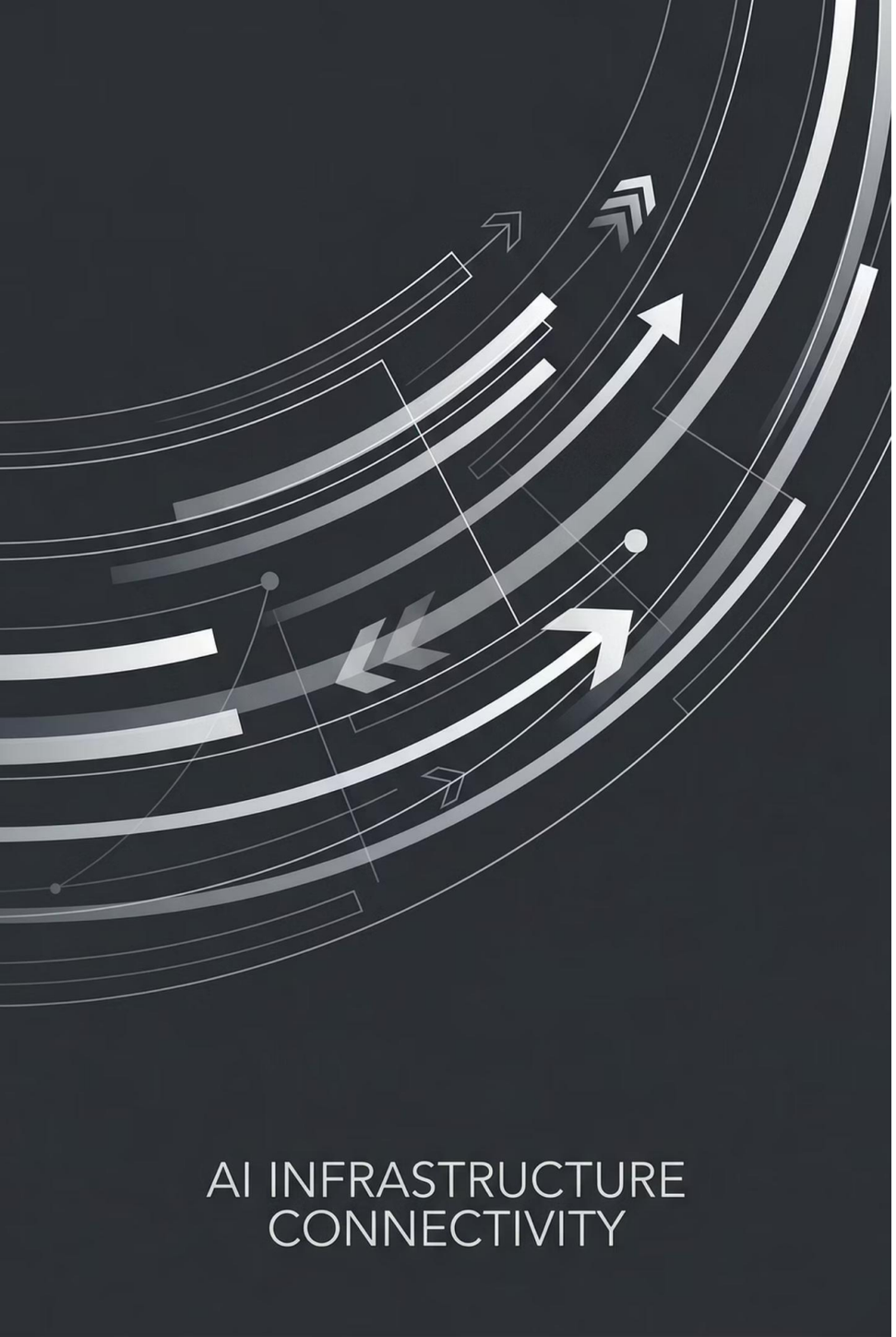
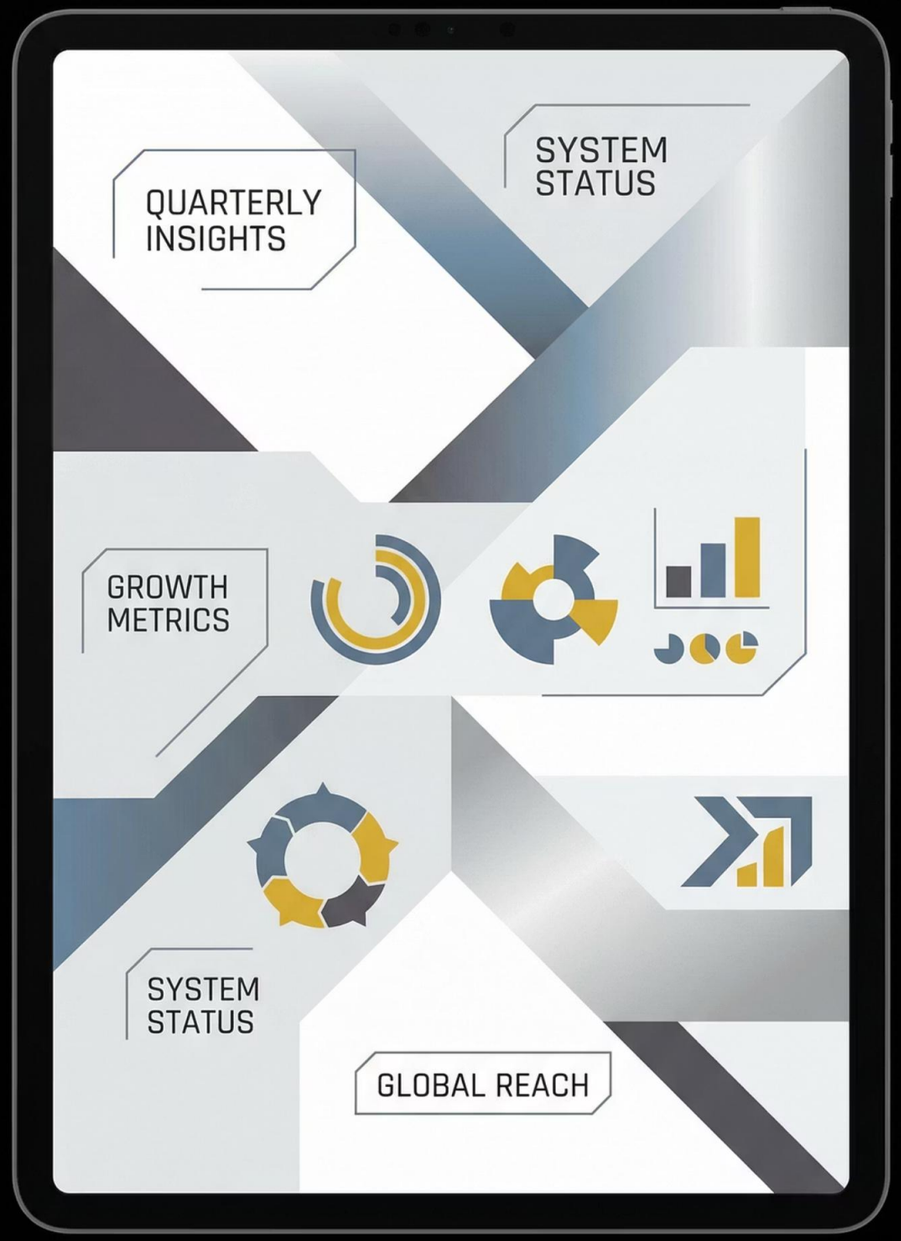


