN XML**, согласованный с регламентом отчуждения.

Связь с FinOps: мультиагентные циклы и длительные прогоны разработки увеличивают **токены и wall-clock**; ориентиры по закладке в TCO — в *«FinOps и юнит-экономика нагрузки»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

4.13.9 Справочно: оценка управляемых песочниц и бенчмарки

Выбор **управляемой** среды для агента удобно вести по трём осям: **модель сессий** (эфемерность, снимки, время жизни), **модель сети** (дефолт egress и способ задания allowlist) и **модель размещения** (регион, контур заказчика, SaaS). Сравнение платформ и поставщиков **не** заменяет прогона на **реальной** нагрузке (репозиторий, зависимости, тесты, файловый и сетевой периметр); тривиальные микробенчмарки и одна лишь задержка **не** измеряют пригодность для **безопасного** исполнения. Детали, примеры **E2B / Modal / Daytona**, критерии приёмки, метрики прода и ссылки на **gVisor** и академическое исследование trade-off runtime — см. *«Управляемые песочницы, сравнение моделей и бенчмарки»*.

4.14 Практический опыт внедрения ИИ (red_mad_robot)

Инженерные сигналы и практики внедрения для оценки технологических рисков — см. Приложение D.

4.15 Российский рынок ИИ: текущее состояние и прогнозы (2024–2026)

4.15.1 Национальная стратегия развития ИИ

Указ Президента РФ №124 (февраль 2024):

- Поправки к Национальной стратегии развития ИИ до 2030 года
- Новые определения: «датасет», «большие генеративные модели», «модель ИИ», «сильный ИИ»

- Федеральный проект «Искусственный интеллект» включен в национальный проект «Экономика данных»
- Цель: более 11 трлн руб. влияния ИИ на ВВП к 2030 году

Финансирование (2025):

- 7.7 млрд руб. на федеральный проект «ИИ»
- Фокус: исследовательские центры, обучение специалистов (15 500 к 2030), здравоохранение, кибербезопасность

4.15.2 Создание офисов внедрения ИИ

Тренд 2025–2026:

- Массовое открытие офисов внедрения ИИ в российских компаниях
- Северсталь: ~30 человек в офисе ИИ, платформа DaVinci
- План на 2026: масштабное внедрение, первые автономные ИИ-агенты
- Рост вакансий с ИИ-навыками: **+89%** в 2025 и **+170%** YoY в I кв. 2026 к I кв. 2025 (hh.ru × PR DEV; таблица — *Приложение D «Общая картина рынка GenAI (red_mad_robot)»*)

4.15.3 Экономический эффект

Прогноз Yakov Partners (2025):

- Экономический эффект ИИ в РФ к 2030: ориентир **~8–13 трлн руб.** в год (по материалам Яндекса и Yakov Partners).
- Фокус: **60% эффекта** приходится на 5 секторов:
  - E-commerce
  - Телеком и медиа
  - IT и технологии
  - Строительство и недвижимость
  - Здравоохранение
- К 2030: ИИ-внедрение станет вопросом выживания для большинства компаний

Российские модели:

- Alice AI (ex-YandexGPT), GigaChat — конкурентоспособные ориентиры в линейке больших диалоговых моделей; у Сбера дополнительно доступны **открытые веса** GigaChat-3.1-Ultra и GigaChat-3.1-Lightning под **MIT** (Хабр, коллекция на Hugging Face).
- **86% компаний** используют open-source модели и fine-tuning

4.15.4 Применение ИИ-агентов

Статистика:

- **46% компаний** уже внедряют или тестируют автономные решения
- Сферы применения: аналитика, логистика, поддержка принятия решений

4.15.5 Карта российского рынка GenAI (обзор red_mad_robot, публичные материалы 2025)

Лаборатория **red_mad_robot** формирует отраслевую повестку через серию открытых исследований рынка GenAI (архив). Публикация ноября 2025 — **карта сегментов** российского рынка — даёт ориентиры для оценки адресуемого рынка и обоснования инвестиционных решений.

- **Управленческая сетка инвестиций.** При планировании бюджета разделяйте четыре зоны: **инфраструктура и вычисления** (CapEx), **базовые и прикладные модели** (лицензии, дообучение), **продукты и сервисы** (подписки, API), **интеграция и внедрение** (проектный бюджет, изменение процессов). Это разграничение задаёт язык смет, контрактов и отчётности для всех стейкхолдеров.
- **Рыночные данные подтверждают зрелость спроса. 86%** компаний, использующих генеративный ИИ, применяют open-source модели с дообучением под собственные задачи. **46%** уже внедряют или тестируют автономные ИИ-агенты, способные выполнять цепочки задач без участия человека. Эти ориентиры по выборке «Яков и Партнёры» и Яндекса (декабрь 2025, 150 технических директоров) определяют масштаб адресуемого рынка для поставщиков корпоративных решений.
- **Стресс-тест портфеля до 2030.** Оценивайте инвестиционные решения через три сценарных ветки:
- **Консервативный рост** предполагает ограниченный доступ к вычислительным ускорителям и усиление регуляторного контроля — портфель требует жёсткой селекции проектов с явным ROI.
- **Консолидация** вокруг ограниченного числа экосистемных игроков меняет конкурентную среду — ставка на партнёрства становится критичной.
- **Ускоренная индустриализация** при снятии узких мест по инфраструктуре и кадрам создаёт условия для массового масштабирования — здесь выигрывают пионеры с готовыми контурами внедрения. Применяйте этот шаблон на каждом переходе: PoC → Пилот → Масштабирование. Количественные параметры сценариев — в *«ИИ-рынок России (оценка IMARC)»*.

Организационный фактор как приоритет. Практика подтверждает: процессы, данные, компетенции и измеримость определяют результат сильнее, чем выбор конкретной модели. Поведенческие риски и барьеры — в *«Стратегия внедрения ИИ и организационная зрелость»* и *«Организационные и поведенческие факторы риска»*. Управление доверием к ИИ-системам (AI TRiSM) и экономика безопасности — в *«AI TRiSM и управление доверием»* и *«OpEx безопасности GenAI и агентов»*.

4.15.6 GenAI в маркетинговых командах крупных брендов РФ (опрос CMO, 2025)

Открытые материалы исследования **red_mad_robot × CMO Club Russia** описывают **массовое тактическое** использование GenAI при **низкой доле** системной интеграции и сильном разрыве между личной цифровой зрелостью маркетологов и корпоративными ограничениями (инфраструктура, ИБ, регламенты). Для корпоративного RAG и агентов это означает: владельцы маркетингового бюджета уже «прогреты» инструментами, но **управляемый контур** (данные, политики, оценка качества, телеметрия) остаётся полем конкуренции поставщиков платформы.

- **PoC → Пилот → Масштабирование:** публичный нарратив совпадает с логикой комплекта — от точечных сценариев к **оркестрации** процессов; при этом **отсутствие стратегии и прозрачного ROI** в открытых выжимках названо главным тормозом системного масштабирования.
- **РФ vs глобальная практика:** по тем же материалам, **ежедневное** использование GenAI у маркетинговых команд в РФ **выше**, чем в приведённых глобальных ориентирах, тогда как **event-, бренд- и бюджетные** сценарии отстают от мировых ориентиров — сигнал для **пилотов без ожидания мгновенного эффекта** и для явного плана обучения и гайдлайнов.
- **Роль CMO как оператора-оркестратора:** в исследовании подчёркивается сдвиг от «исполнителя» к связке **данные + технологии + креатив**; для **Target Operating Model** это стыкуется с выделением владельцев данных, промышленной обвязкой и **Utilization** как KPI зрелости — при этом отраслевые **доли ежедневного использования** из опроса маркетинга не подменяют целевые метрики TOM заказчика, а задают **сегментный ориентир спроса**.
- **Ключевые доли и формулировки опроса** — в *«Зрелость российского рынка GenAI»*; риски **утечки** vs **галлюцинаций** там разведены по разным пунктам опроса — см. *«Организационные барьеры и восприятие рисков (опрос CMO ×red_mad_robot, 2025)»* (связка с LLM02 и минимизацией телеметрии).

4.15.7 Публично описанные паттерны (финсектор)

Переносимые идеи из открытых инженерных и отраслевых материалов (не рекомендация конкретных поставщиков или продуктов) — ориентиры зрелости для внедрений в духе **корпоративный RAG-контур** и агентных сценариев:

- **Жизненный цикл моделей и дрейф:** закладывать деградацию качества (сдвиг признаков, целевой метки, качество данных) и автоматизировать переобучение и вывод в прод через шаблонизированный MLOps-пайплайн и конфигурируемые сценарии, чтобы портфель моделей не съедал растущую долю времени DS — Альфа-Банк, Хабр.
- **Высокая кардинальность тем в текстовых каналах:** масштабировать маршрутизацию обращений и разгрузку операторов через ML на большом числе тематик — классификация диалогов, Хабр.
- **MLOps и каскады моделей:** связывать подготовку данных, обучение и деплой в единый контур; в публикациях как пример стека упоминаются Airflow, Hadoop/Spark, MLflow, Kubernetes — MLOps и каскады, Хабр.
- **Внутренний RAG над регламентами:** для операционных ролей — поисково-дополненная генерация (RAG) по корпоративным базам знаний (инструкции, тарифы, продукты), выделенный RAG-сервис и регулярное обновление источников — «Открытые системы», 2025.
- **Instruction following и вызов инструментов:** в агентских сценариях критичны соблюдение формата ответа и многошаговый вызов инструментов; публично разобраны синтетические обучающие пайплайны, верифицируемые награды и защита от reward hacking — обновление LLM, Хабр.
- **Чат-канал под высокой нагрузкой:** фиксировать SLO по латентности и комбинировать векторизацию запроса, классификаторы и извлечение сущностей; расширять генеративный слой поэтапно, с пилотом на ограниченном наборе сценариев при большой матрице тем — CIO, 2024.

4.15.8 Российские облачные провайдеры для ИИ (экономический срез)

Сводные **цифры по токенам** (Cloud.ru, Yandex AI Studio, комплекты SberCloud, примечания MWS/Selectel) собраны в *«Тарифы российских облачных провайдеров ИИ»* в «Сайзинг и экономика (CapEx / OpEx / TCO)»; в этом документе — роли провайдеров, архитектура доступа к моделям и матрица API vs open weights (подраздел *«Российские облачные провайдеры ИИ»*).

Открытые веса GigaChat-3.1 (MIT, HF/GitVerse — см. Хабр) переносят основную стоимость в **инфраструктуру и эксплуатацию**; вилка TCO — в *«Открытые веса и API: влияние на TCO»*, «Сайзинг и экономика (CapEx / OpEx / TCO)».

4.15.9 Sovereign AI для предприятий

Тренды суверенного ИИ:

- Хранение данных внутри юрисдикции
- Разработка локальных моделей
- Снижение зависимости от иностранных технологий

Российская специфика:

- Государственная поддержка ИИ-инициатив
- Инвестиции в внутреннюю инфраструктуру ИИ
- Политики локализации данных
- Платформа SME.Russia: +35% рост предпринимателей, получивших поддержку через ИИ-рекомендации (2024–2025)

4.16 Методология Enterprise AI (Global Best Practices)

4.16.1 От «vibes» к измеримым результатам

Три измерения ROI:

1. **Вовлечённость** — кто использует ИИ-инструменты, как часто, для каких задач
2. **Компетентность** — качество использования
3. **Бизнес-результат** — связь использования с бизнес-результатами

Ключевые метрики:

- AI Leaders: **3–4× лучше** по продуктивности, инновациям, удовлетворенности сотрудников
- Организации с полным набором измерений: **5.2x выше уверенность** в ИИ-инвестициях

4.16.2 Эмпирика корпоративного внедрения (отчёт OpenAI, 2025; оговорки по выборке)

В «*The state of enterprise AI (OpenAI)*» (декабрь 2025; полный PDF — в параграфе«Источники» этого документа) OpenAI обобщает корпоративное внедрение по двум каналам: агрегированная телеметрия **enterprise-клиентов** и опрос **9 000** сотрудников в почти **100** организациях (данные обезличены и агрегированы, как заявлено на странице отчёта).

По формулировкам первоисточника, внедрение ускоряется в **ширину и глубину**: за год еженедельные сообщения в ChatGPT Enterprise выросли примерно **в 8 раз**, средний работник отправляет **на 30%** больше сообщений; использование структурированных сценариев (Projects, Custom GPTs) выросло **в 19 раз** с начала года — как сдвиг от разовых запросов к повторяемым процессам. Среднее потребление **токенов рассуждения** на организацию за **12** месяцев выросло примерно **в 320 раз** — сигнал для учёта **дорогих** режимов инференса в FinOps.

Отдельно фиксируется **разрыв** между «передовыми» и медианой: у работников около **95-го** перцентиля **в 6 раз** больше сообщений, чем у медианного сотрудника; у «передовых» организаций **в 2 раза** больше сообщений на место и более глубокое **межкомандное взаимодействие**. Для заказчиков в РФ — ориентир неравенства внедрения.

Указанные в отчёте **самооценки** эффективности (например, **75%** респондентов — улучшение скорости или качества, **40–60** минут экономии в день, разрез по департаментам на лендинге) относятся к **опросу** и не заменяют внутренний ROI-контур заказчика.

OpenAI также заявляет о высоком темпе новых возможностей продукта (порядка **одной** примерно **каждые три дня**) и о том, что основные ограничения для организаций смещаются к **готовности и внедрению**, а не к качеству модели или инструментам как таковым. Это согласуется с акцентом настоящего документа на TOM, оценке качества и наблюдаемости — см. *«Инженерия обвязки для агентов»* и *Приложении С «Безопасность, комплаенс и наблюдаемость»*.

Оговорки: материал отражает экосистему OpenAI; для применения в РФ требуется учёт 152-ФЗ, суверенитета данных и выбора контура. Отраслевые и географические акценты полезны для приоритизации гипотез.

4.16.3 Точка безубыточности инфраструктуры (break-even)

Количественное сравнение **TCO** для **GPU** (облако РФ vs закупка ускорителей), воспроизводимые допущения, формулы и таблица ориентиров — в *«TCO GPU: облако РФ против закупки»*, «Сайзинг и экономика (CapEx / OpEx / TCO)». Там же — вилки по провайдерам и оговорка, что **полный** on-prem TCO включает ЦОД, сеть, персонал и налоговую амортизацию, а не только цену карт.

Порог утилизации (управленческий ориентир): при **низкой** или сильно **пиковой** нагрузке выгоднее **аренда** облака; при **устойчивой высокой** утилизации (**порядка 60–70 %** и выше) и горизонте **нескольких лет** экономика чаще смещается к **собственной** инфраструктуре или colocation — см. *«TCO облако vs on-prem для ИИ»*.

Для КП пересчитывайте суммы по **фактическому** прайсу провайдера и курсу на дату сметы, см. *Курс USD*.

4.16.4 Методология внедрения ИИ (IBM Sovereign Core)

Ключевые компоненты суверенного контура

- **Идентичность и ключи внутри периметра:** учётные данные агентов, сервисные токены и криптографические ключи не покидают контур заказчика. Вся аутентификация и авторизация выполняются на инфраструктуре, физически и юридически находящейся в российской юрисдикции.
- **Управляемый инференс ИИ:** языковые модели и вспомогательные компоненты (эмбеддинги, реранкеры, защитники) работают на локальных GPU — в собственном или арендованном контуре заказчика. Запросы пользователей и корпоративные данные не передаются сторонним облачным провайдерам.
- **Журналы аудита и комплаенс внутри суверенной границы:** полная трассировка операций агентов, вызовов инструментов и обращений к данным хранится исключительно в аттестованном контуре. Это закрывает требования 152-ФЗ к локализации обработки персональных данных и обеспечивает готовность к аудиту регулятора без передачи логов за пределы периметра.

4.17 Практические кейсы из каналов

Практические кейсы внедрения и пользовательские сценарии для обоснования ROI — см. Приложение D.

4.18 Рекомендации по внедрению ИИ для клиентов

4.18.1 Методология «двенадцать факторов» для ИИ

1. **Кодовая база:** одна кодовая база — много развёртываний
2. **Зависимости:** явное объявление всех зависимостей
3. **Конфигурация:** все параметры — через переменные окружения
4. **Подключаемые ресурсы:** векторные хранилища, API LLM — как внешние сервисы
5. **Сборка, релиз, запуск:** строгое разделение этапов
6. **Процессы:** процессы без сохранения состояния (stateless)
7. **Привязка к порту:** самодостаточные сервисы
8. **Параллелизм:** масштабирование за счёт процессов
9. **Устраняемость:** быстрый старт и корректное завершение работы

10. **Паритет сред:** минимизация различий между разработкой и эксплуатацией
11. **Журналирование:** поток событий как единый источник для эксплуатации и разбора инцидентов
12. **Административные задачи:** разовые операции в том же стеке

4.18.2 Фазы внедрения

Фаза	Продолжительность	Цель	Результат
PoC	2–4 недели	Проверка гипотезы	MVP, данные для ROI
Пилот	1–3 месяца	Валидация в продуктивной среде	Интеграция, первые пользователи
Масштаб	3–12 месяцев	Масштабирование	Внедрение по всей организации
Оптимизация	Постоянно	Оптимизация совокупной стоимости владения (TCO)	Снижение затрат, повышение ROI

4.19 Рекомендованный план 30/60/90 дней

- **0–30 дней:** выбор 2–3 приоритетных бизнес-кейсов; PoC на связке **корпоративный RAG-контур** + **сервер инференса MOSEC**; при кейсах платформы — пилот **агентного слоя Comindware Platform**; замер базового ROI.
- **30–60 дней:** расширение пилота (**корпоративный RAG-контур**, при необходимости **агентный слой Comindware Platform**) на департамент, внедрение наблюдаемости (ориентир по стеку — **Arize Phoenix**), начало обучения команды клиента.
- **60–90 дней:** финализация масштабирования LLM на **инференс на базе vLLM**, стабилизация **корпоративный RAG-контур** и (при внедрении) **агентный слой Comindware Platform**, подготовка комплекта отчуждения, аудит на соответствие новому закону об ИИ.

4.19.1 Справочно: узкий безопасный MVP контура исполнения агента (ориентир ~30 дней)

Если в фокусе **исполнение кода** или широкие **инструменты** агента, первый месяц разумно трактовать как вывод **безопасного MVP** под **один** узкий сценарий (например, только PR-агент или только аналитика), а не как строительство универсальной платформы. Поэтапный ориентир по неделям (модель угроз → брокер секретов, тип среды, deny-by-default сеть → снимки, артефакты, аудит →

враждебные сценарии и пилот), **критерии готовности**, **вопросы для дискуссии** о двух классах сред и **выводы** по доверию к исполнению в инфраструктуре — см. *«Безопасный MVP контура исполнения за 30 дней, дискуссия по средам и выводы»*.

4.20 Обоснование рекомендаций (метод исследования)

Управленческий цикл формирования рекомендаций:

- **Границы** — фиксируем вопрос заказчика, допущения по контуру данных и целевые KPI.
- **Доказательства** — привлекаем нормативные и отраслевые источники, документацию стеков (LangChain, vLLM, MOSEC), публичные кейсы.
- **Триангуляция** — на каждый существенный тезис не менее трёх независимых опор, как минимум одна из них первого приоритета (регулятор, стандарт, официальная документация вендора). Количественные оценки (ROI, CapEx, эффекты SLA) сопровождаем ссылкой на первоисточник или пометкой «оценочная модель».
- **Синтез** — формируем варианты решений с компромиссами, пригодные для управленческих решений.

Журнал доказательств (шаблон строки): тезис | источник | тип (норма / стандарт / вендор / исследование / кейс) | дата | надёжность (высокая / средняя / низкая) | комментарий.

Конфликт источников: фиксировать расхождение и условия, при которых верна каждая оценка (например, разные границы TCO или дата тарифа).

4.20.1 Сигналы из открытых каналов и сообществ

Иногда в тексте используются выдержки из **публичных** профессиональных каналов (в том числе мессенджеров). Они показывают **рыночную и инженерную повестку** и по возможности снабжены ссылкой на первоисточник.

Как этим пользоваться на уровне решений: считайте такие фрагменты **дополнительным сигналом** — повод уточнить позицию своей команды, юристов и поставщиков. Они **не** являются нормой права, официальным тарифом, обязательством вендора или заменой договору. Перед утверждением бюджета или подписанием контракта **перепроверьте** дату первоисточника, актуальные прайс-листы и соответствие вашим требованиям комплаенса (в том числе 152-ФЗ, режим критической информационной инфраструктуры, реестры программного обеспечения и иные применимые нормы).

4.20.2 Что передаётся клиенту при отчуждении знаний

В рамках отчуждения заказчик получает **согласованный комплект**: методологию внедрения, перечень передаваемых артефактов (при необходимости — код, эксплуатационная и проектная документация, регламенты, программа обучения под контур заказчика), модель сопровождения и требования по комплаенсу — то, что можно зафиксировать в договоре и передать как часть поставки.

Три измерения отчуждения (модель BOT):

Измерение	Состав	Типичный срок
Техническое	Исходный код, конфигурации, модели, IaC	18–42 мес.
Методологическое	Эксплуатационные регламенты, процессы, тестирование, оценка качества	12–24 мес.
Человеческое	Обучение, сертификация, передача компетенций	После масштабирования

Примечание

Для сложных агентных AI-систем типичный горизонт полного BOT-цикла составляет **24–60 месяцев**.

Модель Build–Operate–Transfer обеспечивает поэтапную передачу: от проектирования через эксплуатацию к полному владению активом.

Смысл для руководства: учебные подборки, внутренние справочники по стеку и рабочие материалы исполнителя **не входят** в объём передачи **сами по себе**, пока они **отдельно** не перечислены в соглашении. Без явного включения не стоит исходить из того, что «передаётся вся внутренняя база знаний» вместе с решением.

4.21 Методология ROI для ИИ-проектов

4.21.1 Три измерения ROI

1. **Вовлечённость:**

- Кто использует ИИ-инструменты?
- Как часто?
- Для каких задач?

1. **Компетентность:**

- Качество использования
- Глубина применения
- Сокращение времени на задачи

1. **Бизнес-результат:**

- Связь использования с бизнес-результатами
- Измеримые метрики
- ROI в денежном выражении

4.21.2 Метрики успеха

Лидеры против отстающих в ИИ:

- Продуктивность: **3–4× выше**
- Инновации: **3–4× выше**
- Удовлетворённость сотрудников: **3–4× выше**

Уверенность в ИИ-инвестициях:

- Организации с полным набором измерений: **в 5,2 раза выше уверенность**

4.21.3 Экономический эффект ИИ в РФ

Прогноз Yakov Partners (2025):

- Экономический эффект к 2030: ориентир **~8–13 трлн руб.** в год.
- 60% эффекта — 5 секторов: E-commerce, Телеком, IT, Строительство, Здравоохранение
- К 2030: ИИ-внедрение станет вопросом выживания

Применение ИИ-агентов:

- 46% компаний уже внедряют или тестируют автономные решения
- Сферы: аналитика, логистика, поддержка принятия решений

5. Сайзинг и экономика (CapEx / OpEx / TCO)

5.1 Обзор

Финансовая база для обоснования инвестиций в ИИ: тарифные сетки российских облачных провайдеров, диапазоны CapEx/OpEx, дерево факторов стоимости и сценарный анализ TCO для различных архитектурных моделей.

Практический смысл: обеспечение CFO/CIO доказательной базой для выбора модели владения и защиты TCO-обоснования. Приведенные показатели являются оценочными диапазонами для принятия стратегических решений. Финальные бюджетные обязательства фиксируются после верификации нагрузочного профиля.

Для переговоров и управленческих брифов: переносите бюджетные вилки и пороги утилизации, не отдельные технические таблицы. Для составления смет проводите замеры на стенде заказчика и получайте актуальные прайсы провайдеров.

5.2 Концепция финансовой модели

- **Ситуация:** совокупная стоимость владения (TCO) GenAI-системами формируется из затрат на токены, инфраструктуру GPU, интеграционные работы и непрерывное сопровождение. Профиль нагрузки **Comindware** определяется взаимодействием корпоративного RAG-контура и агентного слоя платформы.
- **Проблема:** стоимость вспомогательных моделей (эмбеддинг, реранкинг, модерация) и мульти-бэкенд архитектур часто недооценивается. Колебания валютных курсов и скрытые статьи (безопасность, регрессионное тестирование) критически влияют на выбор между облачной арендой и собственным контуром.
- **Задача:** определение оптимальной модели финансирования (Cloud RF / On-prem / Hybrid) и расчет целевых диапазонов CapEx/OpEx на горизонте 1–3 лет.
- **Решение:**
  - Инициируйте внедрение с облачного PoC.
  - Переход к оценке безубыточности on-prem целесообразен при устойчивой утилизации **>60%** и горизонте планирования от 3 лет.
  - Используйте сценарный сайзинг, включая в OpEx затраты на наблюдаемость (накопление трасс, аудит) и пре-LLM обработку (маскирование ПДн, предиктивная фильтрация).
  - При измерении бизнес-эффекта учитывайте организационную зрелость.

5.2.1 Управленческие компромиссы

Выбор модели финансирования определяет долгосрочный TCO и операционную гибкость — ключевые компромиссы по каждому варианту:

- **SMB / департамент** — низкий CapEx, оплата по токену vs меньше контроля над данными и вендор-лок на API.
- **Enterprise** — инвестиция в GPU и команду LLMOps vs снижение долгосрочного TCO и суверенитет.
- **Гибрид** — гибкость vs сложность учёта и сопровождения двух контуров наблюдаемости.
- **Открытые веса (MIT) vs управляемый API** — без счётчика токенов vs CapEx/аренда GPU и инженерия; вилка разобрана в разделе *«Открытые веса и API: влияние на TCO»*.

Примеры метрик: стоимость 1 млн токенов по выбранному провайдеру, утилизация GPU, полные 3-летние TCO on-prem vs cloud, чувствительность к курсу и пошлинам (раздел по РФ).

Ключевые выводы:

- Для SMB оптимальны российские облака (ориентир от ~12 руб./млн токенов на отдельных линейках, см. таблицы тарифов).
- Для крупного Enterprise решение о переходе к on-prem или гибриду принимайте не по «быстрому сроку окупаемости», а по **устойчивой утилизации**, горизонту владения и **полному TCO** с учётом энергии, персонала и амортизации.
- Формат TOON используется вместо JSON и значительно снижает расход токенов при структурированном обмене данными; фактический эффект зависит от данных и токенизатора — см. *«Оптимизация затрат на инференс»*.

5.2.2 Ролевой фокус ЛПР

Сводная матрица приоритетов по ролям ЛПР — *Стратегическое резюме: матрица аргументов*.

Финансовые акценты для решения по модели владения и бюджету:

- **CFO:** границы CapEx/OpEx, пороги утилизации, чувствительность к курсу и прайсу.
- **CIO/CTO:** стоимость архитектурных вариантов (cloud/on-prem/hybrid), эффект на SLA и операционную нагрузку.

- **CEO/CRO/CPO/CISO:** экономические последствия выбранной модели владения и ограничений комплаенса.

Курс USD — *«Валюта и правила для коммерческих предложений»*.

5.2.3 Матрица стратегических решений (рыночные ориентиры)

Модель	Преимущества	Ограничения	Бюджетный коридор	Рекомендация
Управляемый сервис (SaaS)	Высокая скорость запуска, отсутствие CapEx	Зависимость от вендора, ограничения ИБ	**~200–250 тыс. руб./мес.**	Сценарии PoC и пилот
Локальное размещение (On-Premise)	Полный контроль данных, соответствие КИИ	Значительный CapEx, потребность в экспертизе	**CapEx узла: ~7–11 млн руб.**	Enterprise и госсектор
Гибридная модель	Оптимизация затрат, гибкость масштабирования	Сложность сетевой оркестрации	**Интеграция: ~0,3–1,5 млн руб.**	Средний и крупный бизнес

Ориентир РБК (рынок): вилки в таблице отражают публичный разбор *РБК Education — во сколько обойдётся ИИ-агент*: порядок величин для **переговоров о бюджете**, а не готовая смета под **отдельный** контур ПДн/КИИ, полный объём интеграций и постгарантийное сопровождение.

5.2.4 Компоненты стоимости внедрения ИИ-агента

1. **Инфраструктура:** облачные токены vs GPU серверы.
2. **Оркестрация:** разработка логики управления.
3. **Бизнес-сценарии:** проработка промптов и workflow.
4. **Интеграции:** CRM, ERP, внутренние БД.
5. **Эксплуатация:** техническая поддержка и дообучение.
6. **Эффективность:** юнит-экономика одного запроса.

5.2.5 Контрольные метрики для инвестиционного решения

Метрика	Значение / Диапазон	Значение для принятия решения
Порог перехода к on-prem	>60% утилизации GPU	При устойчивой утилизации выше 60% следует сравнивать полное TCO on-prem и cloud. Ниже этого порога облако, как правило, экономически выгоднее.
Ориентир SaaS (рынок)	200–250 тыс. руб./мес	Используйте как рыночный бюджетный коридор для этапа PoC и пилота. Цифра не является сметой под конкретный проект.
Ориентир CapEx on-prem узла	7–11 млн руб.	Это входной инвестиционный порог для одного производительного GPU-узла. Умножайте на количество узлов и добавляйте стоимость инфраструктуры и интеграции.
3-летний break-even по GPU	55–75% утилизации	Ключевой управленческий ориентир. При утилизации выше 65–70% и горизонте владения 3+ года on-prem или гибрид обычно выигрывает по TCO.

Решения по бюджету и модели владения принимаются на основании этих метрик. Отклонения допустимы только при явной фиксации причины и даты пересчёта.

5.3 Рыночный контекст

5.3.1 Рынок AI: статистика a16z (март 2026)

Ориентир рынка: рейтинги трафика и долей показывают глобальную динамику потребления GenAI. Эти данные помогают оценить **относительную популярность** моделей для понимания рыночных трендов, но для клиентского бюджета применяйте **фактические тарифы РФ** и специфику локального или on-prem-контура — глобальные метрики не транслируются напрямую в стоимость российских решений.

Источник: *a16z Top 100 AI Apps 6th Edition*

5.3.2 Распределение глобального рынка (веб-трафик, январь 2026)

Модель	Трафик vs ChatGPT	Доля рынка	Комментарий
ChatGPT	1,0x (база)	**52%**	900 млн еженедельных активных пользователей
Gemini	~0,37x	**19%**	#2 по трафику, рост платных подписчиков +258% г/г
DeepSeek	~0,35x	**18%**	#3 по трафику, 316 млн визитов/мес (янв 2025)
Qwen (Alibaba)	~0,12x	**6%**	>180 млн MAU в Qwen Chat + 700+ млн загрузок на HF (март 2026)
Kimi (Moonshot)	~0,04x	**2%**	~36 млн MAU (окт 2025), китайский «ChatGPT»
Claude	~0,033x	**2%**	#4, рост платных подписчиков +200% г/г

Примечание: Qwen и Kimi — преимущественно API/developer-модели, не представленные в a16z Top 100 (ориентирован на потребительские веб-приложения). Их значимость — в экосистеме разработчиков и корпоративном API. Сумма долей ≈ 99% (округление).

Ключевой вывод: ChatGPT доминирует (в 2,7 раза больше Gemini по веб-трафику), но Gemini и Claude растут быстрее. При этом ~20% пользователей ChatGPT также используют Gemini — использование нескольких платформ становится нормой.

Для КП: используйте эту информацию для контекста о трендах, но бюджетные расчёты стройте на российских тарифах (см. *«Тарифы и провайдеры РФ»*).

5.3.3 География использования ИИ

Страна	Ранг
Сингапур	1
ОАЭ	2
Гонконг	3
Южная Корея	4
США	20

Разрыв между созданием и использованием ИИ

США создали большинство ИИ-продуктов, но по использованию стоят на 20-м месте.

5.3.4 Структурные изменения рынка

- **Три мира:** Запад, Китай, РФ (политика формирует контуры)
- **Китайский подход:** массовое внедрение ИИ в экономику — см. *«AI + Economy: китайская модель»*
- **Генерация изображений:** Midjourney выпал из топ-10 (#46) — рынок перенасыщен
- **Генерация видео:** консолидация, сокращение числа игроков
- **Аудио-генерация:** стабильный сегмент — Suno, ElevenLabs сохраняют позиции

5.3.5 Рынок GenAI в России

Ориентир: российский рынок GenAI значительно меньше глобального, но растёт опережающими темпами благодаря регуляторной поддержке и локализационным требованиям.

5.3.6 Распределение AI-сервисов среди россиян (ВЦИОМ, 2025)

ИИ-сервис	Доля пользователей	Комментарий
ChatGPT (OpenAI)	27%	Лидер, несмотря на ограничения
YandexGPT / Алиса AI	23%	#2, экосистема Яндекса
DeepSeek	20%	Китайский прорыв 2025
GigaChat (Сбер)	15%	Суверенный AI
Шедеврум (Яндекс)	11%	Генерация изображений
Прочие (Qwen, Kimi, Perplexity, Grok, Claude, Midjourney и др.)	~4%	Многие требуют обходных путей (прокси, агрегаторы)

Примечание по методологии: ВЦИОМ фиксирует множественный выбор — респондент мог указать несколько сервисов. Сумма 96% означает, что часть пользователей применяет 2–3 сервиса одновременно. Китайские модели (Qwen, Kimi) популярны именно через российские агрегаторы (AITUNNEL, AllTokens) как обход санкций.

Выводы: ChatGPT лидирует даже в России, но локальные модели (YandexGPT, GigaChat) занимают значимую долю (~49% суммарно). Китайские модели — третья сила после американских и российских. Для корпоративных внедрений в РФ рекомендуется **двойной трек**: публичные API для PoC, российские провайдеры для production.

5.3.7 Барьеры и эффекты внедрения (глобальные показатели)

По данным глобальных исследований (McKinsey, Stanford, EY, OECD), ключевые барьеры и эффекты:

- **Барьеры (качество и безопасность): 40–50%** компаний отмечают проблемы качества и галлюцинаций как значимый барьер; **45–60%** выражают опасения по поводу утечки данных и безопасности (Stanford: +56,4% инцидентов год к году; Cisco: 60% обеспокоены уязвимостями).
- **Эффекты (производительность):** опросы EY и Deloitte фиксируют **восприятие** роста продуктивности в диапазоне **40–66%**, но такие оценки нельзя напрямую сопоставлять с макроуровнем. OECD моделирует **долгосрочный** вклад ИИ в annual aggregate labour productivity growth на горизонте **10 лет**: для наиболее AI-exposed экономик G7 диапазон составляет примерно **0,4–1,3 п.п. в год** (OECD, 2025). Для пилотов и КП это означает:

компании чаще видят локальное ускорение задач раньше, чем подтверждённый макроэффект.

5.3.8 Зрелость российского рынка GenAI

Ключевые метрики (сверены с глобальными показателями):

- **Внедрение: 85–90%** компаний используют GenAI (глобальный бенчмарк: McKinsey 88%); системно интегрировали в процессы — **около трети**.
- **Барьеры (дополнительно к глобальным): 53%** отмечают необходимость доработки контента; **49%** — шаблонность результатов.
- **Перспективы: 85%** респондентов считают GenAI ключевым фактором трансформации на горизонте **трёх лет**.

Полный разбор — в *Приложении D «Рыночные и технические сигналы»*.

5.3.9 Объём и динамика рынка GenAI и ИИ в России

Ниже приведены оценки именно **сегмента генеративного ИИ**; их не следует смешивать с более широкими оценками **ИИ-рынка в целом** и тем более с оценками **экономического эффекта** от внедрения ИИ, которые рассматриваются в следующем разделе.

Год	GenAI (млрд руб.)	Весь AI (млрд руб.)	Источник / Примечание
2024	**13**	**~423**	IMARC (базовый год)
2025	**58**	**~533**	IMARC forecast, рост GenAI ×4,5
2026	**~95**	**~680**	Statista + экстраполяция
2030	**778**	**~2100**	Прогноз (IMARC/РБК)
2033	**~1360**	**~3455**	IMARC

Ключевые метрики: - CAGR (весь AI): **26,5%** (IMARC, 2025–2033) - CAGR (GenAI): **19,04%** (IMARC, 2025–2033) - GenAI составляет ~11% от всего AI-рынка в 2025 году (58/533)

Концентрация рынка: топ-5 игроков (Яндекс, Сбер, Т-Технологии, ВК, Касперский) контролируют **~95%** выручки AI-сегмента. Остальные работают в нишевых сегментах или B2B.

Ограничения данных по российскому рынку: ряд западных аналитических агентств (Gartner, IDC, Statista) прекратили публикацию отдельных отчётов по российскому рынку после 2022 г. Часть приведённых цифр по рынку РФ основана на

ограниченном объёме данных из российских источников (Коммерсантъ, ВЦИОМ, Yakov & Partners, NTI/MIPT).

Источники: ВЦИОМ, Коммерсантъ, Digital Budget, Smart Ranking, CNews, IMARC, NTI/MIPT. Оценку **экономического эффекта** ИИ к 2030 году (в трлн руб.) см. отдельно в *«Методологии разработки и внедрения ИИ»*.

Разные методологии

- **IMARC** оценивает весь рынок ИИ в **~423 млрд руб.** в 2024 году
- **NTI/MIPT** включает смежные технологии — **1,15 трлн руб.** в 2024 году
- **Consainsights** даёт альтернативную оценку **~196 млрд руб.** (2024)
- **Statista** оценивает GenAI в **~20 млрд руб.** (2024) → **~353 млрд руб.** к 2030

Для КП используйте диапазон **0,4–1,2 трлн руб.** с обязательным указанием источника. Разница объясняется границами охвата (весь ИИ / только GenAI / ИИ + смежные технологии).

5.3.10 Драйверы роста

- Государственные инвестиции
- Enterprise adoption
- Технологические стартапы
- Приоритетные отрасли: финансы, здравоохранение, промышленность, оборона

Согласование с сегментными оценками: агрегированная оценка IMARC выше и сегментные ориентиры РФ ниже (в рублях) — разные методологии; не суммируйте без сверки границ рынка.

5.3.11 Сегментные ориентиры РФ (GPU-облако, B2B LLM)

- **Облачные сервисы с GPU:** по данным **Межведомственного научно-аналитического центра** (МНИАП), приводимым в Ведомостях, рынок в **2024** мог вырасти примерно в 1,5 раза до **~17,1 млрд руб.** — ориентир **ёмкости** инфраструктурного слоя с ускорителями в облаке, не дублирование CAGR всего рынка ИИ из IMARC.
- **Продукты на базе LLM для бизнеса (B2B):** по материалам РБК со ссылкой на оценку **MTS AI**, объём российского рынка LLM-продуктов для бизнеса в **2024** оценивался примерно в **35 млрд руб.**; структура on-prem vs облако и темпы роста — в том же первоисточнике при планировании **отдельной** строки портфеля.

5.3.12 Суверенный ИИ в России

Ключевые тренды:

- Хранение данных внутри юрисдикции
- Разработка локальных моделей
- Снижение зависимости от иностранных технологий
- Интеграция с государственными платформами (Gosuslugi, SME.Russia)

5.3.13 Агентный код-ревью (на примере Claude Code Review)

Для engineering-контура это сигнал появления отдельной статьи расходов на **автоматизированный review**: часть проверки кода начинает тарифицироваться как запуск агентного анализа, а не как «встроенная» функция IDE. Практический смысл для экономики внедрения — учитывать такие операции в **OpEx разработки** и не переносить цену review один в один в контур заказчика без поправки на доступность сервиса, режим обработки кода и требования ИБ.

Иллюстративный ориентир по цене: **~1275–2125 руб.** за проверку.

5.4 Тарифы и провайдеры РФ

Единый источник цифр: все таблицы с **руб./млн токенов** и **руб./час GPU** в этом разделе — опорный ориентир. Для КП всегда используйте актуальные прайсы.

5.4.1 Российские модели

Провайдер	Модель	Вход (₽/млн)	Выход (₽/млн)	Примечания
Cloud.ru	GigaChat3-10B-A1.8B	12,2	12,2	Evolution FM
Сбер	GigaChat Lite	65	65	Sync; async: 32,5
Сбер	GigaChat Pro	500	500	Sync; async: 250
Сбер	GigaChat Max	650	650	Sync; async: 325
Яндекс	YandexGPT Lite	200	200	
Яндекс	YandexGPT Pro 5.1	800	800	Sync; async: 410
Яндекс	Alice AI LLM	500	1200	

Сбер и Cloud.ru: не путать

Строки **Сбера** в таблице — тарифы **GigaChat API** для юрлиц.

Строка **Cloud.ru** — это тариф **конкретного SKU** в **Evolution Foundation Models**, а не общий эквивалент тарифов Сбера по всем моделям.

Линейка **GigaChat 3** также доступна как открытые веса **MIT** на Hugging Face; для ТСО это отдельный сценарий self-hosted.

5.4.2 Китайские модели (доступны в РФ)

Модель	Вход (₽/млн)	Выход (₽/млн)	Контекст	Источники
GLM-4.6	57	228	200K	Cloud.ru
GLM-5 / GLM-5.1	50	291	204K	AITUNNEL, OpenRouter
MiniMax-M2.5	21	103	200K	Cloud.ru, OpenRouter
MiniMax-M2.7	36	141	204K	AITUNNEL, OpenRouter
Qwen3.5-32B	14	14	128K	OpenRouter
Qwen3.5-235B (MoE)	18	18	128K	Cloud.ru
Qwen3.6 Plus Preview	0	0	1M	OpenRouter
Qwen3 Max	66	332	262K	OpenRouter
Kimi K2.5	64	255	200K	AITUNNEL
Xiaomi MiMo-V2-Flash	8	25	262K	OpenRouter
Xiaomi MiMo-V2-Pro	85	255	262K	OpenRouter
Xiaomi MiMo-V2-Omni	26	77	262K	OpenRouter

Расчёт средних цен

Цены рассчитаны как среднее между российскими (AITUNNEL, Cloud.ru) и глобальными (OpenRouter) провайдерами.

Qwen3.5 и GLM-5.1 — актуальные версии 2026.

Qwen3.6 Plus — **бесплатно** в OpenRouter в тестовом режиме (1M контекст, март 2026).

Xiaomi MiMo-V2-Pro — флагманская модель (1T+ параметров, #7 на Artificial Analysis, март 2026).

5.4.3 Глобальные модели (требуется VPN в РФ)

Модель	Вход (₽/млн)	Выход (₽/млн)	Провайдер
GPT-5.4	850	2550	OpenAI
Claude Sonnet 4.6	255	1275	Anthropic
Claude Opus 4.6	425	2125	Anthropic
Gemini 3.1 Pro	170	1020	Google

5.4.4 Российские API-агрегаторы (аналоги OpenRouter)

Для доступа к глобальным моделям из РФ без VPN используются российские агрегаторы, выступающие маршрутизаторами (роутерами) API:

Провайдер	Модели	Особенности	Сайт
AITUNNEL	400+ (GPT, Claude, Gemini, DeepSeek, Qwen, MiniMax)	OpenRouter-совместимый API, оплата в ₽, без VPN	aitunnel.ru
AllTokens	80+ провайдеров, 400+ моделей	Единый API, оплата в рублях	alltokens.ru
RouterAI	GPT, Claude, Gemini, Grok	Доступ без VPN, оплата картой или юрлицам	routerai.ru
ZvenoAI	LLM	OpenAI-совместимый API, токены по себестоимости	queryrouter.ru
ЦАРЬ ROUTER	Российские модели	Единый ключ ко всем российским AI-моделям	tsarrouter.ru

Использование в КП

Российские агрегаторы — решение для **разработки и PoC** с глобальными моделями.

Агрегаторы удобны при миграции с OpenRouter (AITUNNEL совместим на уровне API).

Для продуктового развёртывания в контуре с ПДн/КИИ используйте российских провайдеров (GigaChat, YandexGPT, Cloud.ru Evolution).

5.4.5 Инфраструктурные провайдеры РФ

Провайдер	Услуги	Особенности
Cloud.ru	Evolution Foundation Models (20+), GPU-инфраструктура	Единый контур: API моделей + GPU
GPUCloud	Облачные серверы с GPU, K8s	IaaS/PaaS для ML-проектов
MWS GPT (МТС)	Корпоративные AI-пакеты под ключ	Полный цикл: инфраструктура + модели + поддержка
Reg.cloud (Reg.ru)	Bare metal GPU, виртуальные машины	Маркетплейс аренды GPU в РФ
Skystark	Выделенные серверы RTX 4090, Tesla, VPS	Фокус на gaming/ML GPU
Selectel	Каталог базовых моделей, GPU-инфраструктура	Оплата по ресурсам, широкий выбор GPU
VK Cloud	GPU-инфраструктура, Managed ML	Интеграция с экосистемой VK
СберОблако	VM с GPU, GigaChat API, AI-сервисы	Экосистема Сбера: инфраструктура + GigaChat API
Яндекс Облако	AI Studio, Compute Cloud (GPU A100/H100), Vertex AI	Полный стек: GPU-инфраструктура + YandexGPT API

5.4.6 Открытые веса и API: влияние на TCO

Self-hosted убирает счётчик токенов, перенося затраты в GPU, энергию, персонал — *описание модели (Хабр, Сбер)*.

- **GigaChat 3.1** доступен под MIT на Hugging Face (ai-sage/GigaChat3.1-702B-A36B, ai-sage/GigaChat3.1-10B-A1.8B).
- **YandexGPT 5 Lite** доступен под кастомной лицензией (бесплатно до 10 млн токенов/мес).

5.5 Модель затрат

Стоимость ИИ-решения складывается из трёх уровней:

1. **Инфраструктура:** вычисления на GPU (VRAM, пропускная способность), хранилище векторных БД (ChromaDB, Qdrant, PostgreSQL + pgvector), сеть (исходящий трафик, задержки).
2. **Инференс:** стоимость токенов (вход/выход), формирование эмбеддингов и ранжирование, перерасход памяти фреймворков (vLLM/MOSEC).
3. **Эксплуатация:** поддержание LLMOps/DevOps, мониторинг и наблюдаемость (Arize Phoenix, Langfuse, LangSmith), хранение трасс и аудит-журналов агентских цепочек, обновление и дообучение моделей.

5.5.1 Инфраструктура и наблюдаемость: статьи затрат

Ниже — **статьи OpEx/CapEx** вокруг сервисов, которые отдают LLM/RAG наружу: потоки событий, хранение журналов, нагрузочное тестирование. Иллюстрация переносима на любой стек; источники — открытые описания проектов Ozon Tech.

Инструменты для журналов и событий:

- **Хранение и поиск журналов:** seq-db позиционируется как масштабируемая СУБД для журналов событий — в дереве затрат это **диск, реплики, запросы** и связанный **персонал** сопровождения наблюдаемости.
- **Пайплайны событий:** file.d — сбор и обработка событий (в т.ч. в Kafka, Kubernetes), интеграция с Prometheus; влияет на **OpEx** вычислений и сети.
- **Нагрузочное тестирование gRPC:** framer — высокопроизводительный генератор нагрузки для gRPC; полезен при проверке **SLO и ёмкости** обвязки инференса.

Наблюдаемость: сценарии размещения (ориентир для РФ):

Сценарий	Статьи затрат	Применимость в РФ
Зарубежный SaaS	Абонентская плата по объёму трасс/сидов, низкий CapEx	Допустим при правовой оценке и ДПО; для ПДн в проде не является решением по умолчанию
Self-hosted	CapEx/аренда ВМ и СХД, занятость специалистов	Соответствует ожиданиям **локализации** и контроля журналов
Облако РФ	OpEx по потреблению + сетевой трафик телеметрии	Компромисс: ниже CapEx, договорный контур РФ
Гибрид	Инжиниринг конвейера + отдельное хранилище	Соотносится с рекомендациями **OpenTelemetry** не хранить полный текст промптов в спанах

Метрики `gen_ai.client.token.usage` и `gen_ai.client.operation.duration` из конвенций **OpenTelemetry GenAI** (спецификация) дают общий язык с биллингом API и упрощают аллокацию FinOps по продуктам. Методологическая связка — см. *«Промышленная наблюдаемость LLM, RAG и агентов»*.

Распределение затрат между интегратором и заказчиком:

Вид затрат	Часто у интегратора	Часто у заказчика
CapEx	Лицензии, стенды пилота	Серверы, GPU, СХД, сетевое оборудование
OpEx (проект)	Аналитика, разработка, интеграции	PM, приёмка, обучение пользователей
OpEx (эксплуатация)	Поддержка по SLA, доработки	Облачные API, электроэнергия/ ЦОД, ИБ
Передача (KT/IP)	Документация, сессии передачи	Владение репозиторием, развитие

5.5.2 Слой перед LLM и режимы нагрузки

Стоимость вызова LLM — не единственная статья OpEx. До попадания в модель запрос проходит через **слой предварительной обработки**, который добавляет собственные затраты на вычисления и персонал.

Два режима эксплуатации запросов:

- **Быстрый путь:** преимущественно детерминированное выделение структурированных идентификаторов без тяжёлого NLP — маршрутизация по ключевым словам, regex, справочникам.
- **Полный каскад:** дополнительные стадии для имён, адресов, смешанного языка, конкурентных интентов — требует малых моделей (NER, классификаторы) и ресурсов CPU/GPU.

На синтетических корпусах IT-поддержки референс-каскад дал **существенно больший recall**, чем только regex, но **ценой** ложных срабатываний на техническом тексте без доменных фильтров.

Что закладывать в смету:

- Долю обращений, проходящих полный каскад, — умножить на стоимость её стадий.
- Затраты на разметку иподдержку blocklist-политик.
- Отдельную строку на совместимость, профилирование и регрессии при каждом крупном обновлении стека.

Инженерный ориентир по бэкендам (внутренние замеры референс-стека, не договор к поставщику): для одной и той же линейки эмбеддеров на ограниченной VRAM корректность pooling и задержка могут отличаться между высокопроизводительным серверным движком и унифицированным мульти-модельным сервисом; **генеративный ранжировщик** может оставаться на «прямом» пайплайне библиотеки inference, если выбранный сервер не поддерживает требуемый формат скоринга. Это неуникальные числа для КП, а напоминание заложить **отдельные строки** на совместимость, профилирование и регрессии при каждом крупном обновлении.

5.5.3 FinOps и юнит-экономика нагрузки

Для управляемости бюджета ИИ-нагрузок закрепите **юнит-экономику** в духе практик FinOps Foundation:

- **Базовые единицы:** стоимость **на 1 млн токенов**, **на пользователя/месяц**, **на успешный диалог** или **на тикет** — в разрезе среды (облако РФ vs on-prem).
- **Связь с P&L: аллокация** затрат по продуктам/департаментам и **наблюдаемость** (токены, латентность, ошибки, объём и срок хранения трасс).

Внешние ориентиры нагрузки: публичные интервью о клиентском чат-боте — порядки **свыше 3 млн** обращений в месяц, **около 1,7 тыс.** клиентских сценариев (CIO, 2024). Внутренний RAG-помощник — пилот на **3 тыс.** сотрудников, **90%** точность ответов, **девятикратное** сокращение времени поиска («Открытые системы», 2025).

5.5.4 Калькуляция расхода и стоимости токенов по классам задач

Для оценки токенов при планировании бюджета используйте классификаторы по агентам и данным.

Классы агентов по длине системного промпта

Базовый тариф расчёта

Расчёт выполнен при медианном тарифе стандарт-сегмента (~300 ₽/млн токенов, вход=выход, типично для YandexGPT Lite/GLM-5/MiniMax-M2.7). Для точного подсчёта используйте токенизатор конкретной модели и актуальный прайс провайдера.

Класс	Слов	Ток. RU	Ток. EN	Ток. ср.	Ориентир, руб.
Простые чат-боты или классификация	200	400	134	266	0,08
Сложные корпоративные агенты	2 000	4 000	1 340	2 660	0,80
Специализированные агенты (юриспруденция, фарма)	5 000	10 000	3 350	6 650	2,00
Среднее	2 400	4 800	1 608	3 192	**0,96**

Классы данных по длине пользовательского текста

Класс	Слов	Ток. RU	Ток. EN	Ток. ср.	Ориентир, руб.
Короткие (абзац, ответ на вопрос)	300	600	201	399	0,12
Средние (описания, инструкции)	1 500	3 000	1 005	1 995	0,60
Длинные (статьи, обзоры, НПА)	6 000	12 000	4 020	7 980	2,39
Среднее	2 600	5 200	1 742	3 458	**1,04**

5.5.5 Ценовые сегменты внедрения ИИ-агентов

Сегмент	Стоимость внедрения	Ежемесячный OpEx	Описание
Базовый	50 000 – 200 000 руб.	5 000 – 15 000 руб.	Чат-бот (Telegram/Сайт), ответы по базе знаний (RAG).
Средний	300 000 – 1 500 000 руб.	30 000 – 100 000 руб.	Мультиканальность, интеграция с CRM, выполнение действий.
Продвинутый	от 5 000 000 руб.	от 200 000 руб.	Полномасштабное Enterprise-решение, сложные AI-агенты.

5.5.6 Целевые показатели эффективности (KPI)

- **Себестоимость транзакции (Unit Cost):**
  - Оптимизированное состояние: **< 1 руб.** за инцидент.
  - Потолок затрат «тяжёлого» цикла: **~2,5–9 руб.**
- **Инвестиционный порог (Break-even):** при **устойчивой утилизации >60–70%** на горизонте **нескольких лет** закупка/colocation выигрывает у облачной аренды; при переменных нагрузках — облачный API РФ.

5.5.7 Примерные расчёты расхода токенов

Ниже — **внутренняя оценка и примерные расчёты**: из имеющихся данных выведены средние длины и ориентиры по токенам и рублям. Исходный корпус: средние длины по **более чем 12 000** заявкам портала поддержки **Comindware**.

Перевод **слов в токены** выполнен по двум эвристикам («как для русского» и «как для английского» текста) и по среднему между ними; для точного подсчёта используйте токенизатор конкретной модели.

> ⚠ **Актуальность данных**
>
> Версии моделей и тарифы обновляются ежемесячно. Настоящие расчёты — снимок на **март 2026 г.** Перед финальным КП сверьте с провайдерами.

Расчёт токенов на слово:

Язык	Токенов на слово
Русский	2,0
Английский	0,67
Среднее	1,33

Методика расчёта:

Компонент	Формула
Слово	~1,5 токена
Извлечённые данные (RAG)	3–10 статей × 500–2500 слов × 1,5 токена/слово
Вход	RAG + системный промпт + текст заявки
Выход	текст ответа + рассуждения
Всего	вход + выход
Ориентир (₽)	`Всего × 300 / 1 000 000`

Базовая стоимость (без системного промпта, контекста и рассуждений):

Категория	Слов	Ток. RU	Ток. EN	Ток. ср.	Ориентир, руб.
Текст заявки	810	1 620	543	1 077	0,32
Текст ответа	2 641	5 282	1 769	3 513	1,05
Вход + выход	3 451	6 902	2 312	4 590	**1,38**

Полная стоимость (с системным промптом, контекстом RAG и рассуждениями):

Сценарий	RAG	Заявка	Ответ	Рассуждение	Всего ток.	Ориентир, руб.
FAQ / навигация	2 250	600	800	200	3 850	**1,16**
Простая справка	6 000	1 000	1 000	300	8 300	**2,49**
Консультация по настройке	11 250	1 400	1 500	500	14 650	**4,40**
Интеграция / процессы	21 000	2 000	2 000	800	25 800	**7,74**
Диагностика ошибки	24 000	2 400	2 500	1 500	30 400	**9,12**
Архитектурный анализ	37 500	3 000	3 500	2 500	46 500	**13,95**
Среднее	16 833	1 733	1 883	967	21 583	**6,48**

5.5.8 Пересчёт под фактические тарифы провайдеров

Таблицы используют **медианный ориентир — ~300 ₽/млн токенов** для сопоставимости. Фактические тарифы:

Сегмент	Диапазон (₽/млн)	Примеры
Эконом	20–65	GigaChat 3.1 Lite, Qwen3-235B
Стандарт	200–500	YandexGPT Lite, GLM-5, MiniMax-M2.7
Премиум	800–1200	YandexGPT Pro 5.1, GigaChat 3.1 Ultra

Формула для КП

Формула: `стоимость из таблицы × (фактический прайс / 300)`

Примеры:

- YandexGPT Pro 5.1 (800 ₽/млн): `2,33 × (800/300) ≈ 6,21 ₽`
- GigaChat 3.1 Lightning (65 ₽/млн): `2,33 × (65/300) ≈ 0,51 ₽`

5.5.9 Учёт токенов рассуждения (reasoning)

Для моделей с рассуждением (GLM-5, MiniMax-M2.7, Kimi-K2.5, Claude 4.6, GPT-5.4, Gemini 3.1 Pro) фактическая стоимость может превышать базовый расчёт из-за внутренних токенов рассуждения.

Уровень задачи	Токены рассуждения	Мультипликатор	Примеры
Без рассуждения	0	1,0×	FAQ, классификация
Лёгкий	0,5–1× от выхода	1,5–2,5×	Типовая диагностика
Средний	2–5× от выхода	4–8×	Технический анализ
Сложный	8–15× от выхода	10–20×	Архитектура ПО

Рекомендация по смете

Закладывать **средний мультипликатор 5×** к базовому расчёту для неизвестного профиля задач, уточнять до 2× или 10× после замеров на пилоте.

5.6 Инфраструктура GPU

5.6.1 Быстрый выбор железа (апрель 2026)

Сценарий	Железо	Модели	CapEx (ориентир)
Разработка	M3 Max 48 ГБ	До 70B (с оптимизациями)	~300 000 ₽
Продакшн малый	RTX 4090 24 ГБ	До 32B	~190 000–255 000 ₽
Продакшн средний	A100 40 ГБ	До 70B	~850 000–1 275 000 ₽
Продакшн крупный	H100 80 ГБ	До 235B	2 125 000–3 400 000 ₽
Альтернатива NVIDIA	AMD MI300X 192 ГБ	До 405B	~850 000–1 275 000 ₽

Провайдер	A100/час	H100/час	Примечания
Selectel	150–500 ₽	900–2 200 ₽	A100 40/80 ГБ, H100
Yandex Cloud	300–500 ₽	—	A100 80 ГБ
Cloud.ru	~300–500 ₽	—	V100, A100

Оптимизация VRAM позволяет запускать модели, превышающие объём памяти — см. *«Кросс-платформенные техники оптимизации памяти»*.

5.6.2 Профиль on-prem-GPU в проектах Comindware

В **реальных развертываниях Comindware** на узлах инференса используются в том числе **2 GPU RTX 4090 с 48 ГБ VRAM**, доступные в продаже и в конфигурациях **аренды** выделенных GPU-серверов (например у 1dedic — аренда серверов с GPU), и ускоритель **NVIDIA RTX PRO 6000 (Blackwell) с 96 ГБ VRAM**.

5.6.3 Цены на GPU-оборудование (покупка и аренда)

Ориентир рынка (март 2026): сводные данные по покупке отдельных карт, готовых серверов и аренде у российских и зарубежных провайдеров. Цены включают НДС и логистику (для РФ).

Модель GPU	VRAM	FP16/BF16	Покупка, руб.	Сервер/WS (1 GPU), руб.	Аренда, руб./час
B200 SXM	192 ГБ	~2250 TFLOPS	~3 500 000–4 500 000	по запросу	—
H200 SXM	141 ГБ	989 TFLOPS	~3 000 000–4 000 000	по запросу	~1000–1500
H100 SXM	80ГБ	989 TFLOPS	~1 290 000 – 2 550 000	~4 000 000 – 4 500 000	~207 – 841
H100 PCIe	80ГБ	756 TFLOPS	~2 427 975	~4 000 000	~200 – 600
A100 80ГБ	80ГБ	312 TFLOPS	~760 000–1 280 000	~3 500 000 – 4 500 000	~150 – 800
RTX PRO 6000	96ГБ	91,1 TFLOPS*	~1 200 000 – 1 800 000	~1 500 000 – 2 000 000	—
RTX 4090	24 ГБ	82,6 TFLOPS	~170 000–220 000	~320 000–512 000	~80–150
RTX 4090 48 ГБ	48 ГБ	82,6 TFLOPS	от ~300 000	~1 000 000–1 300 000	—

RTX PRO 6000 Blackwell

RTX PRO 6000 (96ГБ) показывает производительность выше H100 во многих задачах инференса за счёт большего объёма памяти и новой архитектуры. Данные по аренде — на основе тарифов Hostkey, Cloud.ru, Yandex Cloud и Selectel.

5.6.4 Требования к VRAM при инференсе LLM

Общий VRAM = Веса модели + KV-кэш + Активации + Перерасход памяти фреймворка

```
VRAM ≈ (Параметры × Байт/Вес) / TP + KV-кэш + Перерасход памяти
```

TP — количество GPU для **тензорного параллелизма** (см. *Глоссарий*); **TP = 1** соответствует отсутствию параллелизма.

Расход VRAM на 1B параметров

Точность	VRAM на 1B параметров
FP16	~2 ГБ
BF16	~2 ГБ
INT8	~1 ГБ
INT4	~0,5 ГБ

Для грубой прикидки до замеров на целевом стеке полезен внешний калькулятор *«VRAM calculator — apxml.com»*; он **не заменяет** учёт **KV-кэша**, **батча** (размера пакета запросов) и перерасхода памяти фреймворка на **инференс на базе vLLM/MOSEC** у заказчика.

Ориентиры для глубокого аппаратного сайзинга: опирайтесь на **официальные** и воспроизводимые источники — отраслевые бенчмарки **MLCommons**, документацию и release notes **vLLM**, а не только маркетинговые цифры. Это помогает отделить оценку **железа**, **KV-кэша**, **батча** и **перерасхода памяти фреймворка**.

Юридические и лицензионные нюансы по поставке GPU — см. *Приложение А: аренда GPU и лицензирование NVIDIA*.

Сайзинг моделей по VRAM

Альтернатива облачным подпискам — локальные модели на выделенных GPU-серверах.

Диапазон **100 000 – 250 000 руб./мес** расходов на токены для команды из 5–10 инженеров в **2,5–26 раз дешевле**, чем аренда GPU-серверов (A100/H100) с инфраструктурой. Однако локальный инференс обеспечивает полный контроль над данными и технологический суверенитет.

Квантование, MoE и плотные модели

Квантование Q4 снижает расход VRAM примерно в 4 раза с минимальной потерей качества. Это не единственный вариант квантования, его следует побирать экспериментально под ваши задачи.

Модели MoE (Mixture-of-Experts) имеют высокое общее число параметров, но используют лишь небольшую их часть (sparse activation), что даёт высокую скорость инференса при относительно скромном потреблении памяти по сравнению с плотными (dense) моделями аналогичного качества.

Таблица необходимого объёма VRAM для моделей отсортирована по росту потребления VRAM в Q4 (экономный формат).

Модель	Параметры	FP16	Q4	Контекст	Мин. GPU
Phi-4-mini-instruct	4B	8 ГБ	3 ГБ	8K	RTX 3060
GigaChat 3.1 Lightning	10B / 1,8B active	20 ГБ	4 ГБ	128K+	RTX 3060
Qwen3-8B	8B	16 ГБ	5 ГБ	128K	RTX 3060
Gemma 4 9B	9B	18 ГБ	5 ГБ	128K	RTX 4090
Llama 4 Scout	17B	34 ГБ	9 ГБ	128K	RTX 4090
DeepSeek-R1-Distill-14B	14B	28 ГБ	9 ГБ	64K	RTX 4090
Qwen3.5-14B	14B	28 ГБ	10 ГБ	128K	RTX 4090
Mistral Small 22B	22B	44 ГБ	12 ГБ	32K	RTX 4090
Gemma 4 27B	27B	54 ГБ	14 ГБ	128K	RTX 4090 (24 ГБ) / A100
GPT-OSS 20B	20B	40 ГБ	10 ГБ	128K	RTX 4090
Qwen3.5-32B	32B	64 ГБ	19 ГБ	128K	A100-80G
DeepSeek-R1-Distill-32B	32B	64 ГБ	20 ГБ	64K	A100-80G
Mistral 8x7B MoE	8×7B / 12B active	84 ГБ	26 ГБ	32K	2×A100
GPT-OSS 120B	120B	160 ГБ	40 ГБ	128K	2×H100 (96 ГБ+)
Qwen3.5-235B-A22B MoE	235B / 22B active	140 ГБ	35 ГБ	128K	2×A100
Llama 3.1 70B	70B	140 ГБ	35 ГБ	128K	2×A100
Llama 4 Maverick	128B / 17B active	256 ГБ	38 ГБ	128K	2×H100
DeepSeek V3	671B / 37B active	400 ГБ	80 ГБ	128K	2×H100
GigaChat 3.1 Ultra	702B / 36B active	700 ГБ	80 ГБ	128K+	3×H100

5.6.5 Пропускная способность инференса

Теоретический максимум

```
Макс. ток/сек ≈ Пропускная способность памяти (ГБ/с) / Размер модели (ГБ)
```

GPU	Пропускная способность	7B (Q4)	70B (Q4)
RTX 4090	1 008 ГБ/с	~288 ток/с	~29 ток/с
A100-80ГБ	2 039 ГБ/с	~583 ток/с	~58 ток/с
H100-80ГБ	3 352 ГБ/с	~958 ток/с	~96 ток/с

Стратегия для агентов для программирования

«Мозг + периферия (edge)» — тяжёлая модель в облаке или на суверенном сервере для архитектурного ревью и сложных задач + микро-модель локально для мгновенного автокомплита.

5.6.6 Корректировка TCO для российского рынка

Импортные пошлины и санкции:

Компонент	База, руб. (глоб. прайс)	Наценка в РФ
H100 80ГБ	~2 500 000	+40–60%
A100 80ГБ	~1 000 000	+40–60%
RTX 4090 48 ГБ	~300 000	+50–80%
RTX 4090 24 ГБ	~200 000	+50–80%
Электричество	~8,5 руб./кВт·ч (зарубежные ЦОД)	5–7 ₽/кВт·ч (РФ)

Факторы наценки:

- Параллельный импорт GPU: +30–50% к стоимости
- Логистика: +10–20% к стоимости
- Отсутствие официальной поддержки NVIDIA: гарантия через сторонних дилеров

5.6.7 Перерасход памяти фреймворков инференса

Перерасход памяти вспомогательных моделей

Примечание: данный раздел касается только вспомогательных моделей (эмбеддеры, ранжировщики, защитники). Основные LLM обычно развёртываются через vLLM, Llama.cpp или SGLang.

vLLM имеет более высокий перерасход памяти по сравнению с MOSEC из-за KV-кэша и непрерывной пакетной обработки:

Тип модели	Модель	Расход VRAM	vs. MOSEC
Эмбеддер			
	`ai-forever/FRIDA`	3,6 ГБ	~2,1 ГБ
	`Qwen3-Embedding-0,6B`	1,9 ГБ	~1,1 ГБ
	`Qwen3-Embedding-4B`	8,9 ГБ	~3,0 ГБ
Ранжировщик			
	`DiTy/cross-encoder`	2,3 ГБ	~1,4 ГБ
	`Qwen3-Reranker-0,6B`	1,5 ГБ	~0,9 ГБ
	`Qwen3-Reranker-4B`	8,1 ГБ	~2,8 ГБ
Защитник			
	`Qwen3Guard-Gen-0,6B`	1,8 ГБ	~1,0 ГБ
	`Qwen3Guard-Gen-4B`	8,8 ГБ	~3,0 ГБ
Комбинация (0,6B)	Embed+Rerank+Guard	~10–15 ГБ	~5 ГБ

Комбинации вспомогательных моделей:

Комбинация	vLLM	MOSEC	Свободно из 24 ГБ (vLLM)	Статус
Embed 0,6B + Rerank 0,6B	~3,5 ГБ	~2,0 ГБ	~20 ГБ	✅ Безопасно
Embed 4B + Rerank 0,6B	~10,5 ГБ	~6,5 ГБ	~13 ГБ	✅ Безопасно
FRIDA + DiTy + Guard 0,6B	~7,7 ГБ	~4,5 ГБ	~16 ГБ	✅ Безопасно
Embed 4B + Rerank 4B	~17 ГБ	~11 ГБ	~7 ГБ	⚠ Тесно
Любая модель 8B	~16–18 ГБ	~10–12 ГБ	<8 ГБ	⚠ Риск OOM

Три типа опор для цифр

- **Замер на стенде заказчика** — при жёстких SLO; - **Профиль MOSEC** — типовой эталон вспомогательных моделей **Comindware**, не индивидуальный сайзинг; - **Публичные профессиональные публикации и энтузиастские обзоры** — только порядок величин для разговора с заказчиком, **не** замер **Comindware** и **не** строка КП **без** отдельного прогона на **его** стеке.

5.6.8 Минимальные системные требования

Компонент	Минимум	Рекомендуется	Высокая производительность
GPU	RTX 3060 (12 ГБ) и др. потребительские	RTX 4090 (24/48 ГБ)	RTX PRO 6000 Blackwell (96 ГБ), A100 (80 ГБ)
ОЗУ	16 ГБ	32 ГБ	64 ГБ+
Хранилище	50 ГБ SSD	200 ГБ NVMe	1 ТБ NVMe
CPU	4 ядра	8 ядер	16+ ядер
Сеть	1 Гбит/с	10 Гбит/с	25 Гбит/с+

5.6.9 Анализ чувствительности по нагрузке

Рост длины контекста в 2 раза увеличивает OpEx в облаке в 1,8 раза, но почти не влияет on-prem (до предела VRAM).

Квантование (Q4) снижает требования к VRAM в 4 раза при потере точности < 3%.

Параметр	Small (консервативный)	Medium (базовый)	Enterprise (агрессивный)
Нагрузка (DAU)	10–50 пользователей	50–500 пользователей	500+ пользователей
Запросов/день	~200	~2 500	~10 000+
Средний контекст	4K токенов	16K токенов	32K–128K токенов
Норматив задержки	< 5 с	< 2 с	< 1 с (реальное время)
Рекомендуемое железо	1× RTX 4090 / аналог	2×–4× RTX 4090 или A100	GPU-кластер (H100/ B200, RTX PRO 6000 Blackwell)

Классы решений: edge-кейсы (ноутбук, одноплатник), выбор CLI или толстого протокола инструментов, квантизация и «раздувание» модели относительно VRAM задают **диапазон** CapEx/OpEx, но не заменяют расчёт под профиль **корпоративный RAG-контур** / **агентный слой Comindware Platform** и выбранный инференс-слей.

5.6.10 Рекомендуемые конфигурации для России

Малый бизнес (1–3 пользователя)

- 2×RTX 4090 локально (24/48 ГБ) или RTX PRO 6000 Blackwell 96 ГБ на более требовательном контуре
- Или GigaChat 3.1 Lite API (Cloud.ru) — 65 ₽/млн токенов

Средний бизнес (5–10 пользователей)

- 2× RTX 4090 (в т.ч. 1×48 ГБ) или A100 40ГБ — локально
- Или гибрид: локальный RAG + облачный LLM

Enterprise (50+ пользователей)

- A100/H100 сервер или МГПУ-кластер
- Sovereign AI: локальный инференс + российские модели

5.7 Облачные провайдеры РФ

5.7.1 Облачное развертывание в России

⚠ Актуальность

Таблицы дают справочный ориентир для сравнения сценариев; точные цены для КП сверяйте по актуальному прайс-листу провайдера.

5.7.2 Cloud.ru — Evolution Compute GPU (2026)

Конфигурация	GPU	VRAM	Стоимость/час, ₽ (с НДС)	Стоимость/мес (730ч), ₽
4×V100	V100	128 ГБ	~988	~721 000
5×A100	A100	320 ГБ	~1 586	~1 158 000
5×H100	H100	400 ГБ	~2 745	~2 004 000
5×H100 NVLink	H100 NVLink	400 ГБ	~4 270	~3 117 000
7×H100 NVLink	H100 NVLink	560 ГБ	~5 978	~4 364 000

Источник: Тарифы Cloud.ru «Evolution Compute GPU», Приложение №7G.EVO.1.

🔥 Рекомендация по Cloud.ru

Для старта инференса LLM выбирайте конфигурацию с **4×V100 (~721 000 ₽/мес)** или **5×A100 (~1 158 000 ₽/мес)**. H100 целесообразен для pre-training 70B+ моделей или кластерных задач.

5.7.3 Yandex Cloud — GPU-инстансы (2026)

Конфигурация	GPU	VRAM	Ориентир ₽/час	Ориентир ₽/мес
T4	T4	16 ГБ	~60–120	~44 000–88 000
V100 (1×)	V100	32 ГБ	~200–350	~146 000–256 000
V100 (4×)	V100	128 ГБ	~700–1 200	~511 000–876 000
A100 (1×)	A100	80 ГБ	~300–500	~219 000–365 000
A100 (8×)	A100	640 ГБ	~2 000–3 500	~1 460 000–2 555 000

Источник: оценка на основе публичных конфигураций Yandex Cloud GPU и рыночных данных. Точные тарифы требуют сверки с прайс-листом Yandex Cloud.

Особенность Yandex DataSphere

Для ML-разработки используйте **Yandex DataSphere** — оплата только за время расчётов (обучение, инференс), без оплаты простоя виртуальной машины. Экономия до 40–60% относительно постоянно включённых GPU-инстансов.

5.7.4 Selectel — Cloud GPU (2026)

Конфигурация	GPU	VRAM	Ориентир ₽/час	Ориентир ₽/мес
A100 40 ГБ	A100	40 ГБ	~150–300	~110 000–219 000
A100 80 ГБ	A100	80 ГБ	~250–500	~183 000–365 000
H100 80 ГБ	H100	80 ГБ	~900–2 200	~657 000–1 606 000
RTX 4090	RTX 4090	24 ГБ	~80–150	~58 000–110 000

Источник: Selectel Cloud GPU, публичные конфигурации. Скидки до 44% при долгосрочной аренде.

5.7.5 Справочно: зарубежные облака (AWS/GCP/Azure)

Конфигурация	GPU	VRAM	Стоимость/час ($)	Эквивалент ₽/час*	Стоимость/мес (730ч), ₽
g4dn.xlarge	T4	16 ГБ	$0,526	~44,71	~32 640
g5.xlarge	A10G	24 ГБ	$1,006	~85,51	~62 390
p3.2xlarge	V100	16 ГБ	$3,060	~260,10	~189 890
p4d.24xlarge	A100 (8x)	320 ГБ	$32,773	~2 785,45	~2 033 370

Сравнение с зарубежным рынком

Российские тарифы на GPU облако в 2026 году **выше** американских/европейских на 30–100% для аналогичных конфигураций (при пересчёте по курсу). Это связано с ограниченной доступностью GPU H100/H200, логистикой и требованиями локализации.

5.7.6 Зарубежные API (разработка и песочницы)

Область применения: оценка ниже относится к **зарубежным** управляемым API (агрегаторы вроде **OpenRouter**, **Google Gemini** и аналоги), удобным для **разработки, экспериментов и ассистентов в IDE**. Для **промышленного TCO** решений у заказчиков в РФ с персональными данными и типовыми требованиями локализации **базовый** контур экономики — тарифы **Cloud.ru / Yandex Cloud / SberCloud / MWS GPT** (токены), **Selectel Foundation Models Catalog** (инфраструктура) и (или) **on-prem** из таблиц выше по документу.

- **Gemini Pro (иллюстративно):** ~0,085 руб. за 1к токенов (вход), 0,255 руб. за 1к токенов (выход).
- **OpenRouter (иллюстративно):** переменная, ~8,5–85 руб. за 1M токенов в зависимости от модели — сверять с каталогом моделей и цен.
- **Оценка диапазона:** 4 250 – 42 500 руб./мес при умеренном использовании в сценарии разработки.

5.8 Альтернативный инференс: edge-устройства, потребительское железо

Альтернативы облачному GPU: локальный инференс на потребительском железе и edge-устройства для сценариев автономности, IoT, промышленной автоматизации.

5.8.1 Кейс: Qwen3.5-397B на M3 Max 48 ГБ

Крупные модели на потребительском железе: можно запускать посредством оптимизации памяти и квантования. Это даёт возможность построения суверенных агентов на стеке **Comindware**.

Источник: LLM под капотом (Dan Woods extract 2026-03-18, MLX/"LLM in a Flash").

Метод: Claude Code / OpenCode + auto-research + оптимизация памяти (квантованный MoE, эквивалент ≈72B, работает на 48 ГБ).

Этап	Время	Результат
Базовая реализация	5 часов	1 ток/с
Оптимизации	+3 часа	4,74 ток/с, 5,9 ГБ RAM

Бизнес-применение для рынка РФ:

- **Независимость от зарубежных облаков:** локальный инференс на Apple Silicon (Mac Studio, Mac Pro) позволяет использовать современные LLM без рисков санкций и блокировок.
- **Корпоративные R&D-кластеры:** кластеризация Mac Studio (M3 Ultra до 256 ГБ RAM) для инференса 70B+ моделей — альтернатива недоступным или дорогим NVIDIA-кластерам. См. Apple specs.
- **Конфиденциальность данных:** данные не покидают периметр — критично для разработки под госконтракты и работы с персональными данными
- **Экономика:** при умеренной нагрузке CapEx на Mac-кластер окупается за 12–18 месяцев относительно облачных GPU РФ.

Юридические и лицензионные нюансы по поставке GPU — см. *«Приложение А: аренда GPU и лицензирование NVIDIA»*.

5.8.2 Edge-агенты на минимальном железе

Edge-AI и тонкие клиенты: возможно запускать агентов на минимальном железе с подключением к облачной LLM через прокси.

Источник: Валерий Ковальский. Red Mad Robot

Характеристики Raspberry Pi + PicoClaw:

- Размер: 5×7 см
- Питание: 5 В
- Модель: GPT-5.4 (через прокси)

Функционал:

- Треды и стриминг
- LangFuse для трейсов
- Google Workspace CLI интеграция
- Самомодификация с перезапуском

Архитектура для РФ: при наличии связи — API российских провайдеров (GigaChat 3.1 Lightning через Cloud.ru — MoE с **1,8B** активных параметров, YandexGPT Lite); при автономном режиме — локальные микро-модели вроде **Phi-4-mini-instruct**, **Qwen-1,5B** (3–4 ГБ VRAM) с синхронизацией при восстановлении канала; тарифы — см. *«Тарифы и провайдеры РФ»*.

Бизнес-применение для рынка РФ:

- **Промышленный IoT:** edge-агенты на производстве для локальной обработки данных с сенсоров — низкая задержка, автономность при обрыве связи.
- **Умные датчики и контроллеры:** интеллектуальная предобработка данных на PLC и промышленных контроллерах перед отправкой в центральную систему.
- **Полевые устройства:** автономные агенты для удалённых объектов (нефтегаз, энергетика, транспорт) — работа при отсутствии интернета с синхронизацией при подключении.
- **Стоимость инференса:** ориентир от ~12 ₽/млн токенов (GigaChat 3.1 Lightning) — экономически эффективно для распределённых edge-задач.

5.9 TCO и сценарии развёртывания

5.9.1 Облачный хостинг (Россия)

На основе тарифов Cloud.ru, Yandex Cloud, Selectel (январь–март 2026):

- **Мелкомасштабный:** ~110 000 – 220 000 ₽/мес (Selectel A100 40ГБ / Yandex Cloud A100 1× / Cloud.ru 4×V100 для небольших нагрузок)
- **Среднемасштабный:** ~510 000 – 880 000 ₽/мес (Yandex Cloud 4×V100 / Cloud.ru 4×V100 / Selectel H100)
- **Крупномасштабный:** от 1 160 000 ₽/мес (Cloud.ru 5×A100 и выше, Yandex Cloud 8×A100, кластеры H100)

5.9.2 TCO GPU: облако РФ против закупки

Ниже — **воспроизводимая иллюстрация** для переговоров о модели владения (облако dedicated GPU vs закупка ускорителей): все суммы в **руб.**. **Не** смешивайте в одной строке сметы: облачный ряд — **операционная аренда** по публичным тарифам; on-prem ряд — **CapEx закупки GPU** без полного учёта ЦОД, сети, персонала и налоговой амортизации (их заказчик добавляет в своей модели).

Допущения сценария «непрерывная нагрузка»: 730 ч/мес на весь период (**8 760 ч/год** на GPU) — типовой множитель из прайса Cloud.ru; горизонт **3 года** → **36** таких месяцев. Для спайковой нагрузки умножьте часы фактической занятости.

Шаг 1 — облако (8× H100, РФ): в открытом прайсе Cloud.ru нет строки ровно **8×**; ближайшая опора — **7×H100 NVLink** (**~5 978 ₽/ч** за конфигурацию). Эквивалент **8 GPU** — линейное масштабирование ставки: **(8/7) × 5 978 = 6 832 ₽/ч**. Тогда:

- **3 года:** 6 832 × 730 × 36 = **179 544 960** руб. (округлённо **~179,5 млн**).

Альтернативная вилка по публичным ориентирам «за 1× H100» (Selectel): **~900–2 200 ₽/ч** на GPU → за 3 года при **730 ч/мес**:

- **низ:** 900 × 8 × 730 × 36 ≈ **189 216 000** руб. (**~189 млн**);
- **верх:** 2 200 × 8 × 730 × 36 ≈ **462 528 000** руб. (**~463 млн**).

Шаг 2 — закупка GPU (только карты, 8× H100 80ГБ): диапазон **~1 290 000 – 2 550 000** руб. за карту → **~10 320 000 – 20 400 000** руб. **единовременного CapEx** на восемь ускорителей (без серверного шасси и прочего).

Шаг 3 — сопоставление: приведённый CapEx **~10–20 млн** относится к **закупке**; ряд **~179–463 млн** — к **трёхлетней** аренде при **полной** занятости по часам выше. Отсюда **не** следует коэффициент «экономии 70%» без профиля часов: при снижении фактических часов аренды облако выигрывает; при **устойчивой высокой**

утилизации выигрывает закупка. Полный on-prem TCO заказчик собирает из CapEx узла, электроэнергии, обслуживания и риска устаревания.

Порог утилизации (управленческий ориентир, не норма РФ): при **низкой** утилизации выгоднее облако; **устойчивая** загрузка **выше ~60–70 %** и горизонт **нескольких лет** сдвигают баланс в сторону собственной инфраструктуры или colocation; параметрически порог **3-летнего** break-even по GPU встречается в коридоре **примерно 55–75 %** в зависимости от тарифа, электроэнергии и графика амортизации (Ivchenko, 2026). Для резидентного контура РФ обязательно пересчитать **практическим** прайсом выбранного провайдера и курсом/договором на дату сметы.

Показатель	Значение (ориентир)	Основание
Аренда 8× H100 (3 года, 730 ч/мес, РФ)	**~179,5 млн** руб. (масштабирование Cloud.ru 7× NVLink)	Cloud.ru — Evolution Compute GPU + расчёт
Та же логика, вилка по Selectel/GPU	**~189–463 млн** руб.	900–2 200 ₽/ч ×8×730×36
Закупка 8× H100 (только карты)	**~10,3–20,4 млн** руб. CapEx	Цены на GPU
1 GPU-экв., облако (доля 5×H100 Cloud.ru)	**~4,81 млн** руб./год при 8 760 ч	2 745/5×8 760

5.9.3 Сравнение TCO за 3 года

Развертывание	Начальные CapEx	Годовой OpEx	TCO за 3 года	Пользователи
Локальное (Мелкое)	212 500 ₽	170 000 ₽	722 500 ₽	1–3
Облачное (Мелкое)	0 ₽	1 020 000 ₽	3 060 000 ₽	1–3
Локальное (Среднее)	850 000 ₽	425 000 ₽	2 125 000 ₽	5–10
Облачное (Среднее)	0 ₽	2 550 000 ₽	7 650 000 ₽	5–10
Локальное (Крупное)	8 500 000 ₽+	1 700 000 ₽	13 600 000 ₽	50+
Облачное (Крупное)	0 ₽	10 200 000 ₽	30 600 000 ₽	50+

Ключевой вывод

Локальное развертывание более экономически эффективно для устойчивых рабочих нагрузок (>1 год); облако лучше подходит для переменных нагрузок или быстрого масштабирования.

5.9.4 Повторяющиеся затраты

Электричество (Локальное)

- **RTX 4090, рабочая станция:** ~400 Вт под нагрузкой → ~1 500 – 2 100 руб./мес (24/7, 730 ч/мес)
- **A100, сервер:** ~2000 Вт → ~7 300–10 200 руб./мес
- **РФ:** 5–7 ₽/кВт·ч (коммерческие тарифы для бизнес-потребителей на апрель 2026; в зарубежных ЦОД — ~8,5 руб./кВт·ч)

Поддержка и обслуживание

Деятельность	Частота	Трудозатраты	Влияние на стоимость
Обновление моделей	Ежемесячно	2–4 часа	Низкое
Поддержка Qdrant, Chroma DB, PostgreSQL+pgvector	Ежеквартально	4 часа	Низкое
Мониторинг системы	Непрерывно	1 час/неделя	Среднее
Резервное копирование	Еженедельно	2 часа	Низкое
Установка обновлений безопасности	Ежемесячно	4 часа	Среднее

Оценочная годовая стоимость обслуживания: 425 000 – 1 275 000 руб./год

5.9.5 Примеры расчёта локального сайзинга

Малый бизнес / департамент

- **Видеокарта:** RTX 4090 (24/48 ГБ)
- **Пользователи:** 1–3 одновременных
- **CapEx:** ~400 000–600 000 руб.
- **OpEx:** 170 000 руб./год (обслуживание)
- **TCO (3 года):** ~900 000–1 100 000 руб.

Среднее предприятие

- **Видеокарта:** RTX 6000 Ada (48 ГБ) / RTX PRO 6000 Blackwell (48–96 ГБ) или 2× RTX 4090 (2×24 ГБ)
- **Пользователи:** 5–10 одновременных
- **CapEx:** ~1 200 000–2 000 000 руб. (локально)
- **OpEx:** 425 000 руб./год (локально)
- **TCO (3 года):** ~2 500 000–3 300 000 руб. (локально)

Крупное предприятие

- **Видеокарта:** NVIDIA A100 (40/80 ГБ) или H100 (80 ГБ) — 4–8 GPU
- **Пользователи:** 50+ одновременных
- **CapEx:** от ~12 000 000 руб. (локально)
- **OpEx:** 1 700 000 руб./год (локально)
- **TCO (3 года):** от ~17 000 000 руб. (локально)

5.10 Риски и оптимизация

5.10.1 Опасность устаревания оборудования

- H100 (2022) → ГБ200 (2025): базовое устаревание
- V100 → A100 → H100: **40–60% потери стоимости** за 18–24 месяца после выхода нового поколения
- Полезный срок жизни: **3–4 года** (vs 5–7 лет для традиционных серверов)

5.10.2 OpEx безопасности GenAI

В TCO промышленного ассистента закладывают не только железо и токены, но и **постоянный контур проверки**: периодический **AI red teaming** (внутренний или внешний), обновление сценариев под OWASP LLM Top 10 2025 и при наличии инструментов — OWASP Agentic Top 10 2026, прогон открытых сканеров вроде Garak на изолированных стендах, обучение разработчиков и линии поддержки **безопасной работе с GenAI** (фишинг против пользователя и «промпт против модели»).

Суммы в смете не фиксируем здесь: они зависят от масштаба, требований регулятора и выбора подрядчика; при планировании закладывают **постоянную занятость специалистов ИБ и машинного обучения и периодические закупки услуг** так же, как на классический пентест.

Контекст рынка (не обязательный поставщик): в деловой прессе отмечают рост внимания к атакам на ИИ-системы и смежным услугам. Крупные вендоры усиливают направления тестирования GenAI. Для заказчика это сигнал **дефицита кадров и инструментов**, а не указание закупить конкретный продукт.

Класс OpEx AI TRiSM: в терминологии Gartner — AI TRiSM к**доверию к ИИ** относят объяснимость, защиту моделей и данных, соблюдение требований и устойчивость операций. В смете промышленного GenAI закладывают работы по **интерпретируемости**, контуры **ModelOps** под защитные механизмы и аудит, периодические проверки на соответствие политикам и сопровождение после смены модели.

5.10.3 Риски внедрения ИИ-проектов

Риск	Влияние	Мера снижения
Раздувание контекста	Высокое (рост операционных затрат)	Жёсткие лимиты токенов; сжатие истории диалога; отсечение неактуальных сообщений.
Путаница в цене за тикет	Среднее (ошибка бюджета, недоверие стейкхолдеров)	Разделять: «цена после оптимизаций» и «верхняя оценка полного цикла»; в КП — только пересчёт по прайсу провайдера.
Организационная зрелость и пилот	Высокое (бюджет без эффекта в P&L)	Закладывать расходы на обучение; явные критерии выхода из пилота; мониторинг затрат (FinOps).
Дефицит GPU (санкции)	Среднее (рост CapEx)	Российские облака; потребительские GPU (RTX 4090, Mac Studio до 256 ГБ); китайские ускорители (Huawei Ascend 910C, Moore Threads MTT S4000, Cambricon MLU590).
Галлюцинации	Высокое (крах агента, $67 млрд потерь в 2024, 40% проектов отмен к 2027)	RAG с цитированием; отдельная модель-контролёр; многоуровневая верификация (96% снижение); сквозные тесты.
Регрессии стека инференса	Среднее (снижение качества)	Закреплять версии vLLM/SGLang; регрессионные тесты перед каждым обновлением; не обновлять прод без пруф-оф-концепт.
Мульти-бэкенд (vLLM, SGLang, Triton)	Среднее (рост сопровождения, несовместимость форматов)	Единая матрица «модель → сервер → тест»; проверять пул-реквесты на совместимость; время специалистов на каждый стек.
Композитный инцидент (атака через RAG + шлюз + политики)	Высокое (утечка данных, простой, репутация)	Разделять хранилища политик, секретов и данных; ASVS/WSTG для шлюза; сценарии по OWASP LLM Top 10 2025 и Agentic 2026.
Теневые ИИ-инструменты	Среднее (обход ИБ, вредонос под видом «ИИ-клиента»)	Политика допустимых моделей; обучение пользователей; мониторинг endpoint'ов; внешний бенчмарк угроз (Kaspersky, 2025).
Самопроверка моделью без инструментов	Среднее–высокое (дефекты в проде)	Жёсткие пороги по критериям; эталонные примеры в промпте; тесты через инструменты и сквозные сценарии, а не только моделью.

Риск	Влияние	Мера снижения
Избыточная оркестрация после обновления модели	Среднее (рост токенов без пользы)	После обновления пересматривать шаги агента; упрощать цепочки, где новая модель справляется сама (Anthropic — Harness design).
Рост затрат на хранение артефактов проверки	Среднее (системы хранения, требования 152-ФЗ)	Политика минимизации; ретенция; контур хранения согласованно с требованиями по персональным данным; см. *«Персональные данные и телеметрия»*.

5.10.4 Оптимизация затрат на инференс

Связка с моделью затрат: следующие факторы напрямую влияют на расходы на **инференс** и **эксплуатацию** и снижают требования к GPU при тех же SLO.

Оптимизация	Эффект	Применимость
DTR-фильтрация	-50% compute	Все LLM
SMTL-параллелизм	-70% шагов	Агенты
Memex-память	-50% токенов	Длинные задачи
KARL-стиль агентного поиска (RL)	≈33% дешевле /≈47% быстрее vs указанные модели на корп. знаниях	Агентный поиск / закрытые базы; требуются замеры на данных заказчика
SkillNet-навыки	-30% шагов	Повторяющиеся задачи

Вывод: оптимизации могут снизить требования к железу на 30–70% без потери качества.

Ключевые академические публикации по оптимизации инференса

Снижение вычислительных затрат:

- *«Think@n — Deep-Thinking Ratio»*: метрика активации «мышления» в слоях; экономия ~2× компьюта без потери качества.
- *«SMTL — Search More, Think Less»*: параллельное решение подзадач; на 70% меньше шагов инференса.
- *«Moonshot Attention Residuals»*: переиспользование представлений предыдущих слоёв; экономия компьюта ~1,25×.

Снижение расхода токенов:

- *«Accenture Memex(RL)»*: индексированная память вместо раздутого контекста; пиковые токены -50%; успешность в ALFWorld: 24% → 86%.
- *«Databricks KARL»*: корпоративный поиск с RL; 33% дешевле и 47% быстрее на задачах корпоративных знаний.

Цифры из бенчмарков; для практического применения требуются замеры на данных заказчика.

5.10.5 Дополнительные стратегии оптимизации

1. **Выбор моделей:** используйте модели 0,6B для приложений с чувствительностью к стоимости; резервируйте 4B+ для высокоточных задач.
2. **Смешанная точность:** используйте fp16/bf16 для уменьшения VRAM на 50% по сравнению с fp32.
3. **Обслуживание нескольких моделей:** MOSEC позволяет развертывать несколько моделей на одном сервере, снижая потребности в оборудовании.
4. **Прерываемые инстансы:** Yandex Cloud и другие российские провайдеры предлагают прерываемые (preemptible) GPU-инстансы со скидкой до 40–60% для некритичных рабочих нагрузок.
5. **Автоматическое масштабирование:** масштабируйте облачные ресурсы в зависимости от спроса для минимизации простоев.

5.10.6 Актуальные тренды AI/ML

Актуальные тренды AI/ML для стратегического планирования — см. *Приложение D*.

5.11 Заключение

5.11.1 Обоснование рекомендаций

В тексте местами используются выдержки из **публичных** профессиональных публикаций и отраслевых сообществ. Они показывают **рыночную и инженерную повестку** и по возможности снабжены ссылкой на первоисточник.

Как использовать на уровне решений: такие публикации — дополнительный сигнал для уточнения позиции команды, юристов и поставщиков. Перед утверждением бюджета или подписанием контракта перепроверьте дату первоисточника, актуальные прайс-листы и соответствие требованиям комплаенса (152-ФЗ, режим КИИ, реестры ПО и иные применимые нормы).

5.11.2 Экономика документа и комплект для заказчика

Экономический контур проекта охватывает **бюджетирование, сценарный сайзинг и TCO**; **договорной комплект отчуждения** (артефакты, регламенты, обучение, комплаенс) согласуют с отчётом *«Методология внедрения и отчуждения ИИ»* и *Приложением В «Отчуждение ИС и кода (КТ, IP, лицензии, приёмка)»*. Перечень передаваемых артефактов фиксируется в соглашении.

5.11.3 Итог

Comindware предлагает готовую методологию и референс-стек для внедрения корпоративного ИИ в российских условиях. Мы передаём не «чёрный ящик», а воспроизводимую инженерную практику: от расчёта CapEx/OpEx и выбора облако/on-prem до промышленной наблюдаемости и отчуждения.

5.11.4 Для заказчика это означает

- **Предсказуемый бюджет** — сметы на основе актуальных российских тарифов, а не зарубежных ориентиров
- **Суверенный контур** — соответствие 152-ФЗ, КИИ, локализация данных и моделей
- **Передаваемая экспертиза** — код, конфигурации, регламенты эксплуатации и обучение команды

Следующий шаг: пилот по методологии PoC → Пилот → Масштабирование с фиксацией метрик качества и экономики до промышленного запуска.

6. Приложение А. Отчуждение ИС и кода: KT, IP, лицензии, критерии приёмки

6.1 Обзор

Приложение фиксирует состав минимального пакета передачи (код, конфигурации, данные, модели, регламент), лицензионные границы и критерии приёмки для закрепления в договоре.

Обеспечивается не только развертывание целевого контура, но и полная **операционная автономия** заказчика.

TCO и сметные последствия — в *«Сайзинг и экономика»*.

6.2 Практический смысл для сделки и передачи

6.2.1 Для обоснования инвестиций

- **Правовой суверенитет**: фиксация прав собственности на код, модели и конфигурации. Детерминированные границы лицензионной ответственности.
- **Устранение вендор-лока**: минимизация «фактора поставщика» через стандартизированную передачу компетенций (KT).
- **Коммерческая прозрачность**: что включено в поставку и передачу, а что — внутренние рабочие материалы исполнителя.

6.2.2 Для переговоров

- Демонстрируйте переход от «SaaS-зависимости» к **контролируемому активу**. Акцентируйте внимание на устранении рисков «теневого ИИ» за счет внедрения корпоративного контура.
- Для выбора модели взаимодействия (управляемый сервис / BOT / построение и передача) опирайтесь на *«Модели поставки и передачи»*.

Минимальный комплект передачи (что фиксировать в договоре):

См. *«Пакет отчуждения (минимально целостный)»* и *«Критерии приёмки передачи (чек-лист)»*.

Рекомендуемый подход: закрепить артефакты передачи + критерии приёмки + срок интенсивного сопровождения после передачи (hypercare) + владельцев компонентов на стороне заказчика.

6.3 Детальная методология отчуждения

6.3.1 Ориентиры для заказчика: инструменты ускорения разработки (вне поставки **Comindware**)

Инструменты ускорения разработки для команды заказчика:

- **OpenCode** — открытый агент для программирования и исполнения кода; провайдеры и модели задаются конфигурацией. Каталог плагинов и интеграций сообщества: Ecosystem.
- **OpenWork** — десктоп/UI-слой для команд поверх OpenCode (также перечислен в Ecosystem).
- **OpenCode Zen** — опциональный **платный** шлюз с отобранными моделями (beta).

> ⚠ **Для контуров с 152-ФЗ**
>
> Для контуров с **152-ФЗ не** принимайте OpenCode Zen как дефолт без оценки: хостинг и политики обработки данных определяются провайдерами шлюза (юрисдикция США). Бесплатные линейки могут иметь **ограниченный срок** и **особые условия использования данных** — см. официальный текст Zen.

- **OpenRouter** — агрегирующий **API-шлюз** к множеству зарубежных провайдеров; типичное применение — **IDE, агенты для программирования, прототипирование** (в т.ч. совместимо с конфигурацией **агентного слоя Comindware Platform** в upstream).

> ⚠ **Для продакшн-развёртывания в РФ**
>
> Для **продакшна в РФ** с ПДн OpenRouter **не** является базовым вариантом: маршрутизация за рубеж, биллинг и политики журналирования задаются цепочкой провайдеров; без отдельной **юридической и ИБ-оценки** не подменяет API **Cloud.ru / Yandex Cloud / SberCloud** или закрытый контур.

- **Cursor** — коммерческая IDE с подпиской; ориентиры по токенам для сравнения см. в параграфе *«FinOps и юнит-экономика нагрузки»*.

Практика для РФ: снижение зависимости от зарубежного биллинга — через **локальные модели** и **API в РФ**; проверяйте доступность и условия на дату по официальным источникам. Итоговый контур согласовывается с комплаенсом и владельцем данных.

6.3.2 Теневой GenAI в маркетинге и маршрутизация моделей (ориентир опроса CMO, 2025)

Публичные материалы опроса **red_mad_robot × CMO Club Russia** фиксируют **высокую концентрацию** на универсальных зарубежных чат- и визуальных сервисах среди маркетинговых директоров (порядка **91%** для **ChatGPT** и **59%** для **Midjourney**, с широким разрывом до следующих инструментов; сводные доли и контекст — в *«Зрелость российского рынка GenAI»*).

Для комплекта отчуждения и договоров о **ИС** это аргумент за явный **каталог допустимых моделей и маршрутов данных**, учёт **TOS/API** зарубежных SaaS и разделение **промышленного** контура (**корпоративный RAG-контур**, API РФ, on-prem) от **самостоятельного** использования маркетингом глобальных сервисов; управленческий смысл и перекрёстные ссылки — в *«GenAI в маркетинговых командах крупных брендов РФ (опрос CMO, 2025)»*.

Питч «маркетинг / shadow SaaS / суверенитет ИС» (включать при активном движении продаж): высокая концентрация на универсальных зарубежных сервисах усиливает риски **утечки данных** через неучтённые каналы, непрозрачных **обработчиков данных** и смешения **корпоративных активов** с личными учётными записями. В договоре и комплекте отчуждения полезно явно зафиксировать **реестр ИИ-инструментов**, **операторов/обработчиков**, политику журналирования и запрет **теневого** GenAI вне согласованного контура; количественные доли и барьеры опроса — в *«Зрелость российского рынка GenAI»*.

6.3.3 Отчуждение данных

Отчуждение данных — фиксация того, какие знания и цифровые активы переходят к заказчику, какие нужно хранить для воспроизводимости и аудита, а какие можно удалить без потери операционной способности.

- **Поддержка векторного слоя:** штатные утилиты сопровождения позволяют диагностировать, очищать и мигрировать коллекции без потери управляемости.
- **Удаление векторного хранилища:** коллекции можно удалить через прикладной API или клиентские инструменты, но решение об удалении должно быть связано с политикой хранения, требованиями ИБ и возможностью последующего воспроизведения контура.
- **Архивация документов:** исходные документы и правила индексации должны сохраняться как часть передаваемого актива; векторные представления можно пересобрать, если сохраняется источник и политика ingestion.

6.3.4 Отчуждение моделей

Отчуждение моделей — не только передача права использовать модель, но и передача политики версий, лицензий, ограничений применения и ответственности за деградацию качества после смены модели или весов.

- **Обновление конфигурации:** смена идентификаторов моделей выполняется через параметры окружения и конфигурацию моделей; это позволяет заказчику управлять жизненным циклом моделей без скрытой зависимости от интегратора.
- **Перевод модели в эксплуатацию:** порядок перезагрузки, переключения и возврата к предыдущей версии должен быть формализован в эксплуатационном регламенте и согласован с требованиями по SLA.
- **Версионирование:** модели должны отслеживаться по идентификаторам релизов и правилам фиксации версии; откат — часть управляемого процесса, а не ручное исключение.
- **Открытые веса и лицензия:** при self-hosted чекпойнтах (в т.ч. GigaChat-3.1 под MIT — Хабр, Сбер) в комплект передачи входят идентификаторы релиза (HF/GitVerse), текст лицензии, политика фиксации версий и регрессионные проверки качества при смене весов.
- **Кастомные лицензии на публичные веса:** помимо permissive-лицензий хранить пороги по **выходным токенам**, календарные сроки уведомления правообладателя и условия атрибуции; иллюстративный полный текст — Лицензионное соглашение YandexGPT-5-Lite-8B (файл на Hugging Face).
- **Реестр доверенных моделей:** публикация открытых весов **не заменяет** проверку допуска модели для госсектора и КИИ (см. *«Управление рисками и комплаенс»*).

6.3.5 Отчуждение инфраструктуры

Отчуждение инфраструктуры — не техническая операция, а момент перехода ответственности за работоспособность контура, его затраты, риски и постоянные цифровые активы.

- **Переход операционной ответственности:** зафиксируйте в договоре и комплекте передачи дату и логику прекращения эксплуатации интегратором и начала ответственности заказчика — это исключает спор о том, кто отвечает за активный контур, инциденты и расходы после передачи.
- **Воспроизводимость среды:** заказчик должен получить не только право на контур, но и возможность управляемо его запустить, остановить, перенести или вывести из эксплуатации; без этого инфраструктура остаётся зависимой от поставщика даже при формальной передаче прав.

- **Постоянные данные и цифровые следы:** после остановки вычисления прекращаются, но постоянные артефакты — данные, журналы, кэши и иные накопленные материалы — продолжают создавать требования по ИБ, комплаенсу и затратам. По ним заранее фиксируют решение: сохранить, архивировать, мигрировать или удалить.
- **Критерий завершённой передачи:** инфраструктура считается реально отчуждённой только тогда, когда определены владельцы сервисов, место хранения постоянных данных, порядок восстановления и правила очистки; сама по себе остановка эксплуатации ещё не означает завершённую передачу.

6.3.6 Справочно: аренда GPU и лицензирование NVIDIA (GeForce vs datacenter)

При **аренде BM или сервера с GPU** юридический контур дополняет open-source лицензии на веса моделей: для потребительских линеек (GeForce / RTX и аналоги) и для продуктов, классифицируемых как **datacenter**, действуют **разные** рамки лицензионных условий NVIDIA и сопутствующих ограничений на ПО и сценарии использования — **due diligence** по текстам на дату сделки и по профилю нагрузки (коммерческий инференс, колокация, облако).

Каталоги аренды публично смешивают классы железа (иллюстрации: Intelion Cloud, HOSTKEY — GPU dedicated servers).

Наличие SKU в каталоге не заменяет юридическую и ИБ-проверку сценария заказчика.

Почасовые тарифы для таких каналов см. в параграфах *«Тарифы российских облачных провайдеров ИИ»* и *«Цены на GPU-оборудование (покупка и аренда)»*.

6.3.7 Модели поставки и передачи (интеллектуальная собственность (ИС) и передача знаний)

Модель	Суть	Типичный комплект на выходе	Риски для заказчика
Управляемый сервис	Эксплуатация и развитие у интегратора	SLA, отчёты, доступ к API; ограниченный доступ к коду	Зависимость от поставщика, границы ИС по договору
Совместная разработка	Команды заказчика и интегратора в одном контуре	Репозиторий, CI, совместные регламенты	Согласование скорости и приоритетов
Построение — Эксплуатация — Передача (BOT)	Сначала ввод в промышленную эксплуатацию силами интегратора, затем передача заказчику	Эксплуатационный регламент, обучение, интенсивное сопровождение сразу после передачи (hypercare), права на код и конфигурацию по договору	Качество передачи и полнота документации
Построение и передача	Разработка и передача заказчику «под ключ» без длительной эксплуатации у интегратора	Код, тесты, документация, сессии передачи знаний	Нужна внутренняя эксплуатационная готовность

Модель **BOT** и факторы успешной передачи обобщены, в частности, в материалах Luxoft — Build–Operate–Transfer и InOrg — бесшовная передача (seamless handover) в модели BOT.

5-фазная модель BOT для ИИ-проектов:

Фаза	Продолжительность	Ключевые активности
Pre-Build	1–2 мес.	Соглашение о передаче ресурсов, начальная настройка governance
Build	3–6 мес.	Инфраструктура, найм команды, правовое соответствие
Operate	12–24 мес.	Оптимизация производительности, передача знаний, созревание процессов
Transfer Prep	2–4 мес.	Подготовка к чистой передаче, параллельные операции
Transfer	2–6 мес.	Формальная передача прав, IP-трансфер

Рекомендуемый срок для ИИ-интенсивных проектов: **24–42 месяца** (для платформ/CoE) до **30–60 месяцев** (для agentic AI систем).

Ключевые факторы успеха передачи (2025–2026): - Управление удержанием персонала: программа взаимодействия с командой не менее **90 дней до T-Day** - Параллельные структуры отчётности для непрерывности знаний - Оценка готовности клиента к приёмке до начала проекта - Измеримость эффекта и KPI до развёртывания: передача проходит устойчивее, когда стороны заранее договорились, какие метрики подтверждают ценность и readiness, а не спорят о результате постфактум - Компетенции в области ИИ как стратегическая возможность

6.3.8 Готовность к передаче

Для руководства полезно оценивать **передачу** не как единичный акт подписания, а как подтверждённую **готовность к смене владельца** по четырём контурам:

- **Документация:** архитектура, эксплуатационный регламент, критерии приёмки и журнал ограничений.
- **Доступы и инструменты:** владение контурами, правами, CI/CD, наблюдаемостью, лицензиями и секретами.
- **Люди и организационная готовность:** владельцы ролей, обучение, график hypercare, готовность принимать инциденты и изменения.
- **Стабилизация поставки:** план перехода **Day 0 / Day 30 / Day 90**, при котором заказчик подтверждает не только формальное владение, но и устойчивую эксплуатацию без сервисного разрыва.

Такая логика соответствует рыночной практике оценки готовности к передаче в BOT-моделях: смена владельца считается успешной, когда стабильны люди,

процессы, документация, доступы и ритм поставки, а не только подписан акт передачи (InCommon — BOT Transfer Readiness & Handover Mechanics, 2025/2026).

При выборе **управляемой LLM-платформы** у инфраструктурного провайдера типовой комплект для юридической и закупочной сверки включает **лицензионные и иные условия ПО**, описание режимов **SaaS / hybrid / on-prem** и границ ответственности — иллюстрация: специальные условия для ПО «MWS GPT». Публичная декомпозиция **платформы корпоративных агентов** (разметка, RAG, ops, интеграции) задаёт **чеклист владельцев** и артефактов передачи, переносимый на стек **корпоративный RAG-контур** / **агентный слой Comindware Platform** независимо от бренда (MWS AI Agents Platform).

Коротко для руководства

Для сделки и приёмки здесь важны три вещи: минимально целостный пакет отчуждения, подтверждённая готовность заказчика принять контур и формальные критерии приёмки передачи.

Рыночные сравнительные цены, карты вендоров и продуктовый радар держите вне этого приложения: они уже вынесены в отчёт по сайзингу и в Приложение F.

6.3.9 Пакет отчуждения (минимально целостный)

Артефакт	Назначение
Исходный код и манифест зависимостей	Воспроизводимая сборка
Конфигурация без секретов + описание переменных окружения	Развёртывание у заказчика
Эксплуатационный регламент (старт, стоп, резервное копирование, масштабирование)	Снижение зависимости от ключевых сотрудников
Наборы для оценки качества и регрессионных проверок	Контроль деградации после релизов
Политика наблюдаемости (выборка, ретенция, маскирование ПДн) и схема экспорта телеметрии	Согласованность с 152-ФЗ и воспроизводимость разборов инцидентов
Дашборды и правила алертов (латентность, ошибки, токены, защитные механизмы)	Эксплуатация и FinOps в одном контуре метрик
Описание данных и политика индексации RAG	Повторяемость ingestion
Политики ИБ и защитные механизмы (черновик под ЛНА заказчика)	Согласование с комплаенсом
Матрица ролей и эскалаций	Эксплуатация и аудит
Регламент и реестр **Agent Skills** (версии, условия вызова)	Воспроизводимая агентская разработка и сопровождение
Конфигурация **MCP**, **CI** и **CD** для агентов (allowlist, секреты, политика веток)	Контролируемая среда исполнения
Рубрики и промпты для **модели-контролёра**, **эталонные примеры в промпте**	Меньше завышенных вердиктов, если вердикт выставляет **только модель-контролёр** без инструментальных проверок
Шаблоны промптов для структурированных не-кодовых артефактов (пример: **BPMN 2.0 XML** с согласованными `id` семантики и диаграммы)	Воспроизводимая формализация процессов при KT; меньше правок после генерации LLM при явных правилах и проверке в редакторе
Регламент **синхронизации док ↔ код** (периодические прогоны, ответственный)	Борьба с устареванием знаний в репозитории

Практика по **BPMN 2.0**, шаблонам промптов и валидации XML — в параграфе *«Формализация процессов (BPMN 2.0) и генерация с помощью LLM»* отчёта «Методология разработки и внедрения ИИ».

6.3.10 Справочно: агент в PR и артефакты вместо прямой записи в ИС

Для сценариев **анализа и предложения правок** по **pull request** снижает риск для ИС и приёмки, когда среда исполнения выдаёт **наружу артефакты** (**diff**, отчёты тестов, текст ревью), а **прямая запись** в защищаемую ветку или «истинный» репозиторий выполняется только после **человеческого** или **согласованного CI**-решения. Типовая песочница: репозиторий **только чтение**, временная рабочая область, сеть с **deny-by-default** и allowlist на зеркала и артефакты, **краткоживущие** токены с минимальным scope. Вопрос «разрешать ли запись **напрямую** в репозиторий или ограничиться **артефактом** на проверку» имеет смысл явно вынести в решение владельца продукта и ИБ; см. также *«Безопасный MVP контура исполнения за 30 дней, дискуссия по средам и выводы»*, *«Модель риска, паттерны среды и минимальный состав платформы»*.

6.3.11 Уровни обучения при передаче

Четырёхуровневая модель:

Уровень	Аудитория	Фокус	Формат
1	Все сотрудники	Статическое обучение — комплаенс, обязательное завершение	Онлайн-модуль 30–45 мин
2	Опытные сотрудники	Динамическое обучение — роль-специфичное, мобильное микрообучение	Интерактивный воркшоп 2–4 ч
3	Технические команды	Программы развития — навыки, карьерное выравнивание	Полный день + ресурсы
4	Все сотрудники	AI-Native capability — мышление с ИИ, стратегическая интеграция	Непрерывное

Расширенная модель для ответственного ИИ:

Уровень	Аудитория	Содержание
Руководство	C-suite, VPs	Риски ИИ, ответственность за комплаенс
Технические специалисты	Разработчики, Data Scientists	Безопасный жизненный цикл ИИ, тестирование смещений, XAI
Опытные пользователи	Частые пользователи ИИ	Продвинутые возможности, валидация выводов
Рядовые сотрудники	Весь персонал	Обзор governance, допустимое использование

6.3.12 Организационные условия после передачи

Устойчивость эффекта после KT и интенсивного сопровождения после передачи (hypercare) зависит не только от эксплуатационного регламента и контура оценки качества, но и от **оргмеханики** заказчика: **внутренняя мобильность** между функциями для кросс-функциональных сценариев с ИИ; **поддержка экспериментов** (песочницы, лимиты, этика использования); роль **руководителей** в **масштабировании** повторяемых практик и в снятии блокеров по данным и доступам. Эти меры **дополняют** таблицу уровней обучения и типичные модели BOT/Построения и передачи выше.

6.3.13 Критерии приёмки передачи (чек-лист)

- Сборка из переданных артефактов воспроизводится на стенде заказчика без «скрытых» шагов.
- Пройдены согласованные сценарии оценки качества; зафиксированы базовые метрики.
- Эксплуатационный регламент покрывает типовые сбои и контакты эскалации.
- Определены владельцы компонентов на стороне заказчика и дата окончания интенсивного сопровождения после передачи (hypercare).
- По ИС: зафиксированы лицензии, сторонние компоненты и ограничения использования.

6.3.14 Справочно: открытые стандарты OWASP и внешние программы обучения (не входят в поставку по умолчанию)

В комплект **отчуждения знаний** целесообразно включать **ссылочный каркас**: первичные URL OWASP GenAI Security Project (LLM Top 10 2025, Agentic Top 10 2026, AI Testing Guide), при необходимости — WSTG и ASVS 5.0 RU. Русскоязычные дайджесты сообщества (например, Habr — OWASP LLM TOP 10 2025) удобны для онбординга, но **не** заменяют официальные тексты.

Коммерческие **курсы безопасности LLM** у третьих лиц (иллюстративный пример публичной программы — «Large Language Models Security» у «Лаборатории Касперского», расписание и стоимость — только по сайту поставщика на дату закупки) могут дополнять подготовку ИБ и разработки заказчика; это **опция**, а не часть базовой поставки **корпоративный RAG-контур** / **сервер инференса MOSEC/vLLM** / **агентный слой Comindware Platform** без отдельного соглашения.

Публичная программа Школы управления СКОЛКОВО *«Переход в ИИ: трансформация бизнес-процессов»* (модули, сроки и заявленные результаты — по странице программы) иллюстрирует рынок **обучения руководителей** внедрению ИИ

в процессы; **не** входит в поставку референс-стека без отдельного договора. Управленческий контекст — в параграфе *«Стратегия внедрения ИИ и организационная зрелость»* отчёта «Методология разработки и внедрения ИИ».

7. Приложение B. Имеющиеся наработки Comindware

7.1 Обзор

В приложении представлен **состав референс-стека Comindware**: границы модулей, роли в архитектуре и разграничение между **поставляемыми артефактами** и **методологическими рекомендациями**.

Критерии приёмки при передаче — *Приложение A «Отчуждение ИС и кода»*.

Модель внедрения и экономика — *«Методология разработки и внедрения ИИ»* и *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

7.2 Практический смысл для сделки и границ поставки

Для обоснования инвестиций:

- Даёт ответ на вопрос «что у **Comindware** реально есть» и где проходит граница между референс-стеком и методологией.
- **Comindware** поставляет **передаваемые артефакты** (код, конфигурации, регламент, контур оценки качества, обучение) — см. *Приложение A «Отчуждение ИС и кода»*.

Для переговоров:

- Названия модулей в отчёте — **роли компонентов** и иллюстративный референс; коммерческий состав фиксируется договором.
- Состав стека явно ограничен: нет «магии», каждый модуль снабжён документацией и тестами — управляйте ожиданиями на старте.

> ⚠ **Ограничения**
>
> Не используйте документ как:
>
> - Исчерпывающий перечень без сверки с договором (коммерческий состав определяется отдельно).
> - Гарантию совместимости со всеми версиями внешних компонентов без проверки.

7.3 Обзор текущей архитектуры **Comindware**

Comindware располагает модульным контейнеризованным контуром: RAG-движок, серверы инференса (**MOSEC/vLLM**) и **агентный слой Comindware Platform**.

Это приложение описывает состав и границы компонентов.

В этом пакете MOSEC и vLLM обозначают не только базовые open-source проекты, но и производственную обвязку Comindware вокруг них:

- **MOSEC:** единый HTTP-сервис, реестр и подбор моделей, конфигурация и управление вспомогательными сервисами (эмбеддеры, ранжировщики, защитные модели).
- **vLLM:** управляемый контур промышленного LLM-инференса и pooling-моделей — конфигурация сервера, оптимизация параметров под GPU-профили, проверки доступности, вызов инструментов и интеграция с **корпоративным RAG-контуром**.

См. также *«Сайзинг и экономика (CapEx / OpEx / TCO)»* и *«Профиль on-prem-GPU в проектах Comindware»*.

Для корректного сравнения вариантов интерпретируйте эти данные через *Инфраструктуру GPU и ориентиры для сайзинга*.

Технические детали развёртывания конкретных модулей — в публичной документации соответствующих программных компонентов экосистемы.

Архитектурные принципы:

- **Разделение ответственности:** дискретные слои для обработки данных, поиска, инференса и API-доставки.
- **Гибридный поиск:** векторный поиск + поиск по ключевым словам для оптимальной точности.
- **Агентная архитектура:** агенты LangChain для динамического вызова инструментов и структурированного рассуждения.
- **Гибкость инфраструктуры:** поддержка MOSEC (единая HTTP-точка для вспомогательных моделей) и vLLM (выделенные инстансы LLM и pooling-сценариев).
- **Российский суверенитет:** приоритет российских облачных провайдеров и локального инференса под требования 152-ФЗ.
- **Наблюдаемость: Arize Phoenix** + **OpenInference** для трассировки, онлайн- и офлайн-оценок GenAI — *Приложение C «Безопасность, комплаенс, наблюдаемость»*.

- **Недоверенное исполнение:** для сценариев с кодом проектируйте **изоляцию среды** отдельно; рекомендации — *Приложение C «Граница доверия»*.
- **Организационная зрелость:** наличие модулей **не заменяет** оргпроцессы, операционную модель и обучение — *«Стратегия внедрения ИИ»*.

7.3.1 Компоненты экосистемы

Ключевые проекты Comindware:

Компонент	Бизнес-ценность
Корпоративный RAG-контур	Поиск и генерация ответов на любых базах знаний; MCP-эндпоинты; инкрементальная индексация; конвейер оценки качества; переиспользуемые YAML-конфигурации, CLI для администрирования
Агентный слой Comindware Platform	Низкокодовые ИИ-агенты: событие → промпт → LLM-классификация → детерминированное действие; интеграция с GigaChat и OpenAI-совместимыми провайдерами; переиспользуемые YAML-конфигурации, CLI для администрирования
Сервер инференса MOSEC	Унифицированный сервер специализированных моделей; переиспользуемые YAML-конфигурации; CLI для администрирования
Сервер инференса vLLM	Промышленный инференс LLM; переиспользуемые YAML-конфигурации; CLI для администрирования
Сервер инференса Infinity	Сервер эмбеддеров и ранжировщиков; переиспользуемые YAML-конфигурации; CLI для администрирования
Анонимайзер	Извлечение и маскирование метаданных из потоков LLM; переиспользуемые YAML-конфигурации, CLI для администрирования

7.4 Функциональный арсенал агентного контура

Агентная архитектура **Comindware Platform** базируется на взаимодействии корпоративного RAG-контура и интеллектуального ассистента аналитика.

Возможность	Проблема	Решение	Результат
RAG-контур на любых базах знаний	Разрозненные документы, ручная подготовка контекста	Инкрементальная индексация, стабильные идентификаторы, MCP-совместимые эндпоинты	Воспроизводимый поиск по любому корпусу документов
MCP-серверы на базах знаний	Интеграция знаний со специализированными коннекторами	MCP-совместимые эндпоинты для чата и поиска	Любой MCP-совместимый агент получает доступ к знаниям о платформе или корпоративным знаниям
MCP-серверы для Comindware Platform	Прямое взаимодействие с платформой из внешних агентов	Инструменты как MCP-совместимый набор	Внешние агенты управляют сущностями платформы
Скиллы для Comindware Platform	Ручное создание атрибутов, шаблонов, записей	Инструменты атрибутов и приложений	Пакетная настройка сущностей по техзаданию
Интуитивное проектирование	Высокий порог входа при низкоуровневом использовании API	Трансформация запросов на естественном языке в детерминированные операции **Comindware Platform**	Снижение TTM за счет автономного исполнения конфигурационных задач
Пайплайны индексации и глубоких исследований	Обновление знаний требует полной переиндексации; одношаговый поиск не даёт обоснованных ответов	Инкрементальное обновление, семантическая нарезка, расширенное обогащение метаданных; многошаговый поиск с fan-out запросов, декомпозицией, поиском родительских документов, фильтрацией и усилением по метаданным	Изменённые документы переиндексируются отдельно; обоснованные ответы и действия агентов на корпоративных или веб-данных

Возможность	Проблема	Решение	Результат
Конвейеры структурированного рассуждения (SGR)	Свободные промпты дают непредсказуемые результаты	Schema-Guided Reasoning (SGR, структурированное рассуждение по схеме): оценка спама, уверенность намерения, подзапросы, план действий	5–10 % улучшение точности; воспроизводимое рассуждение
Извлечение метаданных из потоков LLM	Потери контекста при отладке и аудите	Паттерн извлечения из результатов вызова инструментов, контекст-трекер с диагностикой	Полная трассировка каждого шага агента
Фронтенд- и бэкенд-паттерны агентов	Каждый проект создаёт UI и API с нуля	Gradio UI + встраиваемый виджет; REST API на базе FastAPI с потоковой выдачей, диагностикой и автоматической документацией OpenAPI/Redoc	Переиспользуемый интерфейс и готовый бэкенд для любого агента
Организация вызова инструментов	Модели вызывают инструменты непредсказуемо	Принудительный выбор инструмента, дедупликация, кэширование	Детерминированное поведение инструментов

7.5 Фреймворки обвязки серверов инференса

Три сервера инференса (MOSEC, vLLM, Infinity) управляются единым фреймворком Comindware с общими принципами:

- YAML-конфигурации для переиспользования и транспортируемости
- Проверенные наборы параметров для стандартных моделей
- CLI для администрирования и диагностики

CMW MOSEC — унифицированный сервер вспомогательных моделей:

Единый HTTP-сервис на одном порту для эмбеддеров, ранжировщиков и защитных моделей. YAML-реестр с проверенными конфигурациями; автоматическая документация OpenAPI и интерактивная площадка Redoc.

CMW vLLM — промышленный LLM-инференс:

Промышленный инференс LLM и pooling-моделей с трёхуровневой конфигурацией (переменные окружения → реестр → параметры модели). Оптимизация памяти через выгрузку KV-кэша; YAML-реестр с проверенными конфигурациями.

CMW Infinity — сервер эмбеддеров и ранжировщиков:

Серверы на базе infinity-emb с автоматическим определением устройства (GPU/CPU). YAML-реестр с проверенными конфигурациями.

7.5.1 Ассистент аналитика Comindware: проверенный агент

Проект: агентный слой Comindware Platform (отдельный компонент экосистемы **Comindware**, см. таблицу выше).

Бизнес-назначение: интеллектуальный ассистент, переводящий запросы на естественном языке в детерминированные вызовы API Comindware Platform.

Целевая аудитория: аналитики (настройка приложений, аудит, реверс-инжиниринг ТЗ) + конечные пользователи (поиск информации, создание записей, отчёты).

Сценарии использования: поиск информации, создание записей, настройка приложений, адаптация приложений, отчёты и статистика, поиск узких мест и ошибок, документирование.

Варианты развёртывания:

1. Замкнутый контур клиента: Platform + ассистент + LLM локально, без интернета
2. Контур Comindware/партнёра: ассистент + LLM у партнёра, Platform у клиента через VPN
3. Облачная LLM: ассистент у партнёра, LLM у провайдера, Platform у клиента
4. MCP-server mode: ассистент как MCP-сервер у клиента/партнёра

Ключевые архитектурные решения: модульность, нейтральность к среде развёртывания, поддержка любых поставщиков ИИ, тотальная трассировка диалогов, открытость (LangChain).

Уникальное преимущество: компоненты ассистента переиспользуются в любых агентах для прямого взаимодействия с **Comindware Platform** — уникальное преимущество Comindware. RAG-агент интегрируется через эти инструменты в своём потоке поддержки. Оба агента применяют структурированное рассуждение (SGR) для воспроизводимого поведения.

Архитектура:

```
Слой интерфейса (Gradio)
    ↓
Ядро агента (оркестрация, доступ к LLM, состояние сессий)
    ↓
Слой инструментов
    ├── Инструменты Comindware Platform
    │   ├── Приложения и шаблоны
    │   ├── Атрибуты
    │   └── Шаблоны и записи
    └── Утилитарные инструменты
        ├── Поиск и исследования (веб, Wikipedia, arXiv)
        ├── Исполнение кода (Python, Bash, SQL)
        ├── Анализ файлов (CSV, Excel, PDF, OCR)
        └── Математические операции
```

Поддерживаемые поставщики LLM: российские провайдеры — MTC AI, Yandex, GigaChat, Cloud.ru; международные — OpenRouter, Google Gemini, Groq, Mistral, HuggingFace. Поддержка любых OpenAI-совместимых эндпоинтов и возможность добавления кастомных провайдеров.

Для внедрения в РФ: значение по умолчанию (OpenRouter) рассчитано на скорость разработки, а не на промышленный контур с персональными данными и требованиями суверенитета. В таких случаях выбирайте **российские облачные API**, **on-prem** или иной вариант из раздела о комплаенсе и *«Тарифы российских облачных провайдеров ИИ»*.

Ключевые возможности:

- **Многоходовый диалог** — управление памятью в стеке LangChain
- **Потоковая выдача** — ответ по токенам через потоковый API
- **Изоляция сессий** — разделение пользователей и освобождение ресурсов
- **Кросс-языковая поддержка** — ассистент понимает вопросы на английском и отвечает по русской базе знаний; помогает инженерам готовить ответы для англоязычных клиентов и обслуживает иностранных клиентов без отдельной англоязычной базы
- **Восстановление после ошибок** — автоматическая классификация сбоев и адаптация поведения без вмешательства оператора
- **Учёт токенов и бюджета** — фактический расход токенов и оценка стоимости

Агентный слой Comindware Platform применяет краткосрочную память диалога (LangChain) и **корпоративный RAG-контур** для извлечения знаний. Ассистент поддерживает глубокие исследования по корпоративным и веб-источникам. Долгосрочная агентная память — в плане развития, не в текущей поставке.

Текущие ограничения (Comindware Platform 5.0, февраль 2026): сценарии обрабатывают сообщения только из системного чата; привязка сценариев к

пользовательским чатам запланирована в следующих релизах. Данные, сформированные ИИ-агентом, сохраняет пользователь вручную — платформа не записывает изменения автоматически.

Инструменты для отчуждения:

Компонент	Артефакт	Бизнес-назначение
Документация	Руководство по эксплуатации	Инструкции для команд заказчика: развёртывание, настройка, устранение типовых сбоев
Промпты и конфигурации агентов	Набор системных промптов и контрактов вызова инструментов	Воспроизводимое поведение агентов без привязки к конкретной модели
Код	Агентный слой Comindware Platform	Инструменты интеграции с Comindware Platform и утилитарные функции
Тесты	Пакет поведенческих тестов	Регрессионные проверки: стабильность после обновлений модели или конфигурации
Конфигурация	Шаблон переменных окружения	Параметры окружения без секретов — ускорение развёртывания на стороне заказчика
Адаптер LLM	`AIAdapter.zip` (компилируемый)	Единая точка подключения любого LLM-провайдера; передаётся с инструкцией по компиляции и настройке подключения
Режим отладки	Конфигурация `ResponseWithMocks`	Тестирование ИИ-сценариев без обращения к LLM — снижает стоимость приёмочного тестирования на стороне заказчика

7.6 Источники

- Comindware Platform 5.0. Руководство по работе с ИИ — официальная документация, опубликована 13.02.2026
- Comindware — корпоративный сайт

8. Приложение C. Безопасность, комплаенс, наблюдаемость

8.1 Обзор

Приложение определяет **периметр контроля** для промышленного GenAI: модель угроз, комплаенс (ПДн, 152-ФЗ), **наблюдаемость** и паттерны для RAG и агентов.

Основные отчёты содержат **управленческие выводы**; это приложение — детализация для **ИБ и эксплуатации**.

8.2 Практический смысл для ИБ и эксплуатации

Используйте приложение как источник по комплаенсу, периметру данных, наблюдаемости и контрольным требованиям перед промышленным запуском.

Ориентиры по рынку, стеку и дорожной карте — в *Приложении D «Рыночные и технические сигналы»*.

8.2.1 OWASP LLM Top 10 2025

№	Код	Угроза
1	LLM01:2025	Инъекция промптов
2	LLM02:2025	Раскрытие конфиденциальной информации
3	LLM03:2025	Уязвимости цепочки поставок
4	LLM04:2025	Отравление данных и модели
5	LLM05:2025	Неправильная обработка вывода
6	LLM06:2025	Избыточные полномочия
7	LLM07:2025	Утечка системного промпта
8	LLM08:2025	Уязвимости векторного хранилища
9	LLM09:2025	Дезинформация
10	LLM10:2025	Неограниченное потребление ресурсов

8.2.2 OWASP Agentic Top 10 2026

№	Код	Угроза
1	ASI01:2026	Перехват цели агента
2	ASI02:2026	Злоупотребление инструментами
3	ASI03:2026	Злоупотребление идентификацией и привилегиями
4	ASI04:2026	Уязвимости цепочки поставок агентов
5	ASI05:2026	Неожиданное выполнение кода (RCE)
6	ASI06:2026	Отравление памяти и контекста
7	ASI07:2026	Небезопасная межагентная коммуникация
8	ASI08:2026	Каскадные отказы
9	ASI09:2026	Эксплуатация доверия человек–агент
10	ASI10:2026	Вредоносные агенты

Статистика инцидентов (2026): **26,1%** AI-навыков содержат уязвимости *«arXiv: 2601.10338»*; порядка 40 000 экземпляров OpenClaw уязвимы (из 42 665 публично доступных) *«Clawctl»*.

⚠ Границы применения

Не используйте документ как:

- Полную модель угроз без адаптации под профиль заказчика (требуется отдельный анализ).
- Юридическое заключение по 152-ФЗ без согласования с ДПО заказчика.
- Действующую норму по законопроекту об ИИ (на март 2026 г. — черновик, не вступил в силу).

8.2.3 Нормативный контекст (РФ, март 2026 г.):

- **152-ФЗ** (с поправками 2025 г.): основной закон о персональных данных; поправки вступили в силу в июле и сентябре 2025 г. — усилены требования к локализации и анонимизации.
- **123-ФЗ**: федеральный закон об ответственности за вред, причинённый ИИ (действует с 2024 г.).
- **Указ 124/2024**: стратегия развития ИИ до 2030 г.
- **Законопроект о суверенном ИИ**: Минцифры опубликовало проект в марте 2026 г. — требует создания нейросетей в России и обучения на российских данных.

8.3 Сводка доверия для CISO/CIO: что проверять перед промышленным запуском

Здесь дана рамка due-diligence для руководства по ИБ и ИТ для быстрой оценки готовности промышленного GenAI-контура перед запуском.

1. Периметр данных и комплаенс (РФ):

- Зафиксируйте **маршруты данных** (куда уходит контент запросов / контекст / ответы) и режимы размещения (on-prem / облако РФ / гибрид).
- Для ПДн по умолчанию применяйте **минимизацию**: не записывайте полный контент в телеметрию; храните контент отдельно с контролем доступа (см. раздел *«Персональные данные и содержимое в телеметрии (152-ФЗ)»*).

2. Наблюдаемость как контроль качества и бюджета:

- Внедрите обязательные сигналы: токены, задержки, ошибки, глубина агентского цикла, события политик (защитные механизмы, эскалации).
- Для FinOps используйте единый язык метрик (например, OpenTelemetry GenAI semconv) и аллокацию затрат по продуктам / департаментам.

3. Контроль инструментов и сред выполнения (агенты):

- Определите allowlist инструментов / интеграций, модель сети (deny-by-default) и аудит действий агента.
- Для «недоверенного исполнения» (код / широкие инструменты) обеспечьте отдельный контур изоляции и политики доступа.

4. Рамки риска (ориентиры):

- Используйте *«OWASP Top 10 for LLM Applications (2025)»* и *«OWASP Top 10 for Agentic Applications (2026)»* как чек-лист классов рисков, не подменяя ими внутреннюю модель угроз заказчика.

5. Специфические угрозы агентных контуров:

- **Усиление атак:** автономный агент с доступом к инструментам масштабирует компрометацию со скоростью машины — одна успешная инъекция распространяется через все доступные API, файловые системы и интеграции. Изолируйте среду исполнения, применяйте принцип наименьших привилегий, вводите обязательные контрольные точки с участием человека для операций с высоким риском.
- **Дрейф конфигурации:** агенты с доступом к системным инструментам способны непреднамеренно изменить собственные параметры безопасности. Версионируйте политики через неизменяемые артефакты; отделяйте

хранилище политик от хранилища данных (см. *«Паттерны промышленного RAG и защитных контуров»*).

8.4 Стратегия промышленной наблюдаемости ИИ

Система мониторинга стека **Comindware обеспечивает** сквозную прозрачность операций, аудит соответствия 152-ФЗ и контроль влияния на P&L.

Концепция мониторинга:

- **Ситуация:** промышленные агентные системы принимают решения и расходуют бюджет в реальном времени.
- **Вызов:** без сквозной видимости по трассам, метрикам и событиям невозможно доказуемо связать инциденты, качество и P&L — это операционный риск, не технический долг.
- **Задача:** какие сигналы собирать, как связать их с метриками качества и финансовыми показателями, и какие требования учитывать при развёртывании в РФ (152-ФЗ)?
- **Решение:** три уровня AgentOps — наблюдаемость, оценка качества, оптимизация — формируют управляемый контур на стеке Comindware от сбора сигналов до доказанного снижения затрат.

8.4.1 Фреймворк AgentOps

AgentOps — дисциплина управления, мониторинга и улучшения ИИ-агентов в промышленной эксплуатации: инфраструктурный фундамент, без которого агентный продукт остаётся демонстрацией.

Три уровня формируют систему управления — от сбора сигналов до доказанного улучшения результатов:

1. **Наблюдаемость** — восстановление полной траектории каждого решения: вызовы инструментов, обращения к LLM, межагентные взаимодействия. Стек Comindware (Phoenix/Langfuse) обеспечивает детализированную трассировку по всей длине агентского цикла — от входящего запроса до финального ответа.
2. **Оценка качества** — измерение результативности агента по трём ключевым показателям:

- **доля автономного завершения** — доля запросов, выполненных без вмешательства человека;
- **нарушения защитных механизмов** — срабатывания защитных фильтров (утечка данных, неавторизованные действия);
- **точность фактов** — корректность фактических данных (диагностические коды, номера полисов, дозировки).

1. **Оптимизация** — непрерывное совершенствование системы на основе данных первых двух уровней:

- **эффективность токенов промпта** — экономия токенов при сохранении качества вывода;
- **точность извлечения** — доля релевантных документов в top-K результатов поиска;
- **Успешность межагентной передачи** — доля успешных передач задач между агентами;
- **скорость улучшений** — частота внедрения оптимизаций в неделю.

8.4.2 Задача для бизнеса и эксплуатации

Агентная система без **сквозной наблюдаемости** — операционная слепота: инциденты выявляются после факта, регрессии качества не диагностируются, рост нагрузки не связывается с P&L.

Для агентных контуров дополнительно критична **глубина цикла** (число итераций инструмент+LLM) и её корреляция с рисками *«OWASP Top 10 for Agentic Applications»*.

При разработке с агентами учитывайте **инженерные циклы** (план → реализация → проверка → итерация) как самостоятельный источник затрат: суммарные токены, длительность и число циклов стыкуются с OpEx разработки; ориентиры — в *«FinOps и юнит-экономика нагрузки»*.

Долгосрочная агентная память (GAM) — исследовательский класс архитектур (Memorizer/Researcher, многошаговый цикл над памятью), увеличивающий глубину спанов и токены на запрос относительно стандартного RAG. Ключевой риск при внедрении — **отравление памяти и контекста** (Memory and Context Poisoning, OWASP Agentic 2026), требующий аудита записей в долговременное хранилище. В стеке Comindware такие архитектуры рассматриваются как **ориентир НИОКР**, не как обязательная поставка.

8.4.3 Сигналы: трассировки, метрики, события

Ориентир для **единого языка** телеметрии между приложением, шлюзами и бэкендом инференса — семантические конвенции **OpenTelemetry** для генеративного ИИ (статус спецификации на момент подготовки документа: **Development**; при миграции инструментария с более старых версий конвенций используется переменная окружения `OTEL_SEMCONV_STABILITY_OPT_IN`, в т. ч. значение `gen_ai_latest_experimental` — см. *«документацию по спанам»* и *«метрикам»*).

- **Типовое дерево операций для RAG и агента:** получение эмбеддингов, извлечение контекста, генерация ответа, вызов инструментов; для агентских продуктов провайдера — также вызов агента и создание агента там, где применимо к API.
- **Корреляция:** идентификатор трассировки и идентификатор сессии / диалога, чтобы связывать многошаговые диалоги и повторные вызовы.
- **Метрики клиента:** длительность операций и потребление токенов по типу (входные / выходные) — для аллокации затрат и сопоставления с биллингом API провайдера.
- **Метрики сервера инференса (self-hosted, vLLM и аналоги):** длительность запроса, время до первого токена, время на выходной токен — для SLO по задержке, очередям и фазам обработки.
- **События политик:** срабатывания защитных механизмов, отказы по политике, эскалации в режим **человек в контуре** — как отдельные записи или атрибуты, согласованные с матрицей ИБ заказчика.

Ориентиры **токенов/с** из публикаций сообщества и обзоров полезны лишь для грубой прикидки до замеров; в эксплуатации их следует сопоставлять с метриками времени генерации токенов на целевом стенде. Иллюстративные таблицы по локальному железу и MoE — в отчёте *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Дополнительно **OpenInference** описывает инструментирование ИИ-приложений совместимо с OpenTelemetry и поддерживается, в частности, в **Arize Phoenix** (*«OpenInference»*); выбор бэкенда хранения трассов остаётся за заказчиком.

Self-hosted-решения для соответствия 152-ФЗ:

Для российских продакшн-контуров рекомендуются self-hosted-решения наблюдаемости с развёртыванием в российских дата-центрах:

Инструмент	Тип развёртывания	152-ФЗ соответствие
Arize Phoenix	Self-hosted (Docker/K8s)	✅ Полное
Langfuse	Self-hosted (Docker/K8s)	✅ Полное
Helicone	Self-hosted	✅ Полное
SigNoz	Self-hosted	✅ Полное
LangSmith	SaaS (только EU/US)	❌ Не рекомендуется

Для соответствия требованиям локализации приоритет — Self-hosted-решения на базе OpenTelemetry.

8.4.4 Применимость в России: что не блокируется и где нужны оговорки

- **Открытые стандарты (OpenTelemetry, конвенции GenAI)** не привязаны к юрисдикции поставщика; их можно закладывать в проектирование **без ограничений**, сохраняя осознанность, что часть полей спецификации ещё в статусе **Development**.
- **Self-hosted** стеки наблюдаемости (в т.ч. **Langfuse** с открытым ядром, **Phoenix** + OpenInference, корпоративные стеки на Prometheus/Grafana/Tempo и аналоги) при размещении в контуре заказчика или в **облаке РФ** согласуются с типовыми требованиями к **локализации** обработки ПДн — при условии, что **содержимое** промптов и ответов не утекает за пределы разрешённого контура (см. следующий подраздел).
- **Зарубежные SaaS** наблюдаемости (например, **LangSmith**, облако **Arize**) **не запрещены** абстрактно, но для персональных и иных чувствительных данных и для регулируемых отраслей их использование в проде должно проходить через **ДПО, субобработчиков, резидентность** и явное решение по **запрету полного журналирования** промптов/контекста без правовой базы. **Базовая рекомендация** для чувствительного продакшна в РФ: приоритет **on-prem / РФ-размещение** бэкенда телеметрии или режим без передачи текста запросов вне контура.
- **NIST AI RMF** и **профиль GenAI** полезны как **методологический** якорь функции **Measure** (измерение и мониторинг свойств системы) — публикация NIST.AI.600-1; они **не подменяют** 152-ФЗ и отраслевые требования РФ. **EU AI Act** приводить только как **сравнительный** контекст обязанностей к журналированию у поставщиков высокорисковых систем в ЕС, без выдачи этих обязанностей за нормы РФ без отдельного юридического заключения.

8.4.5 Рынок РФ, наблюдаемость LLM и референс-стек Comindware

Для **резидентного** контура РФ приоритетны **self-hosted** или **облако РФ** для телеметрии GenAI: так соблюдается ожидание по **локализации** обработки ПДн и контролю журналов, при этом **OpenTelemetry** и **OpenInference** задают **единый язык** спанов и метрик поверх типовых вызовов LLM/RAG без привязки к единственному коммерческому продукту.

Arize Phoenix в референс-стеке **Comindware** — операционный слой трасс, дашбордов, экспериментов и оценки качества LLM/RAG (в том числе офлайн- и онлайн-оценки) рядом с инференсом (**MOSEC/vLLM**) и корпоративным RAG-контуром. Phoenix не заменяет классический стек **Prometheus/Grafana/Tempo** для инфраструктурных метрик узлов и сети: эти плоскости дополняют друг друга: инфраструктурный мониторинг с одной стороны, GenAI-спаны и контур оценки качества — с другой. Сметные ориентиры — *«Инфраструктура и наблюдаемость: статьи затрат](./20260325-research-report-sizing-economics-main-*

ru.md#sizing_infrastructure_observability_costs)»; связка с контуром оценки — *«Связь с контуром оценки качества»*.

8.4.6 Наблюдаемость в агентах

Наблюдаемость в агентных контурах Comindware выстроена на трёх уровнях:

- **Arize Phoenix:** лёгкий self-hosted мониторинг производительности моделей, визуализация эмбеддеров, отслеживание дрейфа.
- **Langfuse:** более тяжёлый self-hosted вариант с расширенными дашбордами, обработчиком обратного вызова для потоковой передачи и передачей идентификатора сессии.
- **LangSmith:** облачный сервис с мощной трассировкой ключевых функций и наблюдением на уровне проекта.
- **Учёт токенов и стоимости:** встроенный подсчёт токенов + фактические данные от API, стоимость по тарифам провайдеров, коэффициент корректировки накладных расходов.
- **Обработчик ошибок:** классификаторы для каждого провайдера, TF-IDF (частотно-обратная индексная частота) + косинусное сходство для распознавания паттернов ошибок.
- **Система отладки:** категоризированное логирование, потоковые отладчики для каждой сессии, отображение в реальном времени.

8.4.7 Контекст-трекер и диагностика RAG

В корпоративном RAG-контуре Comindware встроенная модель контекста агента фиксирует:

- **Токены диалога и накопленные токены инструментов** — для контроля бюджета и аллокации затрат.
- **Идентификаторы запрошенных статей** — для предотвращения дублирующих запросов.
- **План ответа клиенту и план разрешения инцидента** — оба используют SGR-движок для структурированного рассуждения; аудит каждого шага.
- **Трассировки запросов** — выполнение каждого вызова поиска (исключается из контекста LLM).
- **Диагностику:** время оборота диалога, использованная модель, оценки уверенности, учёт использования.
- **Парсер многоканального ответа (Harmony)** — технология разделения каналов анализа, комментариев и финального ответа для моделей типа GPT-OSS.

8.4.8 AI TRiSM и управление доверием

AI TRiSM (**AI Trust, Risk and Security Management**) — рамка для доверия к ИИ: объяснимость, защита моделей и данных, соответствие требованиям, устойчивость и надёжность. Для **корпоративного RAG-контура** и **агентного слоя Comindware Platform** стыкуется с практиками OWASP (LLM Top 10 2025, Agentic Top 10 2026), минимизацией содержимого в телеметрии, ModelOps и red teaming — **без подмены внутренней модели угроз заказчика**. Финансовые последствия — в *«OpEx безопасности GenAI и агентов»*.

Для CISO/CIO: трактуйте AI TRiSM не как отдельный рынок, а как управленческую рамку: кто допускает модели, как проверяются агенты, граница журналирования и реакция на инцидент.

Рынок подтверждает (2026): 76% организаций называют shadow AI проблемой; только 19% проводят AI red teaming, 29% имеют AI incident response plan. Дефицит не в технологиях, а в зрелости контроля.

Для бюджета: при агентных сценариях и MCP включите мониторинг исполнения, защиту от враждебных запросов, владельца AI governance и отдельный AI incident response план — рядом с наблюдаемостью и комплаенсом.

Model Context Protocol (MCP): стандарт подключения инструментов и внешних ресурсов к агенту через явные контракты вызова. Для бизнеса это означает воспроизводимую интеграцию агентов с любыми системами без написания кастомных коннекторов под каждую модель.

8.4.9 Персональные данные и содержимое в телеметрии (152-ФЗ)

Спецификация OpenTelemetry для GenAI прямо предписывает: **по умолчанию не** записывать системные инструкции, полные сообщения и вывод модели в атрибуты спанов; для зрелого продакшна рекомендуется паттерн **внешнего хранилища контента** с **ссылками** в телеметрии и раздельным контролем доступа (*«раздел о записи контента»*). Это согласуется с минимизацией обработки ПДн: маскирование, сроки хранения журналов, матрица доступа (SRE vs разработка) и исключение из экспорта в недоверенные SaaS без ДПО.

8.4.10 Организационные барьеры и восприятие рисков (опрос CMO × red_mad_robot, 2025)

Публичные материалы исследования **CMO Club Russia × red_mad_robot** фиксируют не только технические, но и **организационные** причины, по которым маркетинговые команды остаются на уровне экспериментов: непрозрачный **ROI**, слабая **стратегия** и опасения по **ИБ**. Полный набор долей и формулировок — в

отчёте *«Зрелость российского рынка GenAI»*; здесь — связка с контуром контроля и OWASP GenAI.

- **Утечка данных (отдельная доля респондентов, не путать с «галлюцинациями»):** в открытых выжимках **~43%** CMO называют **риски утечки** значимой проблемой — это напрямую стыкуется с *«LLM02: утечка конфиденциальной информации»*: минимизация содержимого в журналах и трассах, **одобренные** каналы и модели, запрет **теневого** GenAI с уводом данных вне контура.
- **Качество вывода (другая метрика опроса): ~43%** отмечают **галлюцинации и ошибки** — это риск **качества и доверия**, а не LLM02; закрывается **контуром оценки качества**, защитными механизмами, обязательным подтверждением человеком в критических точках и политиками допустимых сценариев, а не только маскированием.
- **Связка с телеметрией:** даже при корректной архитектуре спанов организация может **записывать в журнал избыточно** из культуры «сохранить всё»; управленческий барьер «безопасность и данные» из того же исследования усиливает требование к **явной политике** выборки и ретенции, согласованной с ДПО.

8.4.11 Связь с контуром оценки качества

Офлайн- и прод-метрики (RAGAS, DeepEval, MERA, LLM-as-a-judge — см. *«Связь с контуром оценки качества»*) усиливаются, если инструментирование по **OpenTelemetry / OpenInference** (в т.ч. **Arize Phoenix** в self-hosted референсе Comindware) даёт стабильные **идентификаторы трассировок и диалогов**, связываемые с оценками, рейтингами пользователей и результатами выборочного разбора. Это даёт регрессии после смены модели, индекса или промпта и обратную связь для дообучения без смешения с необезличенными журналами в публичных облаках.

Связка NIST AI RMF (функция Measure), ISO/IEC 42001 и операционного цикла: профиль *«NIST.AI.600-1 (Generative AI Profile)»* задаёт измерение и мониторинг свойств генеративных систем; *«сопоставление NIST AI RMF с ISO/IEC 42001»* помогает выровнять управленческую **AIMS** с уже принятой рамкой рисков. На уровне практики удобно разделять: **офлайн-оценки** (регрессия на фиксированных наборах до выката) и **онлайн-оценки** на живом трафике без эталонного ответа — см. концепции и гайды *«LangSmith — Evaluation concepts»* и *«онлайн-оценки в LangSmith»*. Рекомендуемый цикл: мониторинг прод → выделение сбоев → превращение в воспроизводимые тест-кейсы → офлайн-оценка качества → выкат → повторная проверка по метрикам и трассам. Направление развития официальных материалов NIST — *«дорожная карта AI RMF»*.

8.4.12 Периметр до LLM: минимизация данных, обезличивание и обратимые подстановки

Для контуров с ПДн и для снижения объёма сырого текста, попадающего к внешней или облачной LLM, целесообразен **каскад до вызова основной модели**:

1. **Детерминированные правила** для структурированных идентификаторов (телефоны, e-mail, реквизиты и т. п.) с валидацией — обеспечивает базовую анонимизацию без потерь.
2. **Контекстные фильтры** и малые модели распознавания сущностей — выявляют и маскируют чувствительные данные в неструктурированном тексте.
3. **Генерация с семантическими плейсхолдерами** — основная работа модели по тексту с плейсхолдерами; отображение пользователю — исходное за счёт обратного мэппинга — гарантирует комплаенс при сохранении UX.

Такой шаблон согласуется с **минимизацией** обработки ПДн по 152-ФЗ и с рекомендациями *«OpenTelemetry GenAI spans»* не писать полный текст запросов и ответов в атрибуты спанов по умолчанию. Инженерные ориентиры по качеству и режимам «быстрый путь / полный каскад» на синтетических корпусах поддержки — в параграфе *«Слой перед LLM и режимы нагрузки (ориентиры для модели затрат)»* отчёта *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Исследовательские аналоги (не норма): адаптивное разделение вычислений между периметром и облаком встречается в работах класса edge–cloud LLM routing (например, *«HybridFlow»*, *«PRISM»*); их имеет смысл цитировать как направления НИОКР, а не как обязательную архитектуру поставки.

8.4.13 Пакет отчуждения: что добавить по наблюдаемости

В *Приложении B «Отчуждение ИС и кода: KT, IP, лицензии, критерии приёмки передачи»* (раздел *«Пакет отчуждения (минимально целостный)»*) к базовому комплекту передачи целесообразно добавить строки по телеметрии: владельца архитектуры наблюдаемости, перечень бэкендов, политику **выборки** (в т. ч. полные трассы для ошибок / канареек), **ретенцию**, эксплуатационный регламент разбора инцидента по `trace_id` и правила маскирования ПДн.

8.5 Паттерны промышленного RAG и защитных контуров

Архитектуры корпоративных ассистентов опираются на классы, не на привязку к одному примеру: локальный vs облачные эмбеддеры/LLM, agentic-цикл с инструментами, CRAG, гибридный поиск, граф знаний, мультимодальное извлечение.

Отчуждение включает класс архитектуры, конфигурацию индекса, политику защиты и сценарии оценки качества.

Ниже — типовые паттерны; примеры из открытых туториалов — в разделе «Справочно».

8.5.1 Классы RAG-агентов (обобщение)

Класс	Назначение для клиента	Типовой технический смысл
Agentic RAG (локальный контур)	Суверенный PoC/Pilot без обязательной отправки данных наружу	Локальные эмбеддеры и LLM, векторное хранилище
Agentic RAG (облако)	Быстрый старт при допустимой передаче данных в управляемый API	Управляемые LLM + динамическая KB
Agentic RAG (reasoning)	Запросы, где важна явная цепочка рассуждений	Модель + инструменты рассуждения
Corrective RAG (CRAG)	Снижение галлюцинаций за счёт самопроверки и повторного поиска	Оркестрация + защитные механизмы
Knowledge Graph RAG	Ответы с опорой на связи сущностей и цитирование	Граф + векторный поиск
Hybrid Search RAG	Точность поиска по смеси лексики и семантики	BM25 + dense embeddings
Vision RAG	Документы, схемы, сканы	Мультимодальные энкодеры + векторный индекс

RAG с графом знаний оправдан при сильной связанности данных (регуляторика, оргструктуры, каталоги). Для плоского документооборота часто достаточна связка векторного и лексического поиска — графовый слой растит сложность без гарантии качества (MTS AI — граф в RAG).

8.5.2 Справочно: примеры из открытых RAG-туториалов

Конкретные связки стеков встречаются в подборках open-source LLM-приложений (модели и БД — разведочный ориентир, не обязательный стек поставки).

Math Tutor Agent с обратной связью (иллюстративная схема):

```
Запрос → входные защитные механизмы (DSPy) → маршрутизатор
    ↓
    ├── совпадение с БЗ (> порог) → векторная БД Qdrant → GPT-4.1
    └── нет совпадения → веб-поиск Tavily → GPT-4o
    ↓
Выходные защитные механизмы → человек в контуре (обратная связь)
```

- **Входные защитные механизмы:** DSPy для фильтрации только академических вопросов
- **База знаний:** Qdrant с OpenAI Embeddings, датасет JEEBench
- **Резервный веб-поиск (Tavily):** при отсутствии ответа в базе знаний
- **Выходные защитные механизмы:** фильтрация галлюцинаций
- **Человек в контуре:** учёт откликов 👍/👎 для улучшения
- **Бенчмарк (в источнике):** 66% точность на 50 случайных JEE-вопросах

Защитные механизмы (типовые слои):

- **Вход:** классификатор типа, фильтрация по области, проверка безопасности контекста.
- **Выход:** блокировка галлюцинаций, соответствие теме, отсечение посторонних ответов.

8.5.3 Пример: RAG для поддержки (по публикации МТС)

МТС описывает рабочий контур на базе **Confluence** и **Jira**: гибридный поиск (pgvector + BM25), локальные эмбеддеры (BGE-m3), нормализация и чанкинг ~3000 символов, раздельные конвейеры индексации и поиска.

Это переносимый ориентир для **RAG гибридным поиском** и **агентным RAG** на **стеке Comindware**, небез использования облака (Хабр, МТС).

В вики-центричных контурах схемы нередко существуют только в редакторе Confluence. Для архива передачи и согласования вне вики полезен отдельный **BPMN 2.0 XML**, а вывод языковой модели требует ревью в стандартном просмотрщике — см. *«Формализация процессов (BPMN 2.0) и генерация с помощью LLM»*.

Угрозы GenAI и поискового слоя (OWASP LLM Top 10): прямые промпт-инъекции, непрямые (вредоносные инструкции в индексируемом контенте) и сохранённые (повторно подмешиваемые инструкции). Атаки на **слой поиска**: отравление корпуса, манипуляция ранжированием, вынуждение извлечь конфиденциальные фрагменты и выдать их в ответе. Таксономия и контрмеры — Kaspersky (prompt injection). Эти материалы — **внешняя разведка угроз**, не обязательный стек.

Антипаттерн (CodeWall инцидент): классические веб-уязвимости (SQL-инъекция) + системные инструкции и защитные механизмы в той же БД, что и пользовательские данные. Компрометация БД позволила подменить политики RAG. **Архитектурный вывод:** разделите хранилище политик, секретов и контента; жёсткое разграничение API и регрессионные проверки после изменений схемы.

8.5.4 Препринты марта 2026: агенты, инструменты, обучение

Срез свежих препринтов помогает сверить архитектурные решения с периметром **OWASP Top 10 for Agentic Applications (2026)**, **контуром оценки качества**, **152-ФЗ** и политикой телеметрии из раздела *«Промышленная наблюдаемость LLM, RAG и агентов»* (в т.ч. рекомендацией OpenTelemetry — semantic conventions for generative AI spans не записывать полный текст промптов в атрибуты спанов по умолчанию).

- **Microsoft Research, arXiv:2603.03205** — фреймворк для повышения **безопасности** агентов при **многошаговых** сценариях с **внешними инструментами**; стыкуется с контролем вызовов инструментов, allowlist и регрессией контрактов MCP/API.
- **OpenAI, arXiv:2603.05706** — набор задач для оценки **контроля рассуждения** при ограничениях на **скрытые** шаги; для эксплуатации — напоминание о **воспроизводимости** оценки качества и различии между наблюдаемым трассом и внутренней цепочкой модели.
- **Princeton University, arXiv:2603.10165** — взаимодействие пользователя с агентом как источник **непрерывного обучения**: для заказчика с ПДн — правовая база, **минимизация**, сроки хранения и **разделение** обучающих журналов от прод-телеметрии; пересечение с политикой маскирования и внешнего хранилища контента в том же разделе про observability.
- **Meta (Экстремистская организация, запрещена в РФ), OpenAI, xAI, arXiv: 2603.01973** — непрерывное улучшение моделей для развлекательных и социальных чатов; **Meta (Экстремистская организация, запрещена в РФ)** на территории РФ **ограничена** — использовать только как **зарубежный НИОКР-контекст**, не как ориентир для закупки или размещения данных без юридической оценки.

8.6 Агенты, инструменты, память и наблюдаемость

Для заказчика в этом блоке значимы только те сигналы, которые меняют требования к **модели угроз**, эксплуатации и комплекту отчуждения; расширенный обзор рынка и продуктовых сигналов вынесен в *«Приложение D — рыночные и технические сигналы»*.

Практический вывод для платформенного ассистента (естественный язык → действия): пользовательский запрос опирается на **устойчивые контракты инструментов** — схемы параметров, семантика ошибок, идемпотентность там, где она уместна. Рост числа инструментов раздувает матрицу регрессии и риск **расхождения версий** между разработкой, приёмкой и продакшном; в ТОМ и комплект отчуждения стоит включить политику **версионирования схем**, поэтапного включения инструментов и явных сценариев **деградации** (безопасный отказ или упрощённый ответ вместо необязательного вызова при сбое бэкенда). Переносимо на любой оркестратор и на внешнее потребление через протоколы вроде MCP: решает не бренд фреймворка, а **управляемость контракта и среды**.

Централизованный реестр MCP-серверов: в открытых материалах red_mad_robot **MCP Tool Registry** позиционируется как способ связать LLM с данными и инструментами в RAG- и агентных сценариях (исходный код — отдельный репозиторий). Для заказчика это усиливает аргумент **управляемого периметра**: **единый перечень допустимых MCP-серверов**, согласованные ревизии и **аудит подключений** в одной плоскости с allowlist и управлением секретами в отчёте отчуждения — см. *Приложение A «Пакет отчуждения (минимально целостный)»*. Первичные ссылки — в параграфе«Источники» этого приложения и отчёта *«Методология разработки и внедрения ИИ»*.

Сопоставление с публичной декомпозицией корпоративной платформы (иллюстрация): на сайте **MWS AI** перечислены модули **MWS AI Agents Platform** — визуальный конструктор сценариев, **LabelX** (разметка), **autoRAG**, **OPS Level** (LLMOps / MLOps / AgentOps / observability), **AutoML**, **NER**, интеграционный хаб, модули продакшна (версии, оповещения, масштабирование, роли) (продукт). В

таблице сведено **логическое** сопоставление с целевой операционной моделью этого документа и комплектом отчуждения, без переноса маркетинговых KPI с лендинга.

Публичный модуль (MWS AI)	Роли TOM (ориентир)	Артефакты отчуждения (ориентир)
LabelX, AutoML	Knowledge Engineer; при необходимости ML-роль	Регламенты разметки, версии датасетов, описание дообучения
autoRAG	Knowledge Engineer, LLMOps / AI Architect	Конфигурация индекса, оценка качества извлечения контекста, сценарии регрессии
OPS Level	LLMOps / AI Architect; AI Security Officer (аудит)	Дашборды, политики алертов, журналы для комплаенса
Интеграционный хаб	LLMOps; владельцы смежных систем	Контракты API, секреты (вне репозитория), эксплуатационный регламент интеграций
Прод-модули (версии, роли, масштабирование)	LLMOps; AI Security Officer	Матрица RBAC, процедуры релиза, SLO

Semantic Gravity Framework: Контроль галлюцинаций через геометрию

Проблема: модель тяготеет к соответствию пользовательским ожиданиям вопреки контексту, генерируя правдоподобные, но неправильные ответы.

Решение: измерение семантического расстояния между ответом, вопросом и извлечённым контекстом с итеративной коррекцией.

Механика (геометрический подход):

1. **Понижение размерности** (Matryoshka Slicing): сжатие векторов до 256 измерений без потери качества (ускорение на 83%).
2. **Физика как движок:**
   - Chain-of-Thought рассматривается как движение частицы в энергетическом поле
   - Высокая энергия = Конфликт между запросом и контекстом
   - Dynamic Beta: «Температура» системы (ниже для сленга, выше для юридических вопросов)
3. **Индекс семантической обоснованности (SGI):** отношение расстояния от ответа до контекста к расстоянию от ответа до вопроса. - SGI > 1.0 → ответ опирается на контекст (желаемое). - SGI < 1.0 → ответ приносится в жертву согласованности с вопросом (галлюцинация).
4. **Итеративная коррекция:** высокая энергия конфликта (низкий SGI) запускает переформулирование с явным требованием придерживаться фактов.

Результаты на испытаниях: 100% соответствие политике безопасности при удержании потребления токенов близким к исходному.

Продуктовый радар: смежные технологии

Краткие обзоры **GraphOS**, **Nested Learning**, **LEANN** и **Perplexica** вынесены в *Приложение D «Рыночные и технические сигналы»*. В этом приложении сохраняются только их последствия для контроля: глубина агентского цикла, изоляция среды, требования к телеметрии, долговременной памяти и минимизации данных.

Защитные механизмы: Архитектурный паттерн для агентов

Типовой стек:

- **Guardrails AI (v2.x):** структурная валидация, кросс-модельный аудит.
- **NeMo Guardrails:** управление диалоговым потоком и тематическими границами.
- **Llama Guard 4 / ShieldGemma:** классификация токсичности и рисков в реальном времени.

Четырёхслойная архитектура:

1. **Input Rails:** фильтрация входа перед передачей модели.
2. **Logic/Tool Rails:** валидация параметров вызовов инструментов.
3. **Output Rails:** проверка вывода на соответствие политике.
4. **HiveTrace:** журналирование всех решений для аудита и разбора инцидентов.

EffGen: Каркас оркестрации для SLM без накладных расходов

Задача: LangChain и AutoGen добавляют значительные накладные расходы на контекст через системные инструкции и схемы, уменьшая эффективное окно для пользовательского ввода.

Решение: четырёхступенчатая компрессия с маршрутизацией по сложности.

1. **Сжатие промптов:** 70–80% сокращение за счёт структурированного кодирования задачи.
2. **Триаж по сложности:** направление простых запросов напрямую в модель, сложные — в многошаговый цикл.
3. **Параллельная декомпозиция:** разбор задачи на независимые подзадачи для одновременного выполнения.
4. **Гибридная память:** векторный поиск для контекста + граф отношений для консистентности решений.

Результаты на 13 бенчмарках:

- Модель 1.5B + EffGen обходит LangChain/AutoGen
- +11,2% эффективности для моделей 1.5B
- +2,4% эффективности для моделей 32B

Протоколы интеграции: MCP, A2A (agent-to-agent), ACP (agent control protocol).

CLI как основа инструментов агентов

Компромисс MCP: возможность в одном сервере служить нескольким сессиям модели сопровождается накладными расходами на схемы инструментов (30–40% контекста) и хрупкой сетевой обвязкой.

Преимущество CLI для агентов:

- **Zero overhead:** флаг `--help` определяет интерфейс без парсинга схем в контексте.
- **Композиция:** нативные Unix-пайпы (`|` , `jq`) для обработки вывода без оберток.
- **Структурированный вывод:** флаг `--json` даёт модели разбираемый результат.
- **Жизненный цикл:** каждый вызов — отдельный процесс с явным кодом выхода.

Паттерн поставки: исполняемый файл + документация (SKILL.md), которую модель читает один раз при инициализации.

Память агентов: от векторного поиска к структурированным графам

Проблема: плоский векторный индекс не различает смысл событий. Запрос похож на два факта одинаково, хотя один критичен (отмена подписки), другой нет (единоразовый платёж).

Решение: слой онтологии и графов знаний поверх извлечения контекста.

Четыре класса графовой памяти:

1. **Графы знаний:** опорные точки фактов (сущности, отношения).
2. **Иерархические графы:** масштабирование через уровни абстракции.
3. **Временные графы:** отслеживание истории и причинности.
4. **Гиперграфы:** сложные N-арные отношения (например, задача связана с несколькими участниками и сроками одновременно).

Почему онтологии критичны:

- Без: Факт «Юзер перешел на Pro-план во вторник» = текстовый фрагмент

- С онтологией: Структурированное событие PlanChange, связывающее Customer + Subscription + timestamp

Практический результат: агент разбирает причину факта, а не механически находит похожие документы.

Мультимодальные модели: интеграция образов и речи

Типовая архитектура:

1. **Энкодеры модальностей:** ViT для изображений, Whisper для речи.
2. **Проектор выравнивания:** трансформация признаков в единое пространство.
3. **LLM:** единая точка рассуждений над объединённым входом.

Два подхода к реализации:

- **Модульный** (LLaVA, Qwen-VL): замороженный LLM + обучаемые энкодер и проектор.
- **Монолитный** (Fuyu-8B): обучение с нуля на мультимодальных примерах.

Применение в RAG и агентах:

- **Document AI 2.0:** OCR и структурное понимание документов в одном проходе.
- **VQA:** визуальные вопросы над графиками, медицинскими снимками, схемами.
- **GUI-агенты:** навигация по веб-интерфейсам и приложениям через пиксельное восприятие (без DOM).

Локальный стек наблюдаемости для агентов

Компоненты:

- **vLLM:** сервер инференса с поддержкой OpenTelemetry.
- **LangGraph:** оркестрация агентского цикла с явными переходами и состоянием.
- **Arize Phoenix:** хранилище трассировок и дашборды качества.

Инструментирование:

- Интеграция телеметрии с агентами.
- Визуализация трассировок операций.
- Мониторинг входных/выходных данных инструментов.
- Метрики потребления токенов и задержек.
- Перехват вызовов LLM и инструментов.

Выходные данные: полная история цепочки мыслей (Chain-of-Thought), параметры вызовов инструментов, потребление токенов, задержки, обнаруженные ошибки.

Типы агентов по стилю рассуждений

1. **Реактивные:** вход → вывод без памяти (термостат).
2. **Делиберативные:** планирование перед действием (шахматист).
3. **Гибридные:** рефлексы плюс медленное мышление (человек-водитель).
4. **BDI-агенты:** явные убеждения, желания и намерения.
5. **Мультиагентные системы (MAS):** специализированные агенты с координацией.

Эталонная архитектура GenAI в продакшне

Пять уровней системы:

1. **Контекст (RAG 2.0)**

Приём документов (Unstructured.io, LlamaParse) → гибридный поиск (BM25 + векторы) → переранжирование (BGE-Reranker) → расширение запроса (multi-query).

1. **Оркестрация**

Управление состоянием через граф переходов (LangGraph) → вызовы инструментов со структурированным выводом → маршрутизация простых и сложных запросов.

1. **Промпты как код**

Шаблоны (Jinja2) → оптимизация под модель (DSPy) → структурированный выход (JSON Schema, Pydantic).

1. **Оценка и защита**

Метрики качества (RAGAS, G-Eval) → фильтры безопасности (NeMo Guardrails) → обязательный разбор критичных действий человеком.

1. **Инфраструктура**

Инференс (vLLM, TGI) → кэш (Redis, GPTCache) → трассировка (Arize Phoenix).

Рекомендуемый стек Comindware:

- **LLM:** локальная
- **Оркестратор:** LangGraph

- **Векторная БД:** Qdrant, Chroma DB, PostgreSQL+pgvector
- **Оценка качества:** DeepEval
- **Мониторинг:** Arize Phoenix (self-hosted)

8.7 MCP, мультиагентная маршрутизация и воспроизводимые навыки

Model Context Protocol (MCP) формализует границу доверия между моделью и инструментами через явные контракты: отдельные серверные процессы с чёткими полномочиями, легко аудируемые и передаваемые клиенту при отчуждении. Маршрутизация — паттерн проектирования сложных агентов: разные наборы инструментов для разных сценариев (код, безопасность, аналитика) без тесной привязки к единому фреймворку.

Ключевой класс риска при росте числа инструментов: легитимный вызов инструмента с вредоносными параметрами (злоупотребление правами, цепочки делегирования, подмена целей). Применяемый стандарт: OWASP Top 10 for Agentic Applications (2026); примеры и противодействие — в Kaspersky: Agentic AI risks и вебинаре Securelist: AI agents vs. injections. Документ не подменяет внутреннюю модель угроз и не предписывает выбор поставщика.

8.7.1 Инспекционный шлюз (AI Firewall)

Весь трафик между пользователем, LLM и инструментами проходит через единый инспекционный слой. Прямой доступ к модели без промежуточной проверки — архитектурная уязвимость, а не упрощение эксплуатации.

Три плоскости контроля:

1. **Входящий трафик:** фильтрация промпт-инъекций до попадания запроса в модель. Каждый запрос проверяется на попытки перехвата управления, подмены системных инструкций или инъекции вредоносных целей агенту.
2. **Исходящий трафик (предотвращение утечек, DLP):** результаты вызова инструментов проверяются на наличие персональных данных (ПДн) и конфиденциальной информации перед возвратом агенту или пользователю — прямое требование 152-ФЗ для контуров с ПДн.
3. **Контроль протокола MCP:** каждый MCP-запрос верифицируется на соответствие политике *«допустимой агентности»*. Инструмент, не входящий в allowlist, блокируется на уровне шлюза — агент не получает даже ответа об отказе от незарегистрированного сервера.

Стек **Comindware** (Phoenix/Langfuse) фиксирует события всех трёх плоскостей в единой трассировке — CISO видит консолидированную картину безопасности, LLMOps сопоставляет события ИБ с токенными затратами и качеством ответа.

> ⚠ **Важно**
>
> Шлюз не заменяет модель угроз заказчика.
>
> Его наличие снижает поверхность атаки, но не устраняет необходимость проведения adversarial testing и red teaming на целевом контуре.

8.7.2 Корпоративный API-слой и MCP

Архитектурный паттерн: корпоративный шлюз управления API и инструментами агентов регулирует доступ к бэкенд-системам через явный реестр MCP-серверов, политики аутентификации и мониторинг.

Пример (МТС): платформа MWS Octapi (продукт) интегрирует ИИ-агентов с корпоративными API через MCP; в публичных материалах описаны практические паттерны подключения и наблюдаемости. Для КП: принять архитектурный класс, адаптировать под свой стек и модель угроз (см. *«Граница доверия и среда исполнения агента»* выше).

Корпоративный LLM-шлюз: управление доступом к облачным и локальным моделям, аудит вызовов, контроль стоимости. Пример: MWS GPT (МТС). Ценообразование — в параграфе *«Тарифы российских облачных провайдеров ИИ»* отчёта по сайзингу.

8.7.3 Классы MCP-интеграций в корпоративном контуре

- **Файловая среда** — доступ только к изолированным корням; обязателен для on-prem KB
- **HTTP / внешние API** — вызовы по allowlist, журналирование
- **Репозитории и CI** — при необходимости через одобренный Git- или API-слой
- **Браузерная автоматизация** — только под внутренней политикой и с трассировкой
- **Кастомные серверы** — закрытый контур (аналог доменных шин, named pipes и т.п.)

8.7.4 Паттерны мультиагентной оркестрации

Класс сценария	Бизнес-смысл	Типовые источники данных / инструменты
Аналитика и отчётность	Сводки, метрики, проверка фактов	БД, файлы, веб-поиск в рамках политики
Комплаенс и документы	Работа с регламентами и договорами	Поиск по корпоративным хранилищам
Операции и CRM	Лиды, коммуникации	Интеграции с корпоративными системами (по согласованию)
Исследование рынка	Внешние сигналы	Новости, открытые API
Подбор / HR	Резюме, вакансии	Изолированные пайплайны с ПДн
Due diligence	Сбор и сверка фактов	Финансовые данные + исследование

Координационные паттерны: последовательное выполнение (pipeline) → параллель с агрегацией (map-reduce) → иерархическая делегация (supervisor) → консенсус (debating agents).

8.7.5 Навыки агента: артефакты для отчуждения

Skill — версионируемый пакет: метаданные (имя, условия вызова, лицензия, версия, автор) + инструкции + контракты инструментов. При отчуждении поставляют: шаблон skill → реестр навыков → процедура ревью и включения в продакшн.

Типовые классы навыков: глубокое исследование, ревью кода, академический поиск, отладка, fact-checking.

8.7.6 Справочно: топологии MCP и маршрутизация

Базовая топология (агент + инструменты):

```
LLM
    ↓
MCP Client
    ↓
MCP Серверы → Инструменты (GitHub, Filesystem, Fetch, etc.)
```

Маршрутизированная архитектура (специализированные агенты):

```
Запрос
    ↓
[Маршрутизатор] → Классификация намерения
    ├── Агент код-ревью → GitHub MCP + Filesystem MCP
    ├── Агент безопасности → GitHub MCP + Fetch MCP
    ├── Агент-исследователь → Fetch MCP + Filesystem MCP
    └── BIM-инженер → Особый MCP (named pipes)
```

Альтернативные оркестраторы: Google ADK, AutoGen и другие SDK предоставляют аналогичные примитивы (инструменты, память, структурированный вывод). Паттерны — смена провайдера модели без смены прикладной логики, вызов инструментов, передача контекста между агентами. Выбор стека закрепляют в архитектурном решении и при отчуждении.

8.7.7 Управление нечеловеческими идентичностями агентов (IAM)

Каждый агент — полноправный субъект доступа.

Управляйте агентными идентичностями по тем же стандартам, что и человеческими учётными записями, с учётом специфики автономного поведения.

- **Уникальные учётные данные:** каждый агент **Comindware Platform** получает отдельные учётные данные. Совместное использование токенов или паролей между агентами недопустимо — компрометация одной идентичности не должна открывать доступ к остальному контуру.
- **Временный привилегированный доступ (JIT-доступ):** выдавайте агенту права исключительно на период выполнения задачи и автоматически отзывайте их по завершении. Постоянные привилегии (standing privileges) с широким scope — архитектурный долг, а не упрощение.
- **Риск-ориентированный контроль доступа:** объём полномочий агента динамически ограничивается уровнем риска текущей операции. Задача с высоким потенциалом ущерба требует минимального scope и явного подтверждения; рутинные операции допускают автономное выполнение в рамках allowlist.
- **Аудит по идентичности:** каждая цепочка решений агента прослеживается до его уникальной идентичности. Это делает возможным разбор инцидента по `trace_id` и подтверждает соответствие требованиям к журналированию (см. *«Персональные данные и содержимое в телеметрии (152-ФЗ)»*).

Рекомендация

Применяйте принцип наименьших привилегий как **архитектурное ограничение**: агент не может получить больше прав, чем явно разрешено его ролью и контекстом задачи.

Отсутствие этого ограничения — пробел безопасности, а не компромисс.

8.8 Управление рисками и комплаенс

8.8.1 Организационные и поведенческие факторы риска

Помимо технических угроз и комплаенса по ПДн, внедрение ассистентов и агентов упирается в **человеческий фактор**:

- команда не верит ответам без **цитирования и оценки**;
- модель допустимого использования **не определена** (какие данные, какие операции);
- страх **ошибки** и **потери контроля** при расширении автономии агента.

Эти барьеры требуют **человека в контуре** при критичных действиях, поэтапное раскрытие прав инструментов, явная **политика использования** и **обучение**. См. *«Стратегия внедрения ИИ»* в методологии + OWASP LLM/Agentic Top 10 + модель угроз заказчика.

8.8.2 Российские правовые аспекты ИИ (Март 2026)

Правовой контур опирается на 152-ФЗ (персональные данные), 123-ФЗ (ответственность за вред ИИ), законопроект о суверенном ИИ (статус на март 2026 — черновик) и отраслевые требования (ФСБ, ФСТЭК для госсектора и КИИ).

Начните готовить архитектурные и комплаенс-решения сейчас — к вступлению законопроектов в силу организация должна быть готова, а не начинать с нуля после принятия.

Важно: статус законопроекта

Проект федерального закона о госрегулировании ИИ (Минцифры, март 2026) — **черновик**, не вступил в силу.

Положения могут измениться. Не используйте документ как действующую норму до официального опубликования.

8.8.3 Проектный контур: законопроект об ИИ (2026)

Статус: на **24.03.2026** речь идёт о **проекте** федерального закона, а не о действующей норме. Первичная фиксация проекта — на портале НПА: проект № 166424 (актуальная редакция, сроки обсуждения и текст — только по материалам портала).

По публичным описаниям проекта (в т.ч. Москва 24, 18.03.2026) обсуждаются направления вроде **маркировки** контента, созданного с применением ИИ, риск-ориентированных требований и возможного **реестра доверенных моделей** для отдельных контуров. **До принятия и вступления в силу** конкретные обязанности нельзя трактовать как действующее право; материалы проекта использовать как **сигнал для дорожной карты** продуктовой и архитектурной готовности.

Ориентиры ниже — по **публичным пересказам проекта** (март 2026 г.); **итоговые** обязательства закрепляются только **принятым** актом и **подзаконными нормами**.

- **Маркировка и уведомление об использовании ИИ:** в обсуждении проекта фигурируют **строгие правила маркировки** контента, созданного с применением ИИ, и логика, при которой **взаимодействие с пользователем** сопровождается **уведомлением об использовании ИИ** (в исходной редакции встречалась формулировка, что уведомление должно **предшествовать или сопровождать** начало взаимодействия). **Обязательный текст, каналы, исключения и санкции** фиксируются только **принятым** актом и подзаконными нормами.
- **Суверенный ИИ (формулировки в повестке проекта):** в обсуждаемых определениях подчёркивается модель, **полностью обученная в РФ на российских данных** при участии **российских субъектов** (точные критерии и исключения — по финальной редакции закона).
  - **Ориентир для референс-стека Comindware:** сочетание **Qwen3** и **GigaChat** с **локальным инференсом** через **сервер инференса MOSEC/vLLM** и (или) с управляемыми API российских облаков **согласуется с этим направлением** при условии локализации обработки и данных в контуре заказчика, соблюдения лицензий на веса, внутренних ЛНА и (для госсектора и КИИ) требований к **допустимым моделям** на дату эксплуатации. Это **не** заменяет самостоятельную правовую экспертизу и **не** тождественно автоматическому включению в будущий реестр до проверки по действующим и вступившим нормам.
- **Реестр доверенных ИИ-моделей, госсектор и КИИ (ФСБ, ФСТЭК):**
  - Использование ИИ в **госсистемах** и на объектах **КИИ** (критическая инфраструктура) в **типовой** практике допуска и аттестации возможно **только** для решений, соответствующих **требованиям к составу средств** и **реестрам**, подведомственным **ФСБ** и **ФСТЭК** на **дату внедрения**

(конкретные перечни, приказы и процедуры — у заказчика и уполномоченных органов).
  - В **проекте** закона об ИИ дополнительно обсуждается логика **реестра доверенных ИИ-моделей**; до принятия норм она **не заменяет** действующие требования ФСБ/ФСТЭК, а **дополняет** дорожную карту.
  - Для иных контуров с повышенными требованиями заказчик **также** опирается на **действующие** нормы на дату проекта.
  - **Модели OpenAI/Anthropic** в варианте **исключительно зарубежного облачного API** для **госсектора** и типовых сценариев **КИИ де-факто запрещены** (не проходят требования к отечественной цепочке поставки и сертификации применительно к ФСБ/ФСТЭК на дату внедрения).
  - Зарубежные облачные API без явного допуска обычно **несовместимы** с ожиданиями госсектора и КИИ без отдельного обоснования.

Трансграничные шлюзы разработки (OpenRouter, OpenCode Zen и аналоги) для **продакшн**-решений с ПДн в РФ **не подменяют** маркировку, локализацию и логику реестра — см. *Приложении B, параграф «Ориентиры для заказчика: инструменты ускорения разработки (вне поставки Comindware)»* и параграф *«Зарубежные API (разработка и песочницы)»*.

8.8.4 Федеральный закон № 152-ФЗ «О персональных данных»

Требования к операторам персональных данных:

С 1 сентября 2025 года все операторы персональных данных обязаны предоставлять анонимизированные наборы данных в государственную информационную систему по запросу Минцифры источник.

Ключевые требования:

- **Локализация данных:** первичный сбор персональных данных граждан РФ должен осуществляться в базах данных на территории России. Локализация данных и обработки в контуре ИИ-агента с локальным хранилищем является обязательной.
- **Аудит соответствия:** операторы должны пройти аудит соответствия 152-ФЗ.
- **Анализ разрывов:** выявление потенциальных нарушений до аудита.
- **Подготовка документов:** соответствие регуляторным требованиям.

С 1 июля 2025 года:

- Ужесточение требований к локализации персональных данных. Запрещена обработка ПДн в иностранных БД.

- Запрет обработки персональных данных с использованием иностранных баз данных для:
  - Сбора
  - Записи
  - Систематизации
  - Накопления
  - Хранения
  - Уточнения
  - Извлечения

Исключения:

- Запрет не распространяется на трансграничную передачу данных (отдельная операция)
- Допускается передача данных за рубеж после завершения целевой обработки

8.8.5 Приказ Роскомнадзора № 140 и жизненный цикл RAG

С 1 сентября 2025 года действуют утверждённые **требования и методы обезличивания** персональных данных (официальное опубликование: publication.pravo.gov.ru — документ 0001202508010002; агрегаторы норматекста — например Контур). Для ИИ-ассистента с RAG это означает явное проектирование процессов: **какие поля** попадают в индекс и обучающие выборки; **какой метод обезличивания** применяется до индексации и при выгрузках; **учёт действий** по обезличиванию; **раздельное хранение** исходных и обезличенных наборов где это требуется политикой. Практика должна быть согласована с ДПО заказчика; настоящий документ не заменяет юридическое заключение.

Методы обезличивания ПДн (Приказ Роскомнадзора № 140)

Метод	Описание	Применимость в RAG
Псевдонимизация	Замена идентификатора на псевдоним	✅ Плейсхолдеры до LLM
Обобщение	Усечение до категории	✅ Маскирование реквизитов
Удаление идентификаторов	Исключение ПДн	✅ Pre-LLM фильтрация
Изменение	Искажение данных	⚠ Ограниченно

8.8.6 Аттестованное облако для ПДн на примере MWS

Для сценариев, где в контуре ИИ обрабатываются **персональные данные** (журналы обращений, обучающие выборки с ПДн, хранилища), заказчик может рассматривать размещение части инфраструктуры в **аттестованном** облачном сегменте вместо полной самостоятельной аттестации собственных площадок — при условии соответствия вида обработки и уровня защиты требованиям **152-ФЗ** и внутренним ЛНА. **MTC Web Services** публикует **IaaS** для персональных данных с уровнем защищённости **УЗ-1** (страница сервиса), специальные условия аттестованного сегмента — в документации MWS; отдельно анонсировался сервис хранения ПДн в облаке (новость MWS). Это **один из** возможных поставщиков на рынке РФ, а не исключительная рекомендация; выбор сегмента и договорных гарантий остаётся за заказчиком и юридической экспертизой.

8.8.7 EU AI Act: Последствия для российских компаний

Принятые запреты (август 2025):

- Биометрическая категоризация
- Распознавание эмоций
- Социальный рейтинг
- Манипуляция поведением через ИИ
- Эксплуатация уязвимостей

Code of Practice (май 2025):

- Добровольный «технический паспорт» для провайдеров моделей
- Конкретные действия для создания безопасных моделей
- Помогает компаниям адаптироваться заранее

Для российских компаний:

- **Компании, работающие с ЕС:** необходимо соответствие EU AI Act для экспорта продуктов.
- **On-premise решения:** критическое условие для работы с данными, которые нельзя выносить за периметр.
- **Комплаенс:** опора на OWASP Top 10 for LLM Applications (2025) и связанные проекты GenAI Security; при агентских сценариях — OWASP Top 10 for Agentic Applications (2026); классическая обвязка HTTP/API остаётся на линейке OWASP WSTG / ASVS — см. *«OWASP AI Testing Guide и граница с классическим веб-тестированием»*; плюс **ЛНА и governance** заказчика по ответственному использованию GenAI (в т.ч. матрица ролей, допустимые данные и модели).

8.8.8 Международный регуляторный контекст: EU AI Act

Раздел — **сравнительный фон** для сделок, где продукт или его компоненты используются клиентами в ЕС. Для резидентного контура РФ приоритет остаётся за **152-ФЗ**, отраслевыми требованиями и действующими ограничениями на дату внедрения; **EU AI Act** становится значимым там, где есть **вывод продукта на рынок ЕС**, роль **поставщика / импортёра / дистрибьютора / оператора** или договорные обязательства перед европейскими заказчиками.

> ⚠ **Юридическая оговорка**
>
> Применимость EU AI Act к конкретному продукту зависит от роли участника, факта вывода на рынок и характера использования в ЕС. Раздел не заменяет юридическую квалификацию сделки и должен использоваться как навигация для бизнес- и архитектурного решения, а не как правовое заключение.

Хронология вступления в силу

- **2 февраля 2025:** общие положения и запрещённые практики.
- **2 августа 2025:** правила для **GPAI-моделей** и связанного governance-контура.
- **2 августа 2026:** основной массив требований, включая прозрачность и большую часть прикладного регулирования для high-risk сценариев.
- **2 августа 2027:** отдельные требования для встроенных регулируемых продуктов и части высокорисковых систем.
- **Digital Omnibus / последующие поправки:** возможны сдвиги по отдельным категориям и срокам; для сделки фиксируйте дату проверки по актуальному официальному тексту.

Территориальный охват

EU AI Act применяется не только к компаниям, зарегистрированным в ЕС: критичны **роль в цепочке поставки**, **факт вывода продукта на рынок ЕС** и **использование системы в ЕС**. Формула «есть клиент в ЕС, значит регламент точно применим целиком» слишком груба; для оффера важнее заранее определить, кто является **provider**, кто — **deployer/operator**, и какие обязательства возникают именно для этой модели поставки.

Структура штрафов (ст. 99)

Для **крупных субъектов (undertakings)** штрафы по ст. 99 обычно считаются как **большее** из фиксированного лимита в евро или процента **мировой** годовой выручки. Для **МСП, включая стартапы**, в соответствующих режимах действует более мягкая

логика с применением **меньшего** из двух ориентиров. Для переговоров это означает простое правило: EU AI Act нельзя оставлять в категории «юридическое потом», если планируется рынок ЕС, high-risk use case или интеграция в регулируемый продукт.

Практический вывод для сделки: при наличии EU-составляющей до вывода продукта на рынок полезно отдельно фиксировать: роль по регламенту, наличие high-risk признаков, требования прозрачности, договорные обязанности по данным и журналированию, а также перечень источников для повторной юридической сверки на дату оффера. Официальные ориентиры: EUR-Lex — Regulation (EU) 2024/1689, EU AI Act Service Desk — implementation timeline, EU AI Act Service Desk — Article 99.

8.8.9 OWASP Top 10 for LLM Applications (2025)

Официальный источник: GenAI Security — LLM Top 10 for 2025.

Краткая русскоязычная выжимка — Habr; для аудита и контрактов используйте официальный текст.

Десять категорий рисков:

1. **Промпт-инъекции (LLM01)** — манипуляция поведением модели через входной текст: прямые, непрямые и сохранённые в данных сценарии.
2. **Утечка конфиденциальной информации (LLM02)** — раскрытие чувствительных данных через ответы, журналы, контуры обучения или поисковый слой.
3. **Уязвимости цепочки поставок (LLM03)** — компрометация моделей, плагинов, наборов данных и интеграций на пути поставки.
4. **Отравление данных и модели (LLM04)** — целенаправленное искажение дообучения, индекса RAG или эмбеддеров.
5. **Небезопасная обработка вывода (LLM05)** — передача ответа LLM в другие компоненты или пользователю без надлежащих проверок и экранирования.
6. **Чрезмерная автономия (LLM06)** — излишние полномочия и действия без ограничений и подтверждений человека.
7. **Утечка системного промпта (LLM07)** — раскрытие скрытых инструкций и внутренней конфигурации приложения.
8. **Слабости векторов и эмбеддеров (LLM08)** — уязвимости векторного поиска и представлений, подрывающие целостность извлечения контекста.
9. **Вводящая в заблуждение информация (LLM09)** — ответы без достаточного обоснования при слабом контроле качества.
10. **Неограниченное потребление ресурсов (LLM10)** — неконтролируемый расход токенов, вычислений и вызовов внешних API, в том числе из-за злоупотреблений.

Регулярный пересмотр: каждый квартал или при смене модели, инструментов, политики.

Финансовые последствия каждого класса риска — см. в параграфе *«Риски внедрения ИИ-проектов»* в отчёте по сайзингу.

8.8.10 OWASP Top 10 for Agentic Applications (2026)

Для агентов (планирование, инструменты, многошаговые цепочки) OWASP выделяет отдельный перечень Agentic Applications 2026. Ниже — краткое резюме.

Десять классов рисков:

1. **Перехват целей агента (ASI01)** — подмена целей, плана или критериев успеха агента вредоносным вводом или подложными данными.
2. **Неправомерное использование инструментов (ASI02)** — вызов разрешённых инструментов с параметрами или в контексте, ведущим к утечке, саботажу или нецелевому расходу бюджета.
3. **Злоупотребление идентификацией и привилегиями (ASI03)** — злоупотребление цепочками делегирования, сервисными учётными записями и правами нечеловеческих идентификаторов.
4. **Уязвимости цепочки поставок агентов (ASI04)** — компрометация сторонних модулей, промптов, навыков, MCP-серверов и иных артефактов среды исполнения; иллюстративный кейс LiteLLM/Telnyx — *«Инцидент LiteLLM и Telnyx (март 2026): цепочка поставок»*.
5. **Неожиданное исполнение кода (ASI05)** — исполнение кода или команд, порождённых или доставленных через недоверенный вывод или цепочку рассуждений.
6. **Отравление памяти и контекста (ASI06)** — устойчивое искажение памяти агента или долгоживущего контекста, влияющее на будущие решения и сессии.
7. **Незащищённое взаимодействие между агентами (ASI07)** — слабая аутентификация, подмена или перехват сообщений между агентами и оркестратором.
8. **Каскадные отказы (ASI08)** — распространение ошибки или компрометации по сети взаимодействующих агентов и сервисов.
9. **Эксплуатация доверия человека к агенту (ASI09)** — манипуляция пользователем через доверие к «авторитету» агента (социальная инженерия через интерфейс или рекомендации).
10. **Вышедшие из-под контроля агенты (ASI10)** — агенты, вышедшие из-под политики: персистентность вредоносного поведения, обход мониторинга или коллюзия.

Интеграция с тестированием: при расширении инструментов, MCP или маршрутизации агентов переработать регрессионные тесты для Agentic Top 10

наряду с LLM Top 10. Для **Comindware**: оба перечня → модель угроз → бюджет безопасности (см. *«OpEx безопасности GenAI и агентов»*).

8.8.11 Инцидент LiteLLM и Telnyx (март 2026): цепочка поставок

В **марте 2026 г.** группировка **TeamPCP** скомпрометировала в **PyPI** пакеты **LiteLLM** и **Telnyx SDK**. По открытым разборам:

- в одном случае вредоносный код активировался через механику `.pth` при импорте;
- в другом — через **стеганографию в WAV**, что затрудняло обнаружение **типовыми SCA-практиками**.

Практический вывод для агентских и LLM-шлюзов: одних **сигнатурных** проверок зависимостей недостаточно; нужны **поведенческий анализ зависимостей**, **жёсткий контроль исходящего трафика** и отдельный **мониторинг цепочки поставок Python-пакетов**.

Роль в документе: кейс иллюстрирует класс **Agentic Supply Chain Vulnerabilities (ASI04)** в OWASP Agentic Top 10 (2026) и **не** задаёт отдельную таксономию рисков.

8.8.12 OWASP AI Testing Guide и граница с классическим веб-тестированием

OWASP AI Testing Guide — рамка проверки ИИ-систем (атаки, методология, связь со стандартами). Применяйте её **наряду с** проверками **HTTP/API-обвязки**, не вместо них.

Важно для команд, привыкших к ZAP и «OWASP Top 10 для веба»:

- **Сканеры (в т.ч. ZAP)** не подменяют проверку по **OWASP Top 10 for LLM**: это другой контур рисков, «готового соответствия» из одного прогона DAST не получить.
- **Классическая обвязка:** чеклисты в духе OWASP Web Security Testing Guide (WSTG) остаются релевантны для **шлюза, админки, REST/WebSocket** вокруг RAG.
- **GenAI:** перечень рисков LLM/агентов требует **отдельных** сценариев — промпт-инъекции, поиск по знаниям, агентские инструменты.
- **Ручное тестирование веб-приложений (вводные):** Habr — OWASP; углубление по отдельным темам — Habr, Habr.

Наборы запросов для adversarial-оценки (только легитимный red team): при проектировании регрессионных тестов опасности полезны открытые корпуса вроде AdvBench / llm-attacks, XSTest, Do-Not-Answer, ToxicChat, WildJailbreak; коллекции

джейлбрейков (например, L1B3RT4S) применять **только** в изолированных стендах и с явной политикой этики — не как инструкции для злоупотреблений.

8.8.13 Machine Unlearning: Право на забвение

Проблема: классические методы защиты (промпты, классификаторы, RLHF) блокируют генерацию контента, но персональные данные остаются внутри модели. Это конфликтует с:

- GDPR (Европа)
- 149-ФЗ (Россия)

Решение: OpenUnlearning с поддержкой LoRA для экономии GPU-памяти.

Набор забывания и набор сохранения:

- Набор забывания (Forget Set) — данные, которые модель должна забыть
- Набор сохранения (Retain Set) — данные, которые должны остаться

8.8.14 Безопасность ИИ-агентов

Моделирование угроз:

Уровни автономности от инференса до полной автономности. Наибольшую опасность представляют агенты с доступом к инструментам: - Отправка email через LLM - Совершение покупок - Физические действия (регулирование температуры)

Чат против агента: обход ограничений в **диалоге** (jailbreak) и компрометация **агента с инструментами** — разные классы последствий: второй вариант масштабирует ущерб через API, файловые системы, биллинг токенов и интеграции. Здесь опираемся на OWASP Agentic Top 10 (2026) и внешние разборы вроде Kaspersky — agentic risks / ASI.

Защитные меры:

- Защитные механизмы как обязательный инфраструктурный слой
- Входные защитные механизмы: Защита от промпт-инъекций
- Логические и инструментальные защитные механизмы: Валидация параметров
- Выходные защитные механизмы: Проверка соответствия политике бренда
- HiveTrace: Полная видимость и отладка

8.8.15 Справочно: граница доверия, сеть и среда исполнения агента

Для сценариев с **вызовом инструментов**, **генерацией команд или кода** и **недоверенным входом** (загруженные скрипты, ноутбуки, внешние MCP) инженерная граница сильнее определяется **моделью доверия и угроз**, чем удобством привычного стека разработки. В NIST SP 800-190 (Application Container Security Guide) отмечено: контейнеры **разделяют ядро ОС** с хостом и в типовой постановке **не обеспечивают** ту же степень сегментации, что аппаратно виртуализованная ВМ через гипервизор; сила барьера существенно зависит от возможностей, пространств имён и монтирований. Отсюда — осознанный выбор класса среды (ужесточенные контейнеры → прослойки перехвата системных вызовов → отдельная ВМ на задачу / microVM) **под сценарий**, а не использование «контейнера рядом с сервисом» как универсального ответа на недоверенное исполнение.

Сеть и удостоверение — типовые рычаги снижения масштаба инцидента: исходящий трафик из среды задачи — **по умолчанию запрещён** или задаётся строгим allowlist под конкретную цель; внутренний периметр **не** отождествляют с безопасностью сам по себе. Доступ к данным и API выдают **краткоживущими** учётными данными с **минимальным scope**, привязанными к **конкретному запуску**, без размещения долгоживущих production-секретов «на всякий случай» внутри песочницы — в связке с OWASP Top 10 for Agentic Applications (2026) (в т.ч. злоупотребление инструментами и учётными записями).

Антипаттерны при проектировании: совмещение **агентского runtime** с критичным production в одной зоне доверия без разделения; выдача недоверенной нагрузке доступа к **сокету Docker**, привилегированным монтированиям каталогов хоста или долгоживущему общему рабочему пространству между несвязанными акторами без политики изоляции.

Принцип управляемой плоскости исполнения: наружу из среды исполнения операционно предпочтительно выводить **артефакты, журналы аудита и результаты проверок**, а не необузданные прямые действия по prod-системам; любой прямой доступ агента к боевым контурам — через явные политики, брокеры доступа и наблюдаемые шлюзы. См. также *«Методология разработки и внедрения ИИ»* (инженерия обвязки) и раздел «**Промышленная наблюдаемость LLM, RAG и агентов**» выше по настоящему документу.

8.8.16 Справочно: модель риска, паттерны среды и минимальный состав платформы

Принцип выбора: класс среды исполнения и сила изоляции задаются **моделью риска** — насколько недоверен вход, нужен ли широкий выход в сеть, **артефакт или**

действие является допустимым результатом, нужен ли **долгоживущий** mutable state, — а не только привычным runtime команды. Ниже — переносимые **инженерные ориентиры** для согласования с ИБ и владельцем процесса; конкретные продукты и SLA заказчик выбирает отдельно.

Ориентиры по классам сценария:

- **Недоверенный пользовательский код** (скрипт, SQL, ноутбук, загруженный материал): максимально сильная граница (ВМ / microVM), исходящий трафик **по умолчанию запрещён**, удостоверение **на один запуск**, наружу — **только артефакты** (результат, журнал, согласованная выборка данных), без произвольных действий от имени prod.
- **Внутренний агент для программирования (PR, CI):** ужесточенный контейнер или изоляция уровня ВМ по политике ИБ, **репозиторий только для чтения**, **без** доступа к prod-контурам, выход — **артефакты** (например, **diff**, отчёты проверок), а не прямая запись в защищаемую ветку без человека или отдельного согласованного конвейера.
- **Долгоживущая dev-среда** (разработчик и агент совместно, часы и дни): **persistent** workspace, снимки и восстановление, разделение **системного** и **пользовательского** слоя состояния, **контролируемый** доступ к state и сети (см. паттерн 2 ниже).
- **Регулируемый корпоративный контур** (чувствительные данные, комплаенс, enterprise): по возможности **self-hosted** control plane, **резидентность** и границы обработки данных, **полный** аудит, строгие удостоверения и **исполнение политик** на каждом шаге.

Паттерн 1 — песочница под pull request (иллюстративный ориентир): типовое **событие** — открытие PR или явная команда в интерфейсе ревью (например, `/ai review`). **Вход:** репозиторий (или зеркало), хэш коммита, диапазон изменений, тип задачи. **Изоляция:** базовый образ и дерево исходников репозитория **только чтение**; **рабочая директория** — временная и **уничтожается** после прогона. **Сеть:** по умолчанию всё запрещено; в allowlist обычно включают **зеркало Git**, **зеркало комплектов** и **хранилище артефактов**. **Удостоверение: краткоживущий** токен (в практике — **порядка минут**) с минимальным scope: чтение кода, запись **только** в хранилище артефактов; **запрет** прямой записи в основную / защищаемую ветку и в «истинный» репозиторий без отдельного контура. **Результат:** файлы **diff**, отчёт тестов, текстовый отчёт. **Смысл для ИБ и КТ:** агент **не** заменяет собой принятие изменений в ИС — в основной контур попадает то, что прошло **ревью** артефактов и согласованный **CI** / политику веток.

Паттерн 2 — долгоживущая среда с агентом: сценарий непрерывной работы **часами и днями**. **Состояние: неизменяемый** базовый образ; **отдельное** пользовательское хранилище; общие данные — только через **явно заданные** точки

доступа (тома, объектное хранилище, ACL). **Жизненный цикл:** автоматическое завершение при простое, **снимки** состояния, возможность **отката**. **Доступ: SSO** и **временная сессия**. **Секреты:** не хранятся в долгоживущем слое среды; выдача **по запросу** через **посредника** (брокер секретов / access broker). **Сеть:** выход в интернет **по умолчанию выключен**, разрешён **allowlist**. **Изоляция:** среда **на пользователя** или **на задачу**, **без** общего файлового пространства между несвязанными акторами. **Компромисс:** выше сложность управления состоянием и стоимость сопровождения, зато сохраняется интерактивная работа с **полной историей** изменений.

Минимальный состав платформы безопасного выполнения задач (ориентир «с чего начать»): точка входа для задач; сервис **политик** (принятие решений по правилам); **планировщик** сред исполнения; **хранилище образов**; сервис **выдачи прав** доступа; **хранилище результатов**; **журналирование и аудит**. Имеет смысл **отложить** до роста зрелости: избыточно сложное горизонтальное масштабирование, тонкая **оптимизация затрат** и полноценный **пользовательский интерфейс** — при условии, что перечисленный **минимум** уже закрывает базовую безопасность и воспроизводимый цикл «задача → среда → результат → аудит».

8.8.17 Справочно: управляемые песочницы, сравнение моделей и бенчмарки

Три измерения выбора: при подборе **управляемой** среды для агента удобно явно зафиксировать **модель сессий** (эфемерная vs сохраняемое состояние и снимки), **модель сети** (что разрешено исходящему трафику по умолчанию и как задаётся allowlist) и **модель размещения** (мультиарендный SaaS, регион, возможность контура заказчика). Ниже **E2B**, **Modal** и **Daytona** приведены как примеры **разных** операционных моделей, а не как рейтинг «лучший поставщик»; лимиты, цены и **дефолты сети** необходимо **сверять по документации на дату внедрения**.

Сжатое сравнение (иллюстрация):

Поставщик	Сессии и снимки (ориентир)	Сеть и контроль (ориентир)	Размещение (ориентир)
E2B	Изолированные sandboxes и шаблоны; непрерывная работа до **24 ч** (Pro) или **1 ч** (Base), пауза и возобновление с сохранением состояния — см. Sandbox lifecycle; публичные материалы также описывают ускорение старта (E2B Blog)	Политику исходящего доступа и ограничений задавать **явно** по документации продукта	Уточнять опции развёртывания и изоляции у поставщика (в т.ч. облако заказчика) в актуальных условиях
Modal	**Sandboxes** для недоверенного или агентского кода; время жизни по умолчанию **5 минут**, до **24 ч** параметром, при необходимости дольше — снимки файловой системы и последующее восстановление — см. Modal — Sandboxes	Исходящий доступ, туннели и секреты настраиваются конфигурацией Sandbox и образа; детали — в документации **Sandbox** и смежных разделах	Управляемая облачная платформа Modal
Daytona	Открытая модель **sandboxes** для исполнения кода с API/SDK — см. Daytona Documentation; политики автостопа, хранения и региона — по разделам продуктовой документации	В публичных сравнениях для Daytona часто подчёркивают **закрытую по умолчанию** сеть и наличие **аудит-журналов** — подтверждать в **текущей** документации	Выбор региона и требования к данным — по документации и договору с поставщиком

Бенчмарки и выбор runtime: сравнение **не** строят на тривиальном запуске вроде одной команды `python -c "print(1)"` и **не** сводят к синтетической нагрузке только на CPU — оценка должна повторять **реальную** рабочую нагрузку (клонирование репозитория, установка зависимостей, прогон тестов, операции с файлами, участки с интенсивным вводом-выводом и обращениями к ядру). Для **gVisor** официально указано: «While gVisor is able to sandbox any application, it should generally **not** be used to sandbox **every** application» — см. gVisor — Production guide; ограничения совместимости и профиль накладных расходов — Compatibility и Performance. **Вывод:** одни лишь **задержка и пропускная способность** отражают

удобство, а **не** пригодность для **безопасного** исполнения агента без учёта политик сети, аудита и лимитов.

Академический ориентир по компромиссу безопасность—производительность: в работе *Security-Performance Trade-offs of Kubernetes Container Runtimes* (Umeå University; IEEE) анализируют соотношение затрат и защищённости для разных container runtime в Kubernetes — см. IEEE Xplore. Агрегированные публичные обзоры и вторичные слайды нередко приводят **порядковые** соотношения между **runC**, **gVisor**, **Kata** и **microVM**; такие цифры полезны как **предварительный** ориентир, но **не** заменяют прогон на **своей** кодовой базе, в **CI** и с учётом **модели угроз**.

Минимальный набор проверок при приёмке среды: холодный старт (от создания образа или снимка до готовности интерфейса командной строки) → тёплый старт (от восстановления снимка до первой команды) → клонирование репозитория → установка зависимостей → запуск тестов → файловые операции → исходящие соединения: **разрешённые** и **ожидаемо запрещённые** сценарии.

Безопасность и эксплуатация (что держать в смете проверок наряду с latency):

- политики **egress**: как разрешаются и блокируются исходящие соединения;
- **наблюдаемость и аудит**: полнота журналов и возможность разбора действий внутри среды;
- **лимиты ресурсов и частоты** запросов: защита от перегрузки и злоупотреблений;
- **интеграция** с сетевой инфраструктурой заказчика и **стоимость одного** успешного выполнения задачи.

Метрики в промышленной эксплуатации: смотреть не только на p50 времени **создания и восстановления** среды, но и на **p95** «длинного хвоста»; дополнительно — доля **успешных** выполнений по типам задач, **частота сбоев** при восстановлении из снимка, **стоимость** одного успешного выполнения. **Контрольные сигналы** для регулируемых контуров: число **заблокированных** исходящих попыток, **отклонённых** действий на уровне API и системных вызовов, **покрытие аудита** (доля задач с полным журналом), попытки **несанкционированного** доступа по токенам и ссылкам, регрессионные проверки **мультитенантной** изоляции. Переносимый критерий: хорошая sandbox-платформа **предсказуема**, **ограничена** политиками и **аудируема**, а не только «быстрая».

8.8.18 Справочно: безопасный MVP контура исполнения за 30 дней, дискуссия по средам и выводы

Ориентир по сроку: за **30 дней** реалистично вывести **безопасный минимально жизнеспособный** контур под **один узкий сценарий** (например, только агент для **pull request** или только агент для **аналитики**), но **не** «универсальную платформу» под все классы задач сразу. Ниже — переносимый **скелет недель** для согласования с ИБ и владельцем продукта; календарь и объём команды уточняют на месте.

- **Неделя 1 — модель угроз:** выбрать **1–2** сценария использования; описать **границы доверия**, входные данные и **масштаб возможного ущерба**; **не** начинать разработку среды исполнения без этого шага.
- **Неделя 2 — посредник, среда, сеть:** поднять или подключить **управление секретами** / брокер доступа; зафиксировать **тип** среды исполнения под угрозу; настроить **сетевые политики** с **запретом по умолчанию** и явным allowlist.
- **Неделя 3 — снимки, артефакты, аудит:** базовые **снимки** (где уместно), передача **результатов** выполнения как **артефактов**, **журналирование** и обработчики **жизненного цикла** среды.
- **Неделя 4 — тесты, враждебные сценарии, пилот:** проверки на **враждебные** кейсы — утечка данных, нежелательное **открытие портов**, доступ к **служебным** адресам, утечки через **общее состояние**, **бесконечные** циклы и чрезмерное потребление ресурсов; без такого набора остаётся **демонстрация**, а не управляемая **платформа**.

Критерии готовности к выводу за пределы «демо»: работает **аварийное отключение** (kill switch / немедленная остановка среды); применяются **лимиты ресурсов**; в среде исполнения **нет** production-учётных данных и долгоживущих **боевых** секретов «про запас».

Вопросы для дискуссии (архитектура заказчика): типичная ошибка — искать **одну** «правильную» среду для всех сценариев; на практике почти всегда нужны **как минимум два класса**: **жёстко изолированная краткоживущая** среда и **удобная** среда с **сохранением состояния**. Имеет смысл явно зафиксировать позицию по:

- **Интернет для агента для программирования?** полный выход в сеть или только **доверенные** зеркала (Git, комплекты) и артефакты;
- **Прямые изменения или только артефакты?** запись в репозиторий **напрямую** vs передача **diff** и отчётов на **ревью** (см. *Приложении B, параграф «Агент в PR и артефакты вместо прямой записи в ИС»*);
- **Предпрогретые среды или контейнеры с более слабой изоляцией?** компромисс **стоимости и надёжности** warm pool против экономии на границе доверия — закладывать в TCO и модель рисков (см. отчёт *«Сайзинг и экономика»*);

- **Граница между изолированным исполнением и полноценной удалённой разработкой?** где заканчивается **песочница задачи** и начинается **долгоживущий** dev-контур с политикой state;
- **Отдельный класс среды для недоверенного входа?** нужен ли **более жёсткий** контур для вредоносного или потенциально вредоносного ввода отдельно от внутреннего PR-агента.

Выводы для руководства и архитектуры (исполнение кода в контуре ИИ):

- **Сдвиг доверия:** проблема не в том, что **LLM генерирует код**, а в том, что ей **доверяют выполнение действий** в инфраструктуре; **командная строка**, **файловая система**, **сеть** и **состояние** превращают модель из «помощника» в **полноценного исполнителя** — архитектуру проектируют исходя из этого.
- **Контейнер — не финальная граница** для недоверенного исполнения: см. NIST SP 800-190; при необходимости — более строгие механизмы (BM / microVM, политики сети) — в связке с подразделом «**Справочно: граница доверия, сеть и среда исполнения агента**» выше.
- **Отдельный контур выполнения:** либо он выделен и отделён от **управляющего** контура (политики, оркестрация, выдача прав), либо риск для **production** принимается **осознанно**; отсутствие такого разделения — **архитектурный долг**.
- **Тестирование на своей нагрузке:** выбор среды без прогона на **реальных** задачах и репозиториях — ошибка **управления**, а не только технический риск (см. также **бенчмарки** в параграфе про управляемые песочницы выше).
- **Измерять масштаб ущерба, а не только задержки: заблокированные** соединения, **отклонённые** действия, тесты **изоляции** между пользователями, **стоимость одной** задачи — это **ключевые** метрики безопасности и эксплуатации, а не «опциональное наблюдение».

8.8.19 NIST AI RMF 1.0 (GenAI Profile)

Рамка NIST AI RMF выделяет четыре функции (английский термин — для сопоставления с первоисточником):

1. **Установление рамок (Govern):** политики ответственности и этики.
2. **Картирование (Map):** идентификация рисков галлюцинаций и утечек данных.
3. **Измерение (Measure):** количественная оценка рисков через бенчмарки (MERA, ruMMLU).
4. **Меры снижения рисков (Manage):** внедрение мер технического контроля (входные и выходные защитные механизмы).

ISO/IEC 42001:2023 определяет требования к системе менеджмента ИИ (AIMS): политика, роли, жизненный цикл, оценка рисков. Для согласования с NIST AI RMF используйте официальное сопоставление. Сертификация по ISO 42001 не подменяет требования 152-ФЗ и отраслевые нормы РФ; рассматривайте её как опцию зрелости для крупных организаций.

8.8.20 Рынок безопасности ИИ в России

Ключевые игроки:

- HiveTrace — безопасность GenAI-приложений
- Raft — инструменты безопасности LLM
- i-legal — юридические услуги для ИИ

Тренды регулирования:

- Мягкое регулирование в России (подход «сначала технология, потом регулирование»)
- Казахстан: принцип «сначала технология, потом регулирование»
- ОАЭ: намеренное отсутствие регулирования для развития технологий

Рекомендуется:

- Разработать сервис мониторинга токсичного контента для чатботов
- Инструменты очистки персональных данных для RAG-систем
- Бюджет на очистку датасетов при подготовке ИИ-систем

8.8.21 Практические рекомендации по комплаенсу

Требование	Действие	Срок
Аудит 152-ФЗ	Проверка локализации данных	До запуска
Очистка данных	Маскирование ПДн для RAG	При индексации
Защитные механизмы	Внедрение защитных слоев	При проектировании
Мониторинг	Токсичный контент, ПДн	Постоянно
Аудит безопасности	Red teaming, пен-тесты	Ежеквартально
Верификация API/ шлюза (классика)	Требования OWASP ASVS 5.0 (русская редакция) к поверхности входа/выхода ассистента	До пром и при крупных релизах
GenAI / агенты	Сценарии по OWASP LLM Top 10 2025 и при наличии инструментов — Agentic 2026	Ежеквартально и при смене модели или MCP

Закон о страховании рисков ИИ: принят для разработчиков в экспериментальном правовом режиме. Консорциум при Минцифры занимается вопросами безопасности ИИ.

8.8.22 Санкции и доступность технологий

Импорт GPU:

- Параллельный импорт: +30–50% к стоимости
- Логистика: +10–20%
- Российские премиум-наценки на H100, A100, RTX 4090 (**24 / 48 ГБ**), профессиональные SKU (**RTX PRO 6000** и аналоги)

Аренда GPU (IaaS / dedicated) — чеклист комплаенса:

- **Резидентность и субпроцессинг:** юрисдикция площадки, оператор ЦОД, цепочка субобработчиков и соответствие **152-ФЗ** / внутренним регламентам.
- **Изоляция:** multi-tenant vs dedicated, периметр сети, доступ администраторов провайдера к данным и образам.
- **Журналирование и наблюдаемость:** что пишет провайдер, ретенция, выгрузка для расследований; согласование с политикой телеметрии до LLM (см. соседние разделы этого приложения).
- **Вывод модели и ПДн:** запрет на обучение на данных заказчика, обработка выходных токенов и облачных журналов — в договоре и DPA.

- **Происхождение железа:** класс GPU и цепочка поставки остаются релевантны и при аренде (санкции, реэкспорт, план Б при ограничении площадки). Сводная матрица поставщиков аренды — в параграфе *«Цены на GPU-оборудование (покупка и аренда)»*.

Аналоги и решения:

- Китайские ускорители (Huawei Atlas 350 на Ascend 950PR — 2.87x быстрее Nvidia H20)
- Российские облачные провайдеры (Cloud.ru, Yandex Cloud, SberCloud, MWS GPT, Selectel — по контуру)
- Open-source модели с локальным инференсом

9. Приложение D. Рыночные и технические сигналы

9.1 Стратегический контекст

Консолидация ключевых технологических индикаторов, рыночных трендов и векторов развития ИИ-инфраструктуры на горизонте 2026 г.

Сметы и тарифы — в *«Сайзинг и экономика»*.

Модель внедрения и качество — в *«Методология разработки и внедрения ИИ»*.

Комплаенс, модель угроз и наблюдаемость — в *Приложении C «Безопасность, комплаенс, наблюдаемость»*.

9.2 Инвестиционные ориентиры

- **Технологический стек:** доминирование архитектур MoE, VLA и edge-инференса определяет структуру будущих обновлений.
- **Модель владения:** устойчивая окупаемость on-prem решений наступает при горизонте планирования свыше 3 лет.
- **Суверенитет:** резидентные модели (GigaChat, YandexGPT) являются базовым стандартом комплаенса для КИИ и госсектора.

Для переговоров:

- Кадровый дефицит (~10 000 профильных специалистов) определяет динамику рынка труда и условия передачи экспертизы.
- Устойчивый спрос на управляемый GenAI-контур в сочетании с концентрацией на зарубежных SaaS формирует риск утечки данных.

⚠ Границы применения

Не используйте документ как:

- Единственный источник показателей для сметы: внешние обзоры **не** заменяют отчёт *«Сайзинг и экономика»*.
- Юридическое заключение по нормам РФ и трансграничной передаче данных без согласования с юридической службой заказчика.

9.3 Продуктовый радар и архитектурные векторы

Модель угроз и 152-ФЗ для **корпоративного RAG-контура** — *Приложение C «Безопасность, комплаенс»*.

Сводка трендов — *«Тренды 2026 года»*.

9.3.1 GraphOS: высокоуровневая архитектура RAG

Многослойный промышленный стек на базе графов знаний, обеспечивающий адаптивную маршрутизацию запросов:

- **Компоненты (по обзору):** граф знаний **Neo4j**, оркестрация **LangGraph**, **LiteLLM**; **Request Routers**; трёхуровневая память (**Redis** + **Neo4j** и др.); наблюдаемость через **Prometheus** / **Grafana** / **LangSmith**; развёртывание **Docker** / **Kubernetes**.
- **Заявленная логика:** маршрутизация запросов к разным стратегиям извлечения контекста вместо пайплайна «всегда полный».
- **Заявленная экономика:** ориентир **30–50%** экономии стоимости запросов; во вторичных обзорах отдельно фигурирует **~47%**. Перенос в КП и SLA — **только** после подтверждения на данных заказчика.

Источник по архитектуре и цифрам: *«Building Agentic GraphOS: The 16-Layer Architecture Behind Production-Ready Knowledge Graphs»* (Medium, вторичный обзор).

9.3.2 Nested Learning: Transformer 2.0

Исследовательский горизонт (концепции 2025–2026 гг., включая отсылки к NeurIPS 2025):

- модели с «быстрыми весами» под текущую задачу и «медленными» под фундаментальные знания;
- модуль **HOPE** (Higher-Order Processing Engine);
- акцент на **continual learning**.

В коммерческом внедрении — **НИОКР**, не обязательная строка сметы до зрелых API и лицензирования в контуре заказчика; юрисдикция размещения данных и обучения — отдельный комплаенс-разбор. Дополнительный внешний обзор — публикация о Nested Learning.

9.3.3 Perplexica: открытый стек в духе Perplexity

Архитектурный пайплайн (как в обзорах):

1. Классификация намерения.
2. Генерация поисковых запросов.
3. Агрегированный поиск (например, SearXNG, Exa, Tavily).
4. Реранжирование.
5. Генерация ответа с опорой на источники.

Архитектурно значимые свойства: поддержка **Ollama** (локальный инференс), API-first (`/api/search`, `/api/chat`). Цепочка «веб-поиск + ранжирование + генерация» предопределяет политику исходящего трафика, режим журналирования запросов к внешним поисковикам и требования к минимизации ПДн в журналах — согласно модели угроз (*«Приложение C»*).

9.3.4 LEANN: компактный векторный индекс

Идея: вместо хранения всех плотных эмбеддеров — индекс, предполагающий **пересчёт представлений по требованию** (компромисс **хранилище ↔ вычисления на запрос**). В открытых обзорах и репозитории приводится пример масштаба **~60 млн** текстовых чанков: классический вариант с хранением векторов **~201 ГБ** против **~6 ГБ** для подхода LEANN (**оценка «~97%» снижения объёма индекса** — требует подтверждения на корпусе заказчика и SLA по latency). Первичные источники: *GitHub — yichuan-w/LEANN*, *Semantic Scholar — LEANN: A Low-Storage Vector Index*; дополнительный обзор — публикация о LEANN.

9.3.5 Тренды 2026 года

Приведённые ниже сигналы служат внешним ориентиром для архитектурных решений, однако не подменяют проверку на данных заказчика, согласование SLA и правовую экспертизу применительно к конкретному контуру.

Тренды архитектуры:

- Замещение «ванильного» RAG — переход к композитным архитектурам со специализированными индексами
- Графовая память (графы знаний, temporal graphs, hypergraphs)
- Мультимодальные VLM вытесняют традиционные OCR-пайплайны
- **MoE (Mixture of Experts):** промышленный стандарт (GigaChat 3.1, Qwen3); активация только необходимых модулей снижает нагрузку на GPU
- **VLA-модели (Vision-Language-Action):** переход от чат-сценариев к агентам, планирующим действия на основе визуального контекста
- DSPy + GEPA для автоматической оптимизации промптов

Тренды оптимизации стоимости:

- BitNet и 1-битный инференс для продакшна на CPU
- Doc-to-LoRA устраняет накладные расходы KV-кэша
- LEANN для сокращения объёма индекса (*«LEANN: компактный векторный индекс»*)
- REFRAG для снижения задержек (latency) до 30×

Тренды инфраструктуры:

- CLI как альтернатива MCP для части агентных задач с минимальным протокольным перерасходом памяти — см. *«CLI vs MCP для корпоративных систем»*
- Фреймворк EffGen для повышения эффективности SLM (+11,2% для моделей класса 1.5B)
- Устойчивый спрос на **сквозную наблюдаемость GenAI** (трассировка, токенизация, оценка качества в продакшне) в **self-hosted** и резидентных контурах — *Приложение C, параграф «Рынок РФ, наблюдаемость LLM и референс-стек Comindware»*
- Защитные механизмы как обязательный инфраструктурный слой

Значения CapEx/OpEx и SLA по данным сигналам подтверждаются через дерево факторов стоимости, сценарный сайзинг и российские тарифы в отчёте *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Агенты для программирования (англ. **coding agents**; Cursor, OpenCode, OpenWork, OpenRouter) — рыночный контекст для оценки OpEx разработки на стороне заказчика.

Методика и SKU референс-стека — в параграфе *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Перечень инструментов — в параграфе *«Агенты для программирования и IDE»*.

9.3.6 Агенты для программирования и IDE

Типовой порядок оценки SKU в РФ (ориентир): SourceCraft (Yandex) → GigaCode (Sber) → Koda → OpenCode Enterprise (on-prem).

Российские решения (российские разработчики):

Решение	Разработчик	Цена	Примечания
SourceCraft	Yandex Cloud	700 руб./пользователь/мес	Agent mode, code review
GigaCode 2.0 + GigaIDE	SberTech	Бесплатно (ограничено)	IDE + AI
Koda	Koda startup	250–1 590 руб./мес	Работает без VPN, 152-ФЗ

Глобальные решения (конкуренты для сравнения):

Решение	Цена (5–10 чел.)	РФ без VPN	Примечания
OpenCode Go	4250–8500 руб./мес	✅ Да	Open-source модели
Trae (ByteDance)	~1700 руб./мес	✅ Да	КНР
Windsurf Teams	17 000–34 000 руб./мес	❌ Возможно	Данные разнятся
Cursor Teams	17 000–34 000 руб./мес	❌ Нет	Требует VPN
GitHub Copilot	8075–16 150 руб./мес	❌ Нет	API заблокирован

Полный on-prem (без облака):

Решение	Цена	Примечания
OpenCode Enterprise	Индивидуально	Локальные модели (Ollama, vLLM)
Tabnine Enterprise	3300+ руб./пользователь/мес	Полный on-prem
Sourcegraph Cody	5000+ руб./пользователь/мес	Enterprise-поиск + AI

Асинхронные каналы для агентов для программирования: интеграция агентных сред с Telegram, Discord и другими внешними точками входа показывает смещение от «IDE-ассистента» к распределённому рабочему процессу. Для корпоративного контура это повышает требования к разграничению доступа, журналированию действий и контролю внешних интеграций.

Конвергенция в сторону «суперприложений» разработки: крупные вендоры стремятся объединить чат, кодовый ассистент, автономные действия на рабочем месте и маршрутизацию моделей в одном интерфейсе. Для закупки это означает, что сравнивать нужно не только цену подписки, но и глубину интеграции, режим обработки кода, переносимость артефактов и возможность замены зарубежного сервиса на локальный контур.

Глобальные шлюзы разработки: OpenRouter и OpenCode Zen

OpenRouter и OpenCode Zen демонстрируют зрелость **зарубежных** шлюзов разработки применительно к прототипированию, IDE-ассистентам и **агентам для программирования**.

Для контура РФ их роль — прежде всего ускорители разработки, но не базовый маршрут промышленного развёртывания при наличии ПДн.

Валютный биллинг, зарубежные обработчики и политики журналирования требуют самостоятельной юридической и ИБ-оценки.

Экономика использования — в *«Сайзинг и экономика (CapEx / OpEx / TCO)»*; ограничения для продакшна — в *«Безопасность, комплаенс и наблюдаемость»* и *«Отчуждение ИС и кода»*.

9.3.7 Инфраструктура ИИ

В ходе пилотов и тендерных процедур стейкхолдеры систематически сопоставляют **актуальные открытые модели**. Для контуров в **РФ** (с учётом 152-ФЗ) принципиальное значение имеют **локальные API**, веса из доступных реестров (**GitVerse**, Hugging Face) и резидентный хостинг (**Cloud.ru**, **Yandex Cloud**).

Верификация санкционных рисков, цепочек поставок аппаратного обеспечения и условий поддержки вендора обязательна до фиксации поставщика в договоре.

Финансовые модели и тарифы: см. *«Сайзинг и экономика (CapEx / OpEx / TCO)»*. Ссылки на источники — в конце документа.

Актуальные открытые веса (ориентир март 2026)

- **GigaChat 3.1 (Сбер): MIT**, **MoE** (**Ultra** / **Lightning**), веса на **Hugging Face** и **GitVerse** — основа для **суверенного on-prem** контура и доработок под 152-ФЗ; нюансы стека vLLM/SGLang — см. *«Российские модели GigaChat (Сбер)»*.
- **MiniMax M2.7:** присутствует в каталогах **российских** облачных провайдеров как конкурирующий **API**; тарифы и политики журналирования целесообразно сравнивать с GigaChat / YandexGPT / Qwen (см. *«Сайзинг и экономика (CapEx / OpEx / TCO)»*).
- **Qwen3 / Qwen3.5 (Alibaba): открытые веса**, линейки **dense и MoE** (включая крупные модели с малым числом **активных** параметров на шаг). Популярная база для **self-hosted** инсталляций и **управляемых API** провайдеров (в РФ — Cloud.ru и аналоги).
- **GLM-5 / GLM-5.1 (Z.ai): GLM-5** — доступная флагманская линейка; **GLM-5.1** — анонсированы **планы** по открытию весов (сроки зависят от релизов Z.ai).
- **NVIDIA Nemotron 3: Super** — **120B** параметров (**12B** активных), **гибрид Mamba и Transformer** (MoE), контекст около **1 млн** токенов, фокус на агентные сценарии; квантизация **FP8** / **NVFP4**. **Nano** — компактные сборки. Доступны по лицензии Nemotron для on-prem при наличии совместимых GPU и легальной поставки.
- **OpenAI gpt-oss:** семейство открытых моделей (**20B** / **120B**, MoE, **Apache 2.0**); не путать с проприетарным ChatGPT. В РФ рассматривается преимущественно в **on-prem** сценариях при наличии вычислительных мощностей и прохождении комплаенс-контроля.

Акселераторы: Huawei Atlas 350 / Ascend 950PR

- Запрос на альтернативы GPU от NVIDIA с целью оптимизации **TCO** и диверсификации цепочек поставок — устойчивый паттерн при проектировании гибридных контуров.

- **Ascend 950PR**: по заявлениям прессы и вендора, производительность достигает **~2,8×** относительно **NVIDIA H20** в ряде сценариев; поддержка **FP4**, до **112 ГБ HBM**.
- Маркетинговые заявления (в частности, «до 70B параметров на карту») подлежат обязательному нагрузочному тестированию до фиксации SLA в контракте.

Китайские альтернативные GPU: Huawei Ascend 910C (~80% H100, $12–18 тыс.), Moore Threads Huashan (прогноз 2026), Cambricon Siyuan 590 (192 GB), MetaX, цепочки поставок для России/CIS и рекомендации по внедрению, см. *Приложение E «Китайские альтернативные GPU для инференса»*.

Инференс: архитектура Mamba и гибриды

Mamba — архитектура на базе **селективных моделей скрытого состояния (SSM)**. В отличие от механизма **self-attention**, где вычислительная стоимость растёт **квадратично** в зависимости от длины контекста, SSM использует компактное рекуррентное состояние. Это позволяет **снизить стоимость инференса** на длинных текстах при сохранении качества.

- **Nemotron 3 Super** уже использует гибридную архитектуру (**Mamba + Transformer**); параллельно развивается экосистема **Mamba-3** (Together и др.) с режимами инференса **SISO/MIMO**, предлагающая компромисс между суммарной латентностью (prefill+decode) и точностью.
- Обсуждение архитектуры актуально при запросах заказчика на **снижение задержек (latency)** и обработку сверхдлинных контекстов как альтернативу чистым трансформерам.

9.3.8 Корпоративные ИИ-сервисы

Раздел охватывает **управляемые** и **экосистемные** сервисы для корпоративных контуров.

Приоритет: **резидентность и российские вендоры** либо **open-source** с развёртыванием в РФ без обходных решений.

Глобальный SaaS — в заключительной части как ориентир по функциональным возможностям и рискам для данных, но не как дефолтный выбор для российского контура.

Нативные решения РФ (резидентный контур)

GigaChat API и корпоративная линейка (Сбер)

- **Концепция:** встроить генеративные модели в продукты и процессы заказчика при опоре на **российскую** инфраструктуру и русскоязычное качество; снизить порог для пилотов и промышленных интеграций без собственного ЦОД для инференса.
- **Решение:** коммерческий **GigaChat API** (*«продукт и тарифы для бизнеса»*, *«обзор API»*); линейка моделей **Lite / Pro / MAX** с разным балансом скорости, стоимости и глубины (*«документация по моделям и тарифам»*); заявление вендора — хранение данных **на серверах в РФ**, по умолчанию **не используются** запросы и ответы для дообучения (*«GigaChat API»*, блок вопросов и ответов). **GigaChat Enterprise** («ГигаЧат Бизнес») — анонс корпоративной платформы для ИИ-агентов (облако / гибрид / on-prem) по сообщениям СМИ, март 2026 (*«МК: GigaChat Enterprise»*); параметры поставки — по актуальному коммерческому предложению Сбера.
- **Результат:** для заказчика — готовый путь «API → интеграция → масштабирование» в резидентном контуре; для интегратора Comindware — опора на документированный API и SDK (**GigaChain**, *«(GitHub)»*). Архитектура **открытых весов 3.1 (MoE)**, **vLLM/SGLang**, SKU управляемого API и **скрытый OpEx** сопровождения — *«Российские модели GigaChat (Сбер)»*. Голосовой слой в экосистеме Сбера закрывает **SaluteSpeech**, а не **GigaChat**: это отдельный продукт и отдельная коммерческая модель.

MWS GPT Model Hub (MWS Cloud)

- **Концепция:** быстро подключить **открытые LLM** в **резидентном** облаке без развёртывания собственного инференс-стека на пилоте.
- **Решение:** сервис **MWS GPT Model Hub** — доступ к **10 открытым LLM** (в т.ч. DeepSeek, Google, Alibaba) через **OpenAI-совместимый API** внутри **MWS Cloud**; подключение за минуты, развёртывание «в один клик», тарификация по токенам (*«GPT Model Hub»*).
- **Результат:** сигнал зрелости **LLM как облачной услуги** в контуре РФ; для суверенных внедрений — учитывать лицензионные условия ПО «MWS GPT» и режимы облака в части 152-ФЗ.

MWS Octapi

- **Концепция:** ускорить подключение ИИ-агентов к корпоративным системам без отдельного слоя «самодельных» интеграций.
- **Решение: MWS Octapi** — интеграционная платформа из реестра российского ПО с **MCP**-поддержкой для связи агентов с корпоративным ландшафтом; в публичных материалах MWS платформа описана как low-code слой для безопасного подключения сервисов и построения мультиагентных взаимодействий (*«MWS Octapi»*, *«MWS: +30% к скорости создания ИИ-агентов»*).

- **Результат:** для интегратора это более зрелый сценарий встраивания в корпоративный ландшафт, чем прямое связывание LLM с разрозненными API; MWS заявляет ускорение создания мультиагентных систем **на 30%**, что трактуем как ориентир платформенной зрелости, а не как норму КП без замеров на контуре заказчика.

Yandex Agent Atelier

- **Концепция:** собрать корпоративного агента не как отдельный поисковый инструмент, а как управляемую платформу с данными, логикой и интеграциями в одном контуре.
- **Решение: Agent Atelier** в **Yandex AI Studio** объединяет **Agent Tools**, **MCP Hub** и **Workflows**; платформа поддерживает поиск по файлам и в интернете, интерпретатор кода, подключение MCP-серверов, low-code и API-режимы, а также коннекторы к российским системам — от **amoCRM** и **Контур.Фокус** до **Яндекс Трекера** и **Битрикс24** (*«Agent Atelier»*). Для поискового слоя отдельно доступен **File Search** с гибридным поиском по документам и мультимодальным файлам (*«создание агента с File Search»*).
- **Результат:** для заказчика это один из наиболее зрелых российских сценариев «агент + знания + интеграции»; Яндекс отдельно указывает, что токены внутри инструментов могут быть **в 4 раза дешевле**, что делает сложные агентные цепочки экономически ближе к массовому пилоту, а не только к R&D.

GitVerse + GigaCode (Сбер)

- **Концепция:** перевести ИИ из «подсветки кода» в **агентный** участник **SDLC** в российской экосистеме разработки.
- **Решение:** официальные материалы **GitVerse** описывают **GigaCode-чат** как AI-ассистента, который умеет создавать и изменять файлы, создавать коммиты и запросы на слияние, а **GigaCode-агент** подключается к merge request как участник ревью с командами `/describe`, `/review`, `/improve`, `/ask` (*«GigaCode-чат»*, *«GigaCode-агент»*).
- **Результат:** зрелость агентных паттернов в SDLC; для заказчика критичны ИБ (доступ к репозиториям, журналирование, контроль действий агента) и модель отчуждения — см. *Приложение А «Отчуждение ИС и кода»*.

Open-source в РФ (развёртывание без VPN)

Ouroboros (Cloud.ru)

- **Концепция:** экспериментальный и пилотный контур **автономного агента** с самомодификацией кода и устойчивым контекстом — в **резидентной** инфраструктуре маркетплейса.

- **Решение:** открытый проект *«Ouroboros»* (самонастраиваемый аналог OpenClaw и Hermes); развёртывание через витрину Cloud.ru и контейнерный вариант; **автономная эволюция** (перепись собственного кода, исправление ошибок, добавление функций без постоянного участия пользователя); **фоновая активность** между задачами; **персистентность** идентичности, задач и решений после перезагрузки; **многомодельное рецензирование** (одна сущность проверяет работу другой); связывание разрозненных сигналов (календарь, файлы, погода и др.) в неочевидные рекомендации.
- **Результат:** сильный **сигнал для R&D и пилотов** по агентности; для продакшна — отдельная проработка ИБ, наблюдаемости и границ автономии.

OpenClaw как ориентир агентного стека

- **Концепция:** самостоятельно размещаемый агентный стек как ориентир для автономных сценариев разработки и многоканальной доставки.
- **Решение:** OpenClaw полезен не как канонический выбор для контура РФ, а как внешний сигнал зрелости по слоям исполняемой среды, вызова инструментов и маршрутизации моделей; в российских пилотах его чаще рассматривают как инженерный ориентир, а не как прямой стандарт закупки.
- **Результат:** для стратегии это маркер темпа рынка агентных ИИ-систем; для промышленного внедрения ключевой вопрос остаётся тем же: изоляция среды, аудит действий агента и границы допустимой автономии. Контур риска и публичные сигналы экспозиции сохранены в Приложении C.

Другие стеки для агентной автоматизации

Ниже перечислены не просто «ещё одни инструменты для ИИ-разработки», а смежный класс решений для того же типа задач, который решает **OpenClaw**: автономное выполнение инженерных задач, запуск подагентов, работа с инструментами и файлами, управляемые разрешения, удалённые воркеры и, в отдельных случаях, многоканальная доставка через веб-интерфейсы или мессенджеры.

- **Claude Code / Agent SDK:** применяют там, где нужны подагенты, параллельная декомпозиция, изоляция контекста и ограничение набора инструментов по ролям; это особенно полезно для сценариев ревью кода, тестовых прогонов и исследовательских подзадач в одном инженерном цикле (Anthropic — Subagents in the SDK).
- **OpenCode / OpenWork:** используют как более открытый и программируемый контур для тех же задач: настраиваемые основные и вспомогательные агенты, политика разрешений, SDK/API для сессий и событий, а в случае **OpenWork** — ещё и настольная, мобильная и серверная оболочка с удалёнными воркерами

и мессенджер-коннекторами поверх **OpenCode** (OpenCode — SDK, OpenCode — Agents, OpenWork — AGENTS.md, OpenWork — ARCHITECTURE.md).

- **OpenHands / Continue / Cline / Roo Code / aider:** рынок уже разошёлся по нескольким классам решений для схожих задач. **OpenHands** даёт платформенный SDK и переход от локального режима к изолированному или серверному развёртыванию; **Continue** — агентный режим в IDE с MCP и локально размещаемыми моделями; **Cline** — модель управления доступом с приоритетом явного одобрения и режимом YOLO; **Roo Code** — многорежимную оркестрацию; **aider** — git-native терминальный поток с авто-коммитами и быстрым откатом (OpenHands SDK Overview, Continue — How to Customize Agent Mode, Cline — Auto Approve & YOLO Mode, Roo Code — Using Modes, aider — Git integration).
- **Практический вывод для пакета:** сравнивать такие решения полезнее не по бренду, а по пяти осям: изоляция исполнения, политика разрешений, расширяемость (skills / MCP / plugins), поверхность взаимодействия (IDE / CLI / web / messaging) и готовность к резидентному или полностью локальному развёртыванию в контуре заказчика.

Голосовой слой для корпоративных ассистентов

Для voice-first сценариев в РФ сформирована зрелая линейка **локальных TTS/ASR-стеков** без зависимости от зарубежных сервисов. Это актуально не только для IVR и чат-ботов, но и для **речевой аналитики клиентских разговоров**, голосовых ассистентов в продажах, транскрибации встреч и контроля качества сервиса.

По публичным тарифам российские решения демонстрируют особенно сильные позиции в сегменте **TTS**; по **ASR** разрыв с глобальными ориентирами сократился, а выбор в большей степени определяется требованиями к задержке, языковому охвату, режиму размещения и обработке персональных данных.

Сравнение публичных ориентиров по голосовому слою

Сервис	Контур	TTS	ASR	Комментарий
Yandex SpeechKit / AI Speech	Cloud / Hybrid	**1342 руб. / 1 млн символов** (API v1)	**0,16266666 руб. / 15 с** (**~39 руб./час**)	**Brand Voice Lite:** 9150 руб. за создание; 101 666 руб./мес хостинг первого голоса
SaluteSpeech (Сбер)	API / On-prem	**0,000186 руб./символ** (**~186 руб. / 1 млн символов**)	**0,01 руб./с** (**~36 руб./час**)	Для юрлиц отдельно описаны API- и on-prem-сценарии
ElevenLabs	Global SaaS	**$0,06-0,12 / 1K символов**	**$0,22 / час**; realtime **$0,39 / час**	Полезный глобальный бенчмарк; сравнение в рублях — по *«Курсу USD для смет»*

Это публичные ориентиры для рыночного сравнения, а не каноническая строка сметы. Для бюджетирования и TCO используйте отчёт «Сайзинг и экономика (CapEx / OpEx / TCO)».

Yandex SpeechKit / AI Speech

- **Концепция:** добавить в корпоративного ассистента голосовой интерфейс для IVR, голосовых ботов, озвучки уведомлений и мультимодального UX без отдельного зарубежного TTS-стека по умолчанию.
- **Решение: Yandex SpeechKit / AI Speech** используется для синтеза и распознавания речи в голосовых интерфейсах, внутренних ассистентах и контактных сценариях; в публичных материалах Яндекса отдельно выделяются корпоративные сценарии, **Brand Voice** и **SpeechKit Hybrid** для размещения в своём контуре (*«AI Speech / SpeechKit Hybrid»*, *«тарифы SpeechKit»*).
- **Результат:** для руководителей это логичное расширение текстового ассистента до голосового канала в российском облачном или гибридном контуре. Особенно уместно там, где один и тот же контур должен закрывать поиск по знаниям, голосовой интерфейс и обработку разговоров с клиентами.

SaluteSpeech (Сбер)

- **Концепция:** закрыть голосовой слой корпоративного ассистента в резидентном стеке Сбера — от IVR и голосовых рассылок до транскрибации и voice UX.
- **Решение: SaluteSpeech** объединяет **TTS** и **ASR**, поддерживает **SSML**, нормализацию текста, несколько языков, анализ эмоций и on-prem сценарии;

продуктовая страница описывает применение в **автоматизации телефонии**, **голосовых рассылках**, **контакт-центрах**, **транскрибации** и **собственных голосовых помощниках** (*«SaluteSpeech»*, *«тарифы для юрлиц»*).

- **Результат:** сильный российский кандидат для voice-first ассистентов и контакт-центров; в отличие от чисто текстового API, позволяет обсуждать единый стек «чат + голос + аналитика качества сервиса» в одном коммерческом контуре.

Глобальный SaaS (технологический фон, не суверенный дефолт)

Google AI Studio — Vibe Coding

- **Концепция:** ускорить прототипирование full-stack приложений с опорой на экосистему Google.
- **Решение: Antigravity Agent** для развёртывания **Firebase**; поддержка **Next.js, React, Angular**; **Gemini 3.1 Pro** для полного цикла разработки в среде AI Studio.
- **Результат:** ориентир по скорости вывода MVP **вне резидентного дефолта РФ**; персональные и корпоративные данные, трансграничная передача и политика журналирования — отдельная правовая и архитектурная оценка перед использованием в контуре заказчика.

Adobe Firefly

- **Концепция:** кастомизация генеративного дизайна и креатива на данных заказчика в экосистеме Adobe.
- **Решение: Custom AI models** на пользовательских данных; **Project Moonlight** — агентный интерфейс к приложениям Adobe.
- **Результат:** полезный **бенчмарк** для креативных и маркетинговых подразделений; для РФ-контуров с 152-ФЗ — не замена суверенным стекам без явного DPA и локализации.

ElevenLabs

- **Концепция:** использовать внешний voice layer для корпоративных чат- и voice-ассистентов, когда нужен широкий каталог голосов, мультиязычность или отдельный фокус на качественном TTS/voice UX.
- **Решение: ElevenLabs** уместно сравнивать не как «музыкальный маркетплейс», а как глобальный слой синтеза голоса и voice-интерфейсов поверх корпоративных ассистентов.
- **Результат:** полезный бенчмарк по качеству voice UX и международному охвату, но не суверенный дефолт для РФ-контура; его стоит сравнивать с российскими стеком прежде всего в сценариях мультиязычности и premium voice UX.

9.3.9 Российский рынок

Общая картина рынка GenAI (red_mad_robot)

Ключевые метрики:

Метрика	Значение	Контекст
B2B-адаптация GenAI	**71%** (крупные компании)	Рост +17 п. п. год к году
Кадровый дефицит	**~10 тыс.**	M&A + обучение
Вакансии с ИИ-навыками	**+89%** (2025), **+170%** (Q1 2026)	По данным hh.ru
Компании с open-source	**86%**	Дообучение моделей
ROI on-prem	**3+ года**	Cloud для пилотов, ЦОД для продакшна

Онтология рынка:

- Ядро: данные (LabelMe), модели (GigaChat/YandexGPT), инструменты.
- Инфраструктура: GPU/облака (3Logic/DataRu); китайские альтернативы.
- Услуги: AI-TRISM/консалтинг.

Показатели рынка на март 2026:

- Цель **влияния ИИ на ВВП** к 2030: **~11 трлн руб.** (Национальная стратегия, Указ №124). Этот **показатель отличается** от оценки **экономического эффекта** из отраслевых исследований (например, вилка Yakov Partners) — см. параграф *«Экономический эффект»* отчёта «Методология разработки и внедрения ИИ».
- Объём рынка ИИ в России: **~170 млрд руб.** (2025), рост +45% год к году
- Рынок GenAI: **58 млрд руб.** (2025), рост ~400% год к году, прогноз 2030 — **778 млрд руб.**
- Прогноз рынка ИИ на 2026: **~250 млрд руб.**, на 2027: **~360 млрд руб.** (CAGR ~45%)
- ИИ-агенты: **46%** компаний уже внедрили или тестируют

Источники: Указ Президента РФ №124; РБК: объём рынка ИИ; Ведомости: ИИ-рывок 2026; Коммерсант: рынок GenAI; CNews: GenAI 2025.

Дефицит кадров и рыночное давление (Talent War)

> **Бизнес-риск: кадровый тупик**
>
> Дефицит в **~10 000** профильных специалистов (RMR) и агрессивный хантинг со стороны Бигтеха делают стратегию «нанять свою команду R&D» для большинства компаний невыполнимой или экономически неоправданной.

Индикаторы перегрева рынка:

- **Монополизация талантов:** Сбер и Яндекс контролируют вход в профессию через олимпиады (7 000+ участников) и профильные кафедры в ведущих вузах (МФТИ, ИТМО).
- **Скорость найма:** практика «One Day Offer» (Сбер, март 2026) фиксирует критическую конкуренцию за каждого специалиста.
- **Инфляция зарплат:** премия за AI-компетенции достигает **+20%** к рыночному уровню на фоне и без того высоких базовых ставок (Известия/hh.ru, 2025).
- **Сложность воспроизводства экспертизы:** «Agents Week» от ШАДа (апрель 2026) фиксирует переход к элитным инженерным навыкам (scaling, evaluation), быстрая передача которых штатным ИТ-специалистам практически невозможна.

Вывод для стратегии: Comindware позволяет привлечь необходимую экспертизу и суверенный стек без выхода на перегретый рынок труда.

GenAI в маркетинге (CMO Club × RMR, 2025—2026)

Исследование:

red_mad_robot совместно с CMO Club Russia провели опрос директоров по маркетингу крупнейших российских брендов о влиянии GenAI на рабочие процессы.

Для сайзинга и TCO это **не** прямая строка в калькуляторе GPU и токенов, а **сигнал спроса** со стороны владельцев маркетингового бюджета и показатель зрелости внедрения вне ИТ-функции.

Доли/формулировки: согласовано с *«Telegram — CMO Club Russia #197»* и перекрёстно сверено с публикацией *«RB.RU — 93% команд в маркетинге используют ИИ…»*. Для договорных формулировок и внешних заявлений сверяйте полный текст и официальную публикацию исследования.

- **Охват и режим использования: 93%** компаний используют GenAI в рабочих процессах маркетинга; системно интегрировали технологию **около трети** респондентов.
- **Интенсивность: 41%** специалистов пользуются GenAI **ежедневно**; в материалах к исследованию это сопоставляют с глобальным ориентиром **10–20%** ежедневного использования.
- **Инструменты: 91%** CMO называют **ChatGPT** основным рабочим инструментом; **59%** используют **Midjourney**; разрыв **32 п.п.** между ними отражает концентрацию на универсальных чат-моделях.
- **Бюджет: 64%** компаний выделяют на инициативы с GenAI **только 1–5%** маркетингового бюджета.
- **Барьеры (качество и безопасность): 53%** отмечают необходимость постоянной доработки контента; **49%** — шаблонность; **40–50%** — галлюцинации и ошибки; **45–60%** — риски утечки данных (перекрёстно сверено с глобальными данными Stanford: **+56,4%** инцидентов год к году, Cisco: 60% обеспокоены безопасностью).
- **Стратегия и зрелость:** ключевой барьер — **отсутствие плана и стратегии**; в публикации приводится контраст оценок «экспериментаторов» (**6,9** из 10) и «интеграторов» (**2,5** из 10) по значимости этого фактора.
- **Зрелость по направлениям и гео-разрыв:** сильнее всего GenAI встроен в **контент-маркетинг**; **ниже** вовлечённость в **event-маркетинг**, **управление брендом** и **бюджетами**.
- **Эффекты (восприятие vs измерение): 77%** отмечают рост скорости и качества контента; **73%** — ускорение процессов; **50%** — рост продуктивности без расширения штата; **40–66%** ожидают снижения нагрузки.

> ⚠ **Разрыв между ожиданиями и результатами**
>
> Глобальные исследования показывают значительный разрыв между воспринимаемой и измеренной продуктивностью:
>
> - **Воспринимаемая продуктивность:** 40–66% (опросы)
> - **Измеренная продуктивность:** <1% (OECD), а по исследованию METR (июль 2025) — **-19%** для разработчиков, использующих ИИ
> - Этот разрыв между ожиданиями и результатом характерен для ранних стадий внедрения

- **Персональная и корпоративная зрелость:** растёт разрыв между личной цифровой зрелостью специалистов и ограничениями корпоративной инфраструктуры и регламентов.

- **Взгляд вперёд: 85%** российских CMO считают GenAI ключевым фактором трансформации на горизонте **трёх лет**; в нарративах исследования фигурирует сдвиг роли маркетолога к **оператору-оркестратору** (данные, технологии, креатив) и масштабирование ассистентов и агентов по функциям маркетинга.

9.4 Локальный инференс: практические кейсы

Классы решений и влияние на TCO: выбор CLI или тяжёлого протокола инструментов, квантизация и «раздувание» модели относительно VRAM задают **диапазон** CapEx/OpEx, но не заменяют расчёт под профиль **корпоративный RAG-контур** / **агентный слой Comindware Platform** и выбранный инференс-слой (MOSEC/vLLM/SGLang).

9.4.1 CLI vs MCP для корпоративных систем

Сигнал: протокольный оверхед и выбор между гибкостью и стабильностью. MCP удобен для личных агентов, но для корпоративных систем с требованиями к наблюдаемости, безопасности и экономии ресурсов CLI остаётся более предсказуемым выбором.

Источник: @llm_under_hood

Подход	Применение	Преимущества
MCP	Личные агенты	Простота, нативность
CLI/терминал	Корпоративные системы	Стабильность, экономия

Пример: libghostty для запуска агентов на серверах через CLI вместо MCP.

9.4.2 Инструменты дообучения

Сигнал: дмократизация fine-tuning. No-code-инструменты снижают порог входа в дообучение открытых моделей, что ускоряет цикл экспериментов и снижает затраты на GPU-ресурсы.

Unsloth Studio — no-code-платформа для дообучения и инференса LLM. Входит в экосистему инструментов, которые сокращают порог входа в fine-tuning и ускоряют итерации при работе с открытыми моделями.

Возможности:

- No-code-веб-интерфейс для LLM
- Подготовка данных, обучение, инференс, экспорт

- Кастомные Triton-ядра с собственной реализацией обратного распространения ошибок

Преимущество: быстрее стандартных CUDA-реализаций.

9.5 Рынок ИИ: глобальная статистика

Приведённые ниже глобальные метрики служат внешним фоном для оценки спроса, бюджетов и зрелости рынка; применительно к офферу в РФ они не подменяют локальные ограничения по данным, доступности сервисов и экономике.

9.5.1 Глобальные метрики внедрения (McKinsey, Deloitte, Menlo 2025–2026)

Метрика	Значение	Источник
Глобальное принятие ИИ	**88%**	McKinsey 2025
Масштабирование за пределы пилотов	**~33%**	McKinsey 2025
Высокая производительность (EBIT >5%)	**6%**	McKinsey 2025
Рост инцидентов ИИ (год к году)	**+56,4%**	Stanford AI Index 2025
GenAI-бюджет enterprises 2025	**$37 млрд**	Menlo Ventures
Рост GenAI-бюджета (год к году)	**3,2x**	Menlo Ventures

9.5.2 Распределение рынка приложений

Источник: a16z Top 100 AI Apps

Модель	Веб-трафик	Доля
ChatGPT	Базовый	100%
Gemini	0.37x ChatGPT	37%
Claude	0.036x ChatGPT	3,6%

Тренд: Gemini, Grok, Claude набирают долю у платных подписчиков.

9.5.3 География использования ИИ

Страна	Ранг
Сингапур	1
ОАЭ	2
Гонконг	3
Южная Корея	4
США	20

> **Примечание**
>
> США создали большинство AI-продуктов, но по уровню использования находятся примерно на 20-м месте.

9.5.4 Структурные изменения рынка

- **Три мира:** запад, Китай, РФ (из-за политики)
- **Китайская модель внедрения:** публичные материалы о массовом использовании ИИ в экономике — полезный контекст при обсуждении «трёх миров». См. «*AI + Economy: китайская модель*».
- **text2img умирает:** Midjourney упал с ТОП-10 до 46-го места
- **text2video сжался:** консолидация рынка
- **Аудио стабильно:** Suno, ElevenLabs сохранили позиции
- **Браузеры:** Atlas, Comet, Claude в Chrome пока не взлетели

9.6 Планирование мощности ИИ-инфраструктуры (2025-2030)

9.6.1 Прогноз McKinsey

Прогноз спроса:

- Инновации могут снизить потребность в GPU на 50% к 2030

Технологии, влияющие на спрос:

- 3x плотность вычислений — физическое сокращение
- Периферийный инференс (edge) — децентрализация нагрузки
- Квантование и дистилляция — снижение требований к VRAM

9.6.2 Слои ИИ-инфраструктуры

Слой	Компоненты	Маржа
Слой 0: фабрики + память	TSMC N4/N3, SK Hynix HBM3e	Высокая (узкие места)
Слой 1: чипы	NVIDIA H100/H200/Blackwell, AMD MI300X	Высокая (близкая к монополии)
Слой 2: серверы	DGX H100, HGX B100, OEM-сборки	Средняя
Слой 3: оркестрация	Kubernetes, Ray, SLURM	Низкая
Слой 4: облако	AWS Bedrock, Azure AI, GCP Vertex	Высокая (наценка на GPU)
Слой 5: модели	GPT-4o, Gemini, LLaMA, Claude	Средняя
Слой 6: приложения	ChatGPT, Copilot, Claude	Переменная

Ключевой инсайт: максимальная маржа сосредоточена на слое 0 (TSMC, SK Hynix) и слое 1 (NVIDIA).

9.6.3 Капитальные затраты крупных техкомпаний (2025)

Сводка по капитальным затратам крупных техкомпаний — см. основной отчёт по сайзингу, раздел «Глобальный рынок ИИ-инфраструктуры».

ROI: разрыв ожиданий и реальности:

- 80–95% ИИ-проектов не достигают целевого ROI
- Значимый ROI фиксируют лишь 10% компаний
- 42% инициатив свёрнуто в 2025 году

9.6.4 Порог утилизации: on-prem и облако

Правило 40–60%:

- При загрузке ниже 40%: облачная модель экономически предпочтительна
- При загрузке выше 60–70%: собственная инфраструктура обеспечивает преимущество по TCO

Расчётные формулы:

- On-prem: `TCO = CapEx + (OpEx × Годы) + (Энергия × PUE × Годы × Часы) + Персонал`
- Облако: `TCO = Почасовая_ставка × 24 × 365 × Годы + Плата_за_исходящий_трафик + Плата_за_хранение`

⚠ Разрыв продуктивности: восприятие vs реальность

Исследования 2025–2026 выявили значительный разрыв между ожидаемой и измеренной продуктивностью:

Метрика	Воспринимаемая	Измеренная	Источник
Скорость разработки	**+20–24%**	**-19%**	METR (июль 2025)
Продуктивность (опросы)	40–66%	<1%	OECD/Deloitte

Вывод: ожидания от ИИ значительно превышают реальный эффект на ранних стадиях внедрения.

9.6.5 TCO-калькулятор (5 лет)

Формула TCO (on-prem):

```
TCO = CapEx + (OpEx × Годы) + (Энергия × PUE × Годы × Часы) + Персонал
```

Формула TCO (облако):

```
TCO = Почасовая_ставка × 24 × 365 × Годы + Плата_за_исходящий_трафик + Плата_за_хранение
```

9.7 Практический опыт внедрения ИИ: верификация результата

Публичные комментарии и материалы **red_mad_robot** полезны как ориентир по организации инженерного контура и проверке эффекта внедрения (источник).

9.7.1 Подход к ИИ-коду в бизнесе

По данным CTO AI red_mad_robot Влада Шевченко источник:

Вместо построчной проверки компании всё чаще переходят к системе верификации с автоматическими тестами, метриками, регрессионными проверками и оценкой качества. Акцент смещается с контроля действий на контроль результата.»

Ключевые принципы:

- Верификация результата, а не контроль действий
- Автоматические тесты и метрики качества
- Регрессионные проверки
- Доверие к AI формируется за счёт среды, где ошибки быстро выявляются

9.7.2 Оптимизация рассуждений моделей

Google: Deep-Thinking Ratio (DTR) источник

- Метрика оценивает активность мышления на уровне внутренних слоёв
- Метод Think@n отбирает ответы с высоким DTR
- Снижение вычислительных затрат примерно в 2 раза

Oppo AI: Search More, Think Less (SMTL) источник

- Разбиение запроса на независимые подзадачи
- Параллельный сбор информации
- Сокращение шагов инференса на 70,7%

9.7.3 Память и контекст в ИИ-агентах

Databricks KARL источник

- ИИ-агент для корпоративного поиска с многошаговым рассуждением по закрытым корпоративным базам; бенчмарк **KARLBench**.
- В отчётах авторов — превосходство над указанными моделями при **≈33%** ниже стоимости и **≈47%** быстрее; **заявления бенчмарка KARLBench**, без подтверждения на задачах заказчика.

Accenture Memex(RL) источник

- Индексированная память; RL решает, когда разгрузить контекст, как озаглавить запись и когда её достать.
- В материалах авторов в **ALFWorld** успешность 24,2% → 85,6%, пик токенов контекста ~вдвое ниже.

Agent0 (Salesforce Research, Stanford) источник

- Ко-эволюция curriculum/executor; исследовательская архитектура.

General Agentic Memory (GAM) источник

- Memorizer/Researcher и исследовательский цикл над памятью.

9.7.4 Инфраструктура навыков ИИ

SkillNet (Alibaba, Ant, Tencent, Oppo) источник

- Трёхуровневая онтология: таксономия → граф связей → модульные наборы
- Средняя награда +40%, количество шагов -30%

9.7.5 Публичные R&D-практики

У red_mad_robot появился отдельный публичный поток материалов по reasoning-архитектурам, RAG-системам, агентным пайплайнам и LLM-инфраструктуре; для заказчика это полезно как внешний ориентир зрелости инженерной повестки, а не как норма стека.

9.7.6 Исследовательские сигналы марта 2026

- **OpenAI** — задачи контроля рассуждения при ограничениях на скрытые шаги источник.
- **Microsoft Research** — усиление безопасности агентных задач с внешними инструментами источник.
- **Princeton University** — взаимодействие пользователя с агентом как источник непрерывного обучения источник.
- **Meta (Экстремистская организация, запрещена в РФ), OpenAI и xAI** — непрерывное улучшение моделей для чатов источник.
- **Microsoft 365** — Copilot Cowork источник (тенант/ИС/лок-ин).

9.7.7 Инструменты и навыки для агентов

openapi-to-cli (ocli) источник

- Конвертация OpenAPI/Swagger в CLI команды на лету
- BM25-поиск по эндпоинтам за 7мс
- 100 MCP tools (~50K токенов) → 1 CLI tool + поиск

SGR Agent Core источник

- Schema-Guided Reasoning для агентов
- RunCommandTool (safe/unsafe режимы)

Навыки NeuralDeep для российских сервисов источник

- Локальная база агентных навыков для российских сервисов (аналог skills.sh)
- Уже загружены интеграции для Яндекс, Битрикс24, 1С и других популярных решений
- Установка одной командой (формат claude-skill)
- Опенсорсный проект: (GitHub) — можно добавить собственные скиллы
- Модерация и базовые проверки безопасности
- Потенциал: стандартный слой для агентных интеграций под российский рынок

9.7.8 События в индустрии (Март 2026)

Сводка моделей, инструментов и рыночных сигналов, релевантных для корпоративного RAG-контура и агентных сценариев.

Сводка по **MWS GPT Model Hub**, **Yandex AI Studio File Search**, **GitVerse + GigaCode**, линейке **GigaChat** для бизнеса, **Ouroboros** на Cloud.ru — в *«Корпоративные ИИ-сервисы»*.

9.7.9 Практики разработки с ИИ

В публичных инженерных разборах Артёма Лысенко рассматриваются практики разработки с ИИ; применимость к контуру заказчика оценивают с учётом ИБ и внутренних регламентов.

9.8 Практические кейсы внедрения

9.8.1 AGORA: Industrial AI и Enterprise

Кейс Норникель:

- Head of ML Данил Ивашечкин
- Стек: сигналы/SCADA → модели/LLM → агенты/оркестрация

9.8.2 AI & грабли: Agile-подход к ИИ-внедрению

Методология проверки гипотез:

1. Сокращение времени проверки: 1 идея → 2 недели
2. Порог чувствительности: 20% премии для мотивации
3. Психологическая безопасность: Google Project Aristotle

9.8.3 Российские модели GigaChat (Сбер)

Продуктовый контекст **API**, резидентности и корпоративных сценариев — *«Корпоративные ИИ-сервисы — GigaChat»*.

- **Open source:** по материалам Сбера на Хабре выпущены обновлённые **GigaChat-3.1-Ultra** и **GigaChat-3.1-Lightning** под лицензией **MIT**; веса и сопутствующие материалы — в коллекции на Hugging Face и в проекте на GitVerse. В статье описаны переход от dense-моделей к **MoE**, переработка постобучения, снижение **зацикливания** генерации, этап **DPO в нативном FP8** и обнаруженный **баг SGLang** при `dp > 1` (исправление — pull request в

SGLang); это влияет на выбор версий инференс-стека и на доверие к внутренним бенчмаркам.

- **Продукт GigaChat и открытые веса:** в том же источнике отдельно описываются данные и потребительские сценарии (поисковая выдача, цитирование, персонализация с **памятью о пользователе**). Self-hosted развёртывание — отдельная проработка архитектуры.
- **Карточки Hugging Face (идентификаторы и масштаб):** флагман **Ultra** — ai-sage/GigaChat3.1-702B-A36B: **702B** параметров всего, **36B** активных при инференсе, лицензия **MIT**; в карточке зафиксированы сценарии **кластера / крупного on-prem** и пример **многоузлового SGLang** (`nnodes`, `tp`, `ep`).
- **Lightning (3.1 vs 3.0 и API):** актуальная ветка **3.1** — ai-sage/GigaChat3.1-10B-A1.8B (**10B** / **1.8B** активных, MIT); линейка **3.0** — ai-sage/GigaChat3-10B-A1.8B. SKU **GigaChat3-10B-A1.8B** в тарифах **Cloud.ru** относится к **управляемому API** и может не совпадать один в один с версией чекпойнта на Hub; self-hosted исключает счётчик токенов провайдера, но добавляет **GPU и инженерию**.
- **Инференс на базе vLLM и совместимость:** для GigaChat3-10B-A1.8B в карточке указано `VLLM_USE_DEEP_GEMM=0` при работе с vLLM из-за конфликта с размерностью скрытого слоя; для GigaChat3.1-10B-A1.8B описаны **MTP** (`speculative-config` в vLLM) и для **вызова инструментов** — требования к минимальным коммитам vLLM и SGLang (см. карточку). Это включают в эксплуатационный регламент и учитывают как **скрытый OpEx** сопровождения и регрессий.

9.8.4 Перспективные технологии оптимизации инференса (2024–2026)

9.8.5 Кросс-платформенные техники оптимизации памяти

Платформа	Техника	Описание	Применимость
NVIDIA	Выгрузка KV-кэша	Выгрузка KV-кэша в CPU RAM (vLLM 0.11+)	Продакшн, длинный контекст
NVIDIA	FP4/FP8 квантизация	TensorRT-LLM, NVFP4 KV-кэш	H100/B200, 20× сжатие памяти
NVIDIA	Speculative Decoding	Малая модель + верификация	2–3× ускорение инференса
AMD	GPU-партиционирование	MI300X: SPX/DPX/CPX режимы	Мульти-тенант, 192 ГБ HBM
AMD	ROCm vLLM	Нативная поддержка с 2026	MI300X/MI355X, потребительская RX 7900
Apple Silicon	LLM in a Flash	Flash → DRAM по требованию	Edge, Mac Studio/Pro
Кросс-платформенные	GGUF/llama.cpp	CPU offloading, квантизация	Любое железо, низкий порог

NVIDIA: KV-кэш и квантизация

- **Выгрузка KV-кэша (vLLM 0.11+):** выгрузка KV-кэша в CPU RAM позволяет запускать модели с длинным контекстом при ограниченном VRAM (vLLM Blog)
- **NVFP4 KV-кэш (TensorRT-LLM):** 4-битный KV-кэш снижает память в 4× без потери качества; H100/B200 — до 10 000 токенов/с при длинном контексте и **крупных батчах** (NVIDIA Technical Blog)
- **Speculative Decoding:** малая «draft» модель генерирует кандидаты, основная верифицирует; 2–3× ускорение на H100

AMD: ROCm и GPU-партиционирование

- **MI300X (192 ГБ HBM):** архитектура CDNA 3, 8 XCD, 4 IOD; режимы партиционирования:
  - **SPX:** весь GPU как одно устройство (крупные модели)
  - **DPX:** 2 логических GPU по 96 ГБ (мульти-тенант)
  - **CPX:** 8 логических GPU по 24 ГБ (максимальная изоляция)

- **vLLM + ROCm:** нативная поддержка с января 2026; Sequence Parallelism для MI300X/MI355X (vLLM Blog)
- **Consumer GPU (RX 7900 XTX):** ROCm 7.2 официально поддерживает потребительские Radeon; производительность сопоставима с RTX 3090 при меньшей цене (XDA)
- **Важно:** модели требуют конвертации из CUDA в ROCm-формат; «CUDA-обёртки» (HIPify) не дают нативной производительности

Apple Silicon: LLM in a Flash

> **Область применения Apple Silicon**
>
> Техники Apple Silicon — **локальный инференс на Mac**; в облачных GPU не переносятся.

Техника «LLM in a Flash» (Apple ML Research, ACL 2024)

- **Суть:** хранение параметров модели во flash-памяти с подгрузкой в DRAM по требованию, что позволяет запускать модели до 2× размера доступной RAM
- **Оптимизации:** «windowing» для повторного использования нейронов, «row-column bundling» для последовательного доступа к flash
- **Ускорение:** 4-5× на CPU, 20-25× на GPU относительно наивной загрузки
- **Реализация 2025–2026:** проект `mlx-flash` — практическая реализация для MLX; 30B+ на 16GB Mac, 70B+ на 32GB+
- **Реальный кейс (2026):** CVS Health запустила Qwen 3.5 397B на MacBook Pro 48GB, используя ~5.5GB RAM

Кросс-платформенные: GGUF и llama.cpp

- **GGUF-квантизация:** INT4/Q4_K_M — стандарт для CPU-инференса; модели 70B запускаются на 40 ГБ RAM
- **GPU offloading:** частичная выгрузка слоёв на GPU при нехватке VRAM
- **Поддержка:** NVIDIA, AMD (Vulkan/ROCm), Apple Metal, CPU-only

Вывод для бюджетирования: продакшн в отраслевых обзорах чаще связывают с **NVIDIA (vLLM/TensorRT-LLM)** или **AMD MI300X (ROCm vLLM)** с учётом конвертации моделей; **Apple Silicon** и **GGUF/llama.cpp** — класс edge и R&D. Тарифы на GPU в РФ — в отчёте *«Сайзинг и экономика (CapEx / OpEx / TCO)»*.

Сравнение подходов к оптимизации инференса

Подход	Платформа	Применение
Выгрузка KV-кэша / FP4 KV-кэш	NVIDIA (H100/B200)	Продакшн, длинный контекст, SaaS
GPU Partitioning	AMD MI300X	Мульти-тенант, суверенные ЦОД
ROCm vLLM	AMD (MI300X, RX 7900)	On-prem РФ, альтернатива NVIDIA
LLM in a Flash / mlx-flash	Apple Silicon	Edge, R&D, автономные станции
GGUF / llama.cpp	Кросс-платформенный	Прототипы, CPU-only, низкий CapEx
vLLM / MOSEC	NVIDIA / AMD	Продакшн-инференс (Comindware)
GigaChat Speculative (MTP)	Cloud.ru / On-prem	Ускорение инференса в суверенных контурах
YandexGPT Lite	Yandex Cloud	Минимальные задержки для простых задач (RU)

10. Приложение E. Китайские альтернативные GPU для инференса

10.1 Обзор

Китайские AI-ускорители достигли производственной готовности для задач инференса больших языковых моделей.

Huawei Ascend 910C обеспечивает около 80% производительности NVIDIA H100 при цене 12–18 тыс. долларов США против 25–40 тыс. долларов для H100 (*«Huawei Ascend 910C Specifications»*).

DeepSeek V3/R1 развёрнут на Ascend 910C — подтверждённый кейс промышленного инференса (*«DeepSeek R1 on Huawei Ascend»*).

Для организаций в России/CIS китайские ускорители — практически единственный жизнеспособный путь формирования AI-инфраструктуры в условиях санкций в обозримой перспективе. Ascend 910C — рекомендуемый выбор с подтверждённым массовым производством (свыше 70 тыс. единиц) и развёртыванием в промышленной среде.

Сметы и тарифы — в *«Сайзинг и экономика»*.

Модель угроз и комплаенс — в *Приложении C «Безопасность, комплаенс, наблюдаемость»*.

10.2 Практический смысл

10.2.1 Для обоснования инвестиций

- **Диверсификация цепочек поставок**: доступ к AI-ускорителям при санкционных ограничениях на западные решения.
- **Оптимизация TCO**: стоимость-производительность китайских GPU на 30–50% выгоднее при сопоставимом качестве инференса — подтверждается характеристиками Ascend 910C (12–18 тыс. USD за 800 TFLOPS FP16).
- **Суверенный контур**: возможность построения полностью независимой AI-инфраструктуры без обходных схем и рисков перебоев поставок.

10.2.2 Факторы для принятия решений

- Ascend 910C — единственный массово производимый ускоритель с подтверждённым промышленным развёртыванием LLM (DeepSeek).

- Схема «обучение на западных GPU → инференс на китайских» — устойчивая модель для российских условий при наличии доступа к H100/H800 для обучения.
- Ограничения: отсутствие независимых бенчмарков, менее 2 лет производственной истории, ограниченная техническая поддержка в России.

10.3 Huawei Ascend 910C

Производственная альтернатива H100 для инференса:

Параметр	Ascend 910C	H100	A100	Примечание
FP16 TFLOPS	~800	990	312	около 81% от H100
Memory	128 GB HBM	80 GB	80 GB	в 1,6 раза больше H100
Memory BW	~3,2 ТБ/с	3,35 ТБ/с	2,04 ТБ/с	около 95% от H100
TDP	600 Вт	700 Вт	400 Вт	—
Цена (ориентир)	12–18 тыс. USD	25–40 тыс. USD	15–25 тыс. USD	—

Статус производства: массовое производство запущено, свыше 70 тыс. единиц в производственном цикле (сентябрь 2025), техпроцесс SMIC (Semiconductor Manufacturing International Corporation) 7nm (N+2) (*«TrendForce — Cambricon Production»*).

Программная экосистема: CANN 8.5.1 + torch-npu 2.9.0 + vLLM-Ascend — готова для промышленного инференса. Платформа ModelArts (Huawei Cloud) предлагает оптимизированное развёртывание DeepSeek R1 (*«vLLM-Ascend Documentation»*).

10.4 Moore Threads Huashan (прогноз 2026–2027)

MTT S4000/S5000 (текущее поколение): MTT S5000 — инференс DeepSeek V3: 1000 токенов/с декодирования, 4000 токенов/с предварительного заполнения. Программный стек MUSA с совместимостью, подобной CUDA.

Huashan (ожидается в середине 2026 года): заявлено увеличение плотности вычислений на 50% относительно предыдущего поколения, десятикратное улучшение энергоэффективности; позиционируется между поколениями Hopper и Blackwell. **Заявления не подтверждены независимыми источниками** — требуется мониторинг бенчмарков после официального анонса (*«TrendForce — Moore Threads Huashan»*).

10.5 Cambricon Siyuan 590

Параметр	Siyuan 590	A100	Примечание
FP16 TFLOPS	~390	312	в 1,25 раза больше A100
Memory	192 GB HBM2e	80 GB	в 2,4 раза больше A100
Memory BW	~2,4 ТБ/с	2,04 ТБ/с	—

Выход годных около 20% ограничивает объём реальных поставок. Основной заказчик — ByteDance (79% выручки). Плановый объём производства на 2026 год: 500 тыс. единиц (*«TrendForce — Cambricon Production»*).

Рекомендация: для моделей свыше 70 млрд параметров приоритет — Siyuan 590 благодаря 192 GB памяти.

10.6 MetaX

C500: около 75% производительности A100 по FP32. Массовое производство с февраля 2024 года. IPO в декабре 2025 года (Shanghai STAR Market), рост стоимости на 693%. Свыше 52 тыс. единиц отгружено в первом полугодии 2025 года (*«Wikipedia — MetaX»*).

Позиционирование: бюджетная альтернатива A100 для инференс-нагрузок средней интенсивности.

10.7 Россия/CIS: цепочки поставок

Канал	Описание	Уровень риска
Прямой	Закупка у китайских дистрибьюторов	Средний
Параллельный импорт	Юго-Восточная Азия → материковый Китай → Россия	Высокий (вторичные санкции)
Совместные предприятия	Локальные СП в России	Низкий

Китай обеспечивает 80–90% российских импортов полупроводниковой продукции (*«AEI — Semiconductor Sanctions on Russia»*). Китайское правительство негласно одобряет реэкспортные схемы.

10.8 Стоимость-производительность

GPU	Цена (оценка)	USD/TFLOPS FP16	Эффективность (TFLOPS/Вт)
H100	25–40 тыс. USD	25–40	1,41
Ascend 910C	12–18 тыс. USD	15–22	1,33
Ascend 910B	8–12 тыс. USD	13–20	1,50
Siyuan 590	8–12 тыс. USD	20–30	~1,6

Модель «два пути»: обучение на западных GPU (H800/H100), инференс на китайских (Ascend). DeepSeek демонстрирует эту модель — компания использует Ascend для инференса при обучении на NVIDIA H100/H800 (*«Tom's Hardware — DeepSeek CANN Support»*).

10.9 Рекомендации

10.9.1 Первоочередные (0–1 месяц)

- **Оценка Ascend 910C** для инференса LLM: направить запрос в Huawei Cloud (ModelArts) или к дистрибьюторам Ascend.
- **Проверка совместимости:** установить CANN 8.5.1 + torch-npu 2.9.0, проверить совместимость с целевыми моделями.

10.9.2 Краткосрочные (1–3 месяца)

- **Мониторинг Moore Threads Huashan** — ожидаются бенчмарки в середине 2026 года.
- **Развитие собственной экспертизы по CANN:** 2–3 месяца для выхода команды на продуктивный уровень.
- **Бенчмарк Siyuan 590** для моделей свыше 70 млрд параметров.

10.9.3 Среднесрочные (3–6 месяцев)

- **Пилотное развёртывание:** тестовое внедрение на Ascend 910C.
- **Оценка MetaX C500** для инференс-нагрузок без обучения.

10.10 Риски и ограничения

Пробел	Влияние	Снижение
Независимые бенчмарки	Только спецификации вендора	Запрос демонстрации PoC от вендора
Долгосрочная надёжность	Менее 2 лет производства, нет данных MTBF	Резерв 15–25% на непредвиденные расходы
Послепродажная поддержка	Ограниченное присутствие вендоров в России	Развитие собственной экспертизы
Механизмы оплаты	Параллельный импорт требует CNY/крипто	Юридическая экспертиза каналов

11. Приложение F. Дополнительные материалы

11.1 Реестр верифицированных источников

Перечень фундаментальных материалов для углубленного анализа: международные стандарты, управленческие фреймворки и инженерные бенчмарки ведущих консалтинговых групп.

11.2 Глобальное регулирование и стандарты

- EU Commission: GPAI Code of Practice (Final Draft July 2025)
- EU AI Act Compliance Guide 2026 (Unorma)
- G7 Hiroshima AI Process: International Guiding Principles
- IEEE P7000 Series: Process Model for Ethical AI Design
- OECD AI Principles and Governance Framework
- OECD Catalogue of Tools for Trustworthy AI
- UK AI Safety Institute: Systemic Safety Framework (2025)
- UNESCO Recommendation on the Ethics of AI (Global Implementation)

11.3 Управленческие методологии внедрения

- Accenture: Making Reinvention Real with GenAI (2025 Blueprint)
- BCG: From Potential to Profit with GenAI (2025 Framework)
- BCG: Closing the AI Impact Gap (2025 Deep Dive)
- Bain: State of the Art Agentic AI Transformation (2025)
- Bain: From Pilots to Payoff in Software Development (2025)
- Deloitte: State of GenAI in the Enterprise (Q3 2025)
- Gartner: Top Strategic Technology Trends for 2025 - AI focus
- KPMG: AI Governance for the Agentic Era (TACO Framework 2025)
- McKinsey: Rewiring the Enterprise for GenAI (2025)
- McKinsey: The GenAI Operating Model Leader's Guide (2025)
- McKinsey: Seizing the Agentic AI Advantage (June 2025 Report)
- PwC: Global AI Study 2025 - The Path to Value

11.4 Технические паттерны production AI и RAG

- Anthropic: Contextual Retrieval - Improving RAG Accuracy (2025)

- DoorDash: How We Built an Internal AI Platform That Works (2025) practical-strategies-vllm-performance-tuning)
- LangGraph: Enterprise Multi-Agent Orchestration Patterns (2025)
- Giskard: Open-Source Evaluation for LLM Agents (2026 Docs)
- OpenAI: Production RAG Best Practices & Evaluation (Cookbook)
- Microsoft: RAG Architecture on Azure AI Search (2025 Update)
- Uber Engineering: Genie - GenAI On-Call Copilot Architecture (2025)
- Uber Engineering: Raising the Bar on ML Model Deployment Safety (2025)
- vLLM: Performance Optimization and Tuning Guide (2025)
- [vLLM: Practical strategies for performance tuning (Red Hat 2026)](https://developers.redhat.com/articles/2026/03/03/

11.5 Экономика ИИ и FinOps

- AI Agent Cost Optimization: Token Economics in Production (Zylos 2026)
- CloudZero: FinOps for AI - Why AI Alters Cloud Cost Management (2026)
- CloudZero: Cloud Unit Economics In 2026 Guide
- Enrico Piovano: LLM Cost Engineering & Token Budgeting (2026)
- FinOps Foundation: Cost Estimation of AI Workloads (2026 Resource)
- FinOps in the AI Era: 2026 Survey Report (CloudZero)
- Mavik Labs: оптимизация стоимости LLM (маршрутизация, кэш, пакетная обработка, 2026)
- OpenAI: Real-time Cost and Token Monitoring (2025)
- OptyxStack Case Study: Reducing Inference Cost by 60% (2026)

11.6 Российский правовой и исследовательский контур

- ALRUD: ИИ и персональные данные — новые вызовы 2026
- BGP Litigation: Законопроект об ИИ — что нужно знать бизнесу (2026)
- Dentons: Регулирование ИИ в России — обзор 2025–2026
- Melling Voitishkin: Legal Alert — Маркировка ИИ контента в РФ
- НИУ ВШЭ: Исследование точности RAG-систем на русском языке (2025)
- ИТМО: Мультиагентная система ProAGI для разработки ПО (2026)
- TAdviser: Рынок ИИ в России — цифры и тренды 2025–2026
- Сколково: Потенциал GenAI для инженерных задач (Июль 2025)
- Yandex Research: Оптимизация инференса LLM для русского языка (2025)

11.7 Модели передачи и внешние кейсы внедрения

- InCommon: Why BOT Wins for AI Infrastructure
- InnovaNews — три кейса банковского ИИ: Альфа-Банк, Т-Банк и другие примеры (2026)
- Devico: Checklist for a seamless BOT transition (2025)
- Knowledge Transfer Framework for Enterprise Software Handover
- Software Handover Checklist 2026: Documentation & IP Guide
- Озон: ИИ как инструмент для 60% малых предпринимателей (2025)
- Самолет: Кейс «Цифровой рабочий» и ИИ в управлении стройкой (2025)
- СИБУР: Экономический эффект от ИИ на «Сибур-Нефтехиме» (200 млн руб)
- Яндекс: Корпоративный DeepResearch по кодовой базе (Кейс 2025)

11.8 Кураторские подборки и постоянный мониторинг

- Arxiv: Agentic RAG Taxonomy, Architecture and Research (March 2026)
- Arxiv: OrchMAS - Orchestrated Reasoning with Multi-Agents (March 2026)
- Arxiv: TreePS-RAG - Tree-based Process Supervision (Jan 2026)
- GitHub: Awesome AI Agents 2026 (300+ resources)
- GitHub: Awesome Production GenAI (Updated March 2026)
- GitHub: Awesome RAG Production Tools (Curated Feb 2026)

12. Приложение G. Перечень источников

Источники, использованные при подготовке отчёта, сгруппированные по темам.

Для расширенного круга чтения — *Приложение F «Дополнительные материалы»*.

12.1 Инженерия агентов и мультиагентные системы

- Agentic RAG / SGR
- Anthropic — Effective harnesses for long-running agents
- Anthropic — Harness design for long-running application development
- arXiv — Agent Skills in the Wild: An Empirical Study of Security Vulnerabilities at Scale, 2601.10338
- arXiv — Agent0: co-evolving curriculum and executor agents
- arXiv — General Agentic Memory (GAM)
- Bain & Company — The Three Layers of an Agentic AI Platform (апрель 2026)
- EffGen / agentic SLM
- GenAI Security — OWASP Top 10 for Agentic Applications for 2026
- Kaspersky Blog — Agentic AI security measures and OWASP ASI Top 10
- LangGraph — документация
- Martin Fowler — Harness Engineering (Thoughtworks)
- Medium — Qwen 3.5 35B A3B (AgentNativeDev)
- Microsoft Research — Fara-7B: An Efficient Agentic Model for Computer Use (PDF)
- MWS AI — MWS AI Agents Platform (описание модулей)
- OpenAI — Harness engineering
- Securelist — webinar: AI agents vs. prompt injections
- SGR Agent Core on GitHub
- vamplabAI/sgr-agent-core
- vamplabAI/sgr-agent-core — ветка tool-confluence
- Yandex Agent Atelier
- Хабр — Инженер будущего строит обвязку для агентов

12.2 Безопасность GenAI: OWASP, стандарты и практики тестирования

- BCG — AI Transformation Is a Workforce Transformation
- CodeWall — How we hacked McKinsey's AI platform (разбор red team)

- Datadog Security Labs — LiteLLM and Telnyx compromised on PyPI: Tracing the TeamPCP supply chain campaign
- GenAI Security — OWASP Top 10 for LLM Applications 2025
- GitHub — NVIDIA Garak (сканер для LLM, только изолированные стенды)
- GitHub — OWASP Application Security Verification Standard 5.0.0 (PDF, RU)
- GitHub — OWASP www-project-ai-testing-guide
- Habr — OWASP (вводные по тестированию и материалам сообщества)
- Habr — OWASP (смежные публикации сообщества)
- Habr — OWASP (смежные публикации сообщества)
- Habr — OWASP: LLM TOP 10 2025 (адаптация)
- HiddenLayer — AI Threat Landscape 2026
- Kaspersky — press release: training Large Language Models Security (описание программы)
- Kaspersky — пресс-релиз: угрозы под видом популярных ИИ-сервисов (бенчмарк тренда)
- Kaspersky Blog — How LLMs can be compromised in 2025
- Kaspersky Resource Center — What Is Prompt Injection?
- OWASP — Web Security Testing Guide (WSTG), stable
- OWASP — проект Top 10 for Large Language Model Applications
- OWASP Gen AI Security Project — Introduction
- OpenAI — приобретение PromptFoo (контекст рынка тестирования)
- The Hacker News — TeamPCP: LiteLLM и Telnyx скомпрометированы через PyPI (март 2026)
- Коммерсантъ — рынок и атаки на ИИ-системы (журналистский контекст)

12.3 Регуляторика ИИ, управление рисками и изолированные среды (песочницы)

- ACSOUR — обязанность операторов передавать анонимизированные ПДн в ГИС (152-ФЗ)
- DataGuidance — поправки к национальной стратегии развития ИИ РФ
- Daytona — Documentation
- E2B — Sandbox lifecycle
- E2B Blog — Up to 5x faster sandboxes
- EU AI Act Service Desk — статья 99
- EUR-Lex — Regulation (EU) 2024/1689 — Artificial Intelligence Act
- GitHub — llm-attacks/llm-attacks

- IEEE Xplore — Security-Performance Trade-offs of Kubernetes Container Runtimes (Viktorsson, Klein, Tordsson)
- ISO/IEC 42001:2023 — Artificial intelligence management system
- Modal — Sandboxes
- NIST — AI RMF to ISO/IEC 42001 Crosswalk (PDF)
- NIST — AI RMF: Generative AI Profile (NIST.AI.600-1, 2024)
- NIST AI Risk Management Framework 1.0
- NIST AIRC — Roadmap for the AI Risk Management Framework
- Redmadnews — AI-first стратегия: подкаст
- gVisor — Compatibility
- gVisor — Performance
- gVisor — Production guide
- Известия (EN) — создание офисов внедрения ИИ
- Национальная стратегия развития ИИ до 2030 года (Указ Президента РФ №124, февраль 2024)
- Официальное опубликование — Приказ Роскомнадзора от 19.06.2025 № 140 (обезличивание ПДн)
- Фонтанка — проект закона о госрегулировании ИИ (Минцифры, 18.03.2026)

12.4 Данные, доверие (AI TRiSM) и обзоры рынка enterprise AI

- Dataoorts — GPU cloud providers in Russia
- Gartner — AI TRiSM (глоссарий)
- Gartner — пресс-релиз: нехватка AI-ready data подрывает ИИ-проекты (26.02.2025)
- ITNext — GPU infrastructure as foundational to enterprise AI strategy
- Larridin — State of Enterprise AI in 2025 (независимый обзор, не первоисточник OpenAI)
- McKinsey — The State of AI 2025 (March 2025)
- NIST — SP 800-190, Application Container Security Guide
- OpenAI — The state of enterprise AI (обзор, декабрь 2025)
- OpenAI — The state of enterprise AI 2025 (PDF)
- a16z — Top 100 Gen AI Apps (6)
- red_mad_robot — исследования и материалы рынка GenAI
- red_mad_robot — мероприятие: тренд-репорт рынка GenAI (2025)
- Ведомости — рынок облачных сервисов с GPU (МНИАП, прогноз ~17,1 млрд руб.)
- Контур — агрегатор норматекста (документ по обезличиванию ПДн)
- Москва 24 — публикация о проекте закона об ИИ, 18.03.2026

- Правила тарификации Anthropic Claude
- РБК — объём рынка B2B LLM в России (~35 млрд руб., MTS AI)
- Сколково — программа «Переход в ИИ: трансформация бизнес-процессов» (CDTO)
- Сколково — событие «Состояние рынка GenAI в России и в мире» (12.02.2025)
- Указ Президента РФ №124 от 15.02.2024 (Национальная стратегия ИИ)
- Хабр — Релиз Claude Opus 4.6 (новость)
- Хабр — red_mad_robot: анонс тренд-репорта и события в Сколково

12.5 Экономика ИИ: рынок РФ, FinOps, облачные тарифы и on-prem

- AKM.ru — доступ к крупнейшей языковой модели на рынке РФ (Yandex B2B)
- Anthropic — Introducing Claude Opus 4.6
- Anthropic — Introducing Claude Sonnet 4.6
- CNews — Российский рынок генеративного ИИ в 2025 году
- CIO — MTS AI перенесла обучение моделей в облако
- Claude Docs — What's new in Claude 4.6
- Cloud.ru — Тарифы «Evolution Compute GPU», Приложение №7G.EVO.1 (январь 2026)
- Elish Tech — Где арендовать GPU-серверы дешевле и выгоднее: сравнение рынка в России и за рубежом
- Elish Tech — Почасовая аренда GPU A100 vs H100: что выгоднее в 2026 году
- FinOps Foundation — Framework: Unit Economics (Capability)
- FinOps Foundation — Generative AI / Unit Economics
- Hugging Face — Qwen/Qwen3.5-35B-A3B
- IMARC — Russia Artificial Intelligence Market
- Introl — планирование мощностей ИИ-инфраструктуры (прогнозы, McKinsey в обзоре)
- Introl — финансирование CapEx/OpEx и инвестиции в GPU
- Ivchenko, O. — Cloud vs On-Premise Economics for AI: A Structured Cost Framework (Zenodo, DOI 10.5281/zenodo.18678386, 2026)
- MWS — GPU On-premises
- MWS — MWS GPT (продукт)
- MWS — тарифы MWS GPT
- MarketsandMarkets — Russia AI Inference Platform as a Service (PaaS)
- McKinsey — The State of AI 2025
- McKinsey — The State of AI: как организации перестраиваются для извлечения ценности

- OECD — макроэкономический эффект ИИ в экономиках G7 (публикация, 2025)
- OpenAI — Prompt caching (снижение стоимости повторяющегося контекста)
- Ouroboros (Cloud.ru Marketplace)
- PitchGrade — AI Infrastructure Primer
- Redmadnews — AI + Economy: китайская модель масштабирования (пост канала, рыночный фон)
- Runpod — LLM inference optimization playbook (throughput)
- SWFTE — экономика частного AI / on-prem
- Sber Developers — Тарифы GigaChat API для юрлиц (февраль 2026)
- Selectel — Cloud GPU (облачные серверы с GPU)
- Selectel — Новости: новые конфигурации GPU-серверов от 50 руб./час
- Slyd — калькулятор TCO (on-prem и облако)
- Trend Micro — Your AI Gateway Was a Backdoor: Inside the LiteLLM Supply Chain Compromise
- TrendForce — Atlas 350 на Ascend 950PR
- VK Cloud — машинное обучение в облаке (документация)
- Yakov & Partners — AI 2025 (экономический эффект)
- Yakov Partners — Прогноз экономического эффекта ИИ в России, 2025
- Yandex AI Studio — доступные генеративные модели
- Yandex AI Studio — правила тарификации
- Yandex Cloud — GPU (графические ускорители), документация
- Yandex Cloud — Прайс-лист (текущие тарифы)
- ИИ в России — 2025: тренды и перспективы (Яков и Партнёры + Яндекс)
- Коммерсантъ — рынок GenAI
- МТС Cloud — виртуальная инфраструктура с GPU
- РБК — объём рынка ИИ (декабрь 2024)
- РБК Education — во сколько обойдётся ИИ-агент: подсчёты экспертов (2026)
- Сбер — портал GigaChat API

12.6 Облачные платформы РФ: модели, API, 152-ФЗ и публичные кейсы

- 1dedic — GPU-серверы
- CNews — кейс: MTS AI и экономия инвестиций за счёт облака MWS (обзор)
- Cloud.ru — Evolution Foundation Models (продукт, перечень моделей)
- Cloud.ru — Evolution Foundation Models, тарифы (2026)
- GigaChat API — обзор документации
- GigaChat API — продукт для бизнеса (Сбер)

- GitHub — sgl-project/sglang, PR #18802
- GitVerse — GigaTeam GigaChat 3.1
- Google — условия использования Gemma
- HOSTKEY — выделенные серверы с GPU
- Hugging Face — LICENSE (YandexGPT-5-Lite-8B), сырой текст соглашения
- Hugging Face — MiniMaxAI/MiniMax-M2
- Hugging Face — ai-sage/GigaChat3.1-10B-A1.8B (Lightning 3.0)
- Hugging Face — ai-sage/GigaChat3.1-10B-A1.8B (Lightning 3.1)
- Hugging Face — ai-sage/GigaChat3.1-702B-A36B (Ultra)
- Hugging Face — карточка модели YandexGPT-5-Lite-8B-instruct
- Hugging Face — организация ai-sage
- Hugging Face — GigaChat 3.1 Collection
- Hugging Face — коллекция GigaAM
- Hugging Face — коллекция GigaChat Lite
- Hugging Face — коллекция GigaEmbeddings
- InCommon — BOT Transfer Readiness & Handover Mechanics
- MERA — бенчмарк русскоязычных моделей
- MWS — MCP в Octapi и ускорение создания ИИ-агентов
- MWS — MWS Octapi (продукт)
- MWS — новость: хранение персональных данных в облаке
- MWS Docs — лицензионные условия ПО «MWS GPT»
- MWS Docs — условия облачного сегмента 152-ФЗ
- MWS GPT Model Hub
- SaluteSpeech (Сбер)
- Yandex AI Studio — агент с File Search
- Yandex AI Studio — тарифы SpeechKit
- Yandex Cloud — AI Speech / SpeechKit Hybrid
- Yandex Research — обзор направлений работ (2025)
- Yandex Research — принятые к ICML 2025 (список, в т.ч. KV-кэш)
- Альянс в сфере искусственного интеллекта
- МК Астрахань — GigaChat Enterprise / «ГигаЧат Бизнес» (СМИ, март 2026)
- МТС Cloud — IaaS 152-ФЗ УЗ-1
- Сбер — GigaChat API: модели и тарифы
- Сбер — SaluteSpeech: тарифы для юрлиц
- Хабр — GigaChat-3.1: большое обновление больших моделей (блог Сбера)
- Хабр — MTS AI: взаимная оценка LLM при улучшении Cotype

- Хабр — MTS AI: граф в RAG
- Хабр — MTC: MCP в Octapi и AI-агенты
- Хабр — MTC: RAG для поддержки (Confluence, Jira, гибридный поиск)
- Хабр — MTC: RAG-помощник для саппорта (смежная публикация)
- Хабр — MTC: архитектура LLM-платформы MWS GPT

12.7 GPU, каталоги моделей и нишевые облачные провайдеры

- Cloud4Y — облачный GPU-хостинг
- Hugging Face — deepseek-ai/DeepSeek-R1-Distill-Llama-70B
- Hugging Face — deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
- Hugging Face — moonshotai/Kimi-K2-Base
- Hugging Face — nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-FP8
- Hugging Face — openai/gpt-oss-120b
- Hugging Face — openai/gpt-oss-20b
- Hugging Face — openai/gpt-oss-safeguard-20b
- Hugging Face — zai-org/GLM-4.6
- Hugging Face — zai-org/GLM-4.7
- Hugging Face — zai-org/GLM-4.7-Flash
- Hugging Face — zai-org/GLM-5
- Hugging Face — организация Qwen
- Hugging Face — организация moonshotai
- Hugging Face — организация nvidia
- Immers Cloud — GPU
- Intelion Cloud
- NVIDIA — GeForce Software License
- NVIDIA Research — Nemotron 3 (обзор семейства)
- Ouroboros Containers (Cloud.ru)
- Selectel — Foundation Models Catalog
- Selectel — облако GPU (калькулятор)

12.8 Глобальные модели, цены API и оптимизация инференса

- AMD ROCm Blog — Best Practices for MI300X Inference
- AMD ROCm Blog — LLM Inference Optimization Using GPU Partitioning
- AMD ROCm Blog — vLLM on ROCm Attention Backend
- Anthropic — Pricing

- Anthropic — Pricing
- Apple Developer — WWDC 2025: Explore large language models on Apple silicon with MLX
- Apple ML Research — LLM in a Flash: Efficient Large Language Model Inference with Limited Memory (ACL 2024)
- Clarifai — llama.cpp: Hardware Choices & Tuning
- GLM-5 Documentation — Zhipu AI
- GitHub — matt-k-wong/mlx-flash (реализация для MLX, март 2026)
- Google — Deep-Thinking Ratio (DTR), arXiv:2602.13517
- Google — Gemini Embedding 2 (блог: нативно мультимодальные эмбеддеры)
- Google AI — Gemini API Pricing
- Hugging Face — Qwen/Qwen3-235B-A22B-Instruct-2507
- Hugging Face — Qwen/Qwen3-Coder-30B-A3B-Instruct
- Hugging Face — Qwen/Qwen3-Coder-480B-A35B-Instruct
- Hugging Face — Qwen/Qwen3-Next-80B-A3B-Instruct
- Hugging Face — State over Tokens
- Hugging Face — allenai/wildjailbreak
- Hugging Face — nvidia/NVIDIA-Nemotron-3-Super-120B-A12B-FP8
- Hugging Face — openai/gpt-oss-safeguard-120b
- Kunal Ganglani — AMD ROCm on Consumer GPUs 2026 Guide
- Luxoft — модель Build–Operate–Transfer (BOT)
- Microsoft — Copilot Cowork (блог Microsoft 365)
- NVIDIA — Nemotron 3 Super (технический блог)
- NVIDIA Technical Blog — оптимизация NVFP4 KV-кэша
- NVIDIA-NeMo/Guardrails
- OpenAI — Pricing
- OpenAI — Reasoning Models
- OpenAI — контроль рассуждения со скрытыми шагами (бенчмарк), 2603.05706
- OpenAI gpt-oss (GitHub)
- OpenAI gpt-oss-20b (Hugging Face)
- VC.ru — гайд по тарифам Claude и доступу из России
- XDA — Best AMD GPU for Local AI (RX 7900 vs RTX 3090)
- Yotta Labs — Best GPUs for LLM Inference 2026
- arXiv — Cache Me If You Must (KV-quantization), 2501.19392
- arXiv — LLM in a Flash: Efficient Large Language Model Inference with Limited Memory (оригинальная статья, 2312.11514)

- arXiv — Meta (Экстремистская организация, запрещена в РФ), OpenAI, xAI: непрерывное улучшение моделей (чаты), 2603.01973
- arXiv — Microsoft Research: безопасность агентов с внешними инструментами, 2603.03205
- arXiv — Moonshot AI: ускорение синхронного RL
- microsoft/markitdown
- run-llama/llama_index
- vLLM — OpenAI-Compatible Server
- vLLM Blog — выгрузка KV-кэша (KV offloading connector)

12.9 Открытые модели, серверы инференса и маршрутизация API

- @ai_archnadzor — CLI вместо MCP
- GitHub — redmadrobot-rnd/mcp-registry (MCP Tool Registry)
- Hugging Face — lmsys/toxic-chat
- LangChain Docs — Evaluation concepts (LangSmith)
- MOSEC — документация
- NeuralDeep — бенчмарки vLLM / RTX 4090
- OpenCode
- OpenCode — Ecosystem
- OpenCode — документация (Intro)
- OpenCode Zen — документация
- OpenRouter — агрегатор API
- OpenRouter — журналирование и политики провайдеров
- OpenWork (different-ai/openwork)
- Phemex News — Z.ai: планы открыть GLM-5.1
- Redmadnews — MCP Tool Registry / RAG
- Semantic Gravity Framework
- infinity-emb — документация
- langchain-ai/langchain — text-splitters
- mosecorg/mosec (GitHub)
- vLLM — документация
- vLLM — репозиторий проекта (GitHub)
- сервер инференса MOSEC — README проекта (пример публичного зеркала)
- Хабр — red_mad_robot: MCP Tool Registry (реестр MCP для RAG/агентов, открытый код)

12.10 Инструменты разработки, телеметрия качества и публичные кейсы внедрений

- @Redmadnews (red_mad_robot) — канал публикаций лаборатории и исследований
- Anthropic — субагенты в Claude Code SDK
- Arize Phoenix — документация
- Cline — Auto Approve и YOLO Mode
- Continue — настройка режима агента
- GitHub — организация ozontech (открытые репозитории)
- Habr — red_mad_robot: кейс RAG для ФСК
- InOrg — бесшовная передача (seamless handover) в модели BOT
- Just AI — корпоративный GenAI (упоминается как практикующий вендор)
- LangSmith — Online evaluations (how-to)
- LangSmith — документация
- Langfuse — документация observability / tracing
- Model Context Protocol — официальный сайт
- OpenCode (open-code.ai) — SDK
- OpenCode (open-code.ai) — агенты
- OpenHands — обзор SDK
- OpenInference — инструментирование ИИ для OpenTelemetry
- OpenTelemetry — OpenTelemetry for Generative AI (блог)
- OpenTelemetry — Semantic conventions for generative AI metrics
- OpenTelemetry — Semantic conventions for generative client AI spans
- OpenWork — AGENTS.md (сырой текст репозитория)
- OpenWork — ARCHITECTURE.md (сырой текст репозитория)
- RB.RU — 93% команд в маркетинге используют ИИ (обзор исследования CMO Club × red_mad_robot, 2025)
- Roo Code — режимы использования
- Schema-Guided Reasoning (SGR)
- Telegram — CMO Club Russia: анонс исследования GenAI в маркетинге (red_mad_robot × CMO Club, 2025)
- YouTube — подкаст «Ноосфера» #129: Илья Самофеев (red_mad_robot), AI-First / AI-Native (полная запись)
- aider — интеграция с Git
- Ведомости — CTO AI red_mad_robot (Влад Шевченко)
- Хабр — Ozon Tech: Query Prediction, ANN и обратный индекс
- Хабр — Ozon Tech: анонс ML&DS Meetup (MLOps, программа докладов)

- Хабр — Ozon Tech: пересборка конструктора чат-ботов (Bots Factory, no-code, масштаб)

12.11 НИОКР, зрелость ИИ, индексы рынка и практики учёта токенов

- @rmr_rnd — R&D red_mad_robot
- Accenture Memex(RL): Indexing Memory with Reinforcement Learning
- Databricks KARL: Knowledge-Aware Reasoning LLM
- Deloitte — State of AI in the Enterprise 2026
- GitHub — ozontech/file.d
- GitHub — ozontech/framer
- GitHub — ozontech/seq-db
- Gu, Dao — Mamba: Linear-Time Sequence Modeling with Selective State Spaces (arXiv:2312.00752)
- LLMoney — калькулятор цен токенов LLM
- METR — AI Developer Productivity Study July 2025
- Menlo Ventures — State of Generative AI 2025
- Oppo AI — Search More, Think Less (SMTL), arXiv:2602.22675
- Princeton — непрерывное обучение из взаимодействия пользователя с агентом, 2603.10165
- SkillNet (Alibaba, Ant, Tencent, Oppo), arXiv:2603.04448
- Stanford HAI — AI Index Report 2025
- arXiv — HybridFlow: Resource-Adaptive Subtask Routing for Edge-Cloud LLM Inference
- arXiv — MoE на стеке AMD (IBM, Zyphra и др.)
- arXiv — PRISM: Privacy-Aware Routing for Cloud-Edge LLM Inference
- Портал поддержки Comindware
- Спецификация TOON
- Хабр — гид по топ-20 нейросетям для текстов (в т.ч. цены)
- Хабр — обзор цен на токены

12.12 Отраслевые кейсы, экономика токенов и регуляторные сроки

- @ai_archnadzor — локальные модели для кодинга и снижения затрат
- AGORA — Industrial AI
- AI Cost Check — Reasoning Model Pricing
- CMO Club Russia
- EU AI Act Service Desk — Implementation Timeline

- LeanLM AI — LLM Cost Optimization
- NeuralDeep — экономика LLM-решений
- PerUnit AI — GPT-5.4 API Pricing Analysis
- Redmadnews — R&D в AI в 2026
- Redmadnews — СП с «ВымпелКом», фабрика ИИ-агентов
- Redmadnews — бизнес-завтрак КРОК
- Systenics: экономия токенов
- Telegram-канал @llm_under_hood
- Telegram-канал @neuraldeep
- Tensorlake: бенчмарки
- Ведомости — ИИ-рывок 2026: что изменится для российских компаний (альтернативная публикация)
- Ведомости — ИИ-рывок 2026: что изменится для российских компаний
- Канал @ai_archnadzor — RAG и архитектуры
- Канал @ai_machinelearning_big_data
- Хабр — MLOps и каскады моделей
- Хабр — автоматизация обучения и обновления моделей
- Хабр — классификация текстов диалогов на большом числе классов
- Хабр — обновление LLM: instruction following и tool calling

12.13 RAG, архитектуры и инженерные паттерны (обзорные материалы)

- CIO — интервью: чат-бот, масштаб обращений и сценарии
- Cog-RAG
- Disco-RAG
- ETL, эмбеддеры, ранжировщики, фреймворки RAG, eval, безопасность
- GenAI в продакшне: технологический манифест
- GitHub — yichuan-w/LEANN
- GraphOS для RAG
- Guardrails как архитектурный паттерн
- HippoRAG 2
- LEANN
- MLCommons — Inference Datacenter
- Medium — Agentic GraphOS: 16-слойная архитектура графов знаний для продакшена
- Nested Learning
- NeuralDeep — рекомендации по кластерам

- NeuralDeep — репозиторий (GitHub, vakovalskii/neuraldeep)
- OpenClaw (ex-Moltbot)
- Perplexica
- REFRAG
- Raft на Habr — чанкование
- Semantic Scholar — LEANN: индекс векторов с низким объёмом хранения
- Topo-RAG
- apxml.com — VRAM calculator
- «ITипичные аспекты Артёма» (Артём Лысенко)
- «Открытые системы» — RAG и LLM для поддержки операционистов
- Обзор локального стека наблюдаемости (канал @ai_archnadzor)
- Типы AI-агентов

12.14 Фреймворки внедрения, оценка качества и нормативный контур

- 152-ФЗ «О персональных данных»
- BCG — Closing the AI Impact Gap (2025)
- BitNet
- DSPy 3 и GEPA
- Doc-to-LoRA; память агентов (пост /191)
- GitHub — yksi12/prompts: generate-bpmn-prompt.md
- Lakera — платформа
- Marker-Inc-Korea/AutoRAG
- Multimodal LLM
- Neuraldeep.ru - база навыков для российских сервисов
- OCR: NEMOTRON-PARSE, Chandra, DOTS.OCR
- RAGAS — документация
- Shubhamsaboo/awesome-llm-apps
- Stanford HAI — AI Index Report 2025
- bpmn.io — нотация BPMN и инструменты моделирования
- chonkie-inc/chonkie
- datalab-to/marker
- docling-project/docling
- langgenius/dify
- mastra-ai/mastra
- openapi-to-cli (ocli) on GitHub
- protectai/rebuff

- stanford-futuredata/ARES
- Портал НПА — проект ФЗ об основах госрегулирования применения технологий ИИ (ID 166424)

12.15 Сообщество, рынок труда и вспомогательные вендоры

- Clawctl — 42 665 exposed OpenClaw instances (оценка уязвимых экземпляров)
- ElevenLabs
- GigaChain SDK (GitHub)
- GigaCode 2.0
- GigaIDE
- GitHub — Libr-AI/do-not-answer
- GitHub — elder-plinius/L1B3RT4S
- GitHub — paul-rottger/xstest
- GitVerse — GigaCode-агент
- GitVerse — GigaCode-чат
- Koda
- Ouroboros (GitHub)
- RB.RU — вакансии с ИИ-навыками: +170% YoY в I кв. 2026 (~2,7× к I кв. 2025; hh.ru × PR DEV, март 2026)
- SourceCraft
- Sourcegraph Cody
- Tabnine
- Together — Mamba-3 (блог)
- Tom's Hardware — Huawei Atlas 350 / Ascend 950PR
- state-spaces/mamba (GitHub)
- Банк России — официальные курсы валют
- Приложение C — управляемые песочницы, сравнение моделей и бенчмарки

12.16 Китайские альтернативные GPU для инференса

- AEI — Semiconductor Sanctions on Russia
- DeepSeek R1 on Huawei Ascend
- Tom's Hardware — DeepSeek CANN Support
- Tom's Hardware — Huawei Ascend Roadmap
- TrendForce — Cambricon Production
- TrendForce — Iluvatar CoreX Roadmap

- TrendForce — Moore Threads Huashan
- USCC — China's Facilitation of Sanctions Evasion
- vLLM-Ascend Documentation
- Wikipedia — MetaX
- Huawei Ascend 910C Specifications
- Huawei Central — Ascend 910C Specs