AiST Jetstream: Единая платформа контроля AI-расходов

*Превращаем хаос LLM-интеграций в прозрачную управляемую систему. Брандмауэр для вашей AI-инфраструктуры — **прокси-слой** между вашими разработчиками и всеми LLM-провайдерами, обеспечивающий полный контроль расходов, доступов и безопасности.*



Функционал AiST Jetstream

Система создана для решения ключевых проблем, возникающих при работе с любыми LLM API. В простой и удобной форме предоставляет инструменты разработчикам, службе безопасности, руководителям и бухгалтерии — чтобы у каждой роли был свой способ контроля ключевых показателей безопасности и финансов.

Важно: *В данной презентации представлены ключевые разделы и функции. Для получения полной документации или личного ознакомления с сервисом запросите демо, написав на info@deasoft.ru или нашим менеджерам в Telegram @TriggerBase*

ПРОБЛЕМА #1

Процесс закупки и управления ключами LLM

Любая компания стремится получить доступ к самым передовым LLM и генеративным моделям, однако на пути к их использованию возникает множество проблем и ограничений, которые значительно усложняют процесс интеграции и эксплуатации.

Геополитические ограничения

Уход с рынка РФ мировых поставщиков LLM, невозможность оплаты сервисов рублями, сетевые ограничения на прямое подключение к API провайдеров.

Сложность управления

Необходимость контролировать доступ ко всем платформам — OpenAI, Anthropic, Google, Mistral, Alisa Ai, GigaChat. У каждой свой биллинг, своя карта и свой личный кабинет, что создает операционную сложность.

Непрозрачное ценообразование

Множество моделей даже в рамках одного поставщика (у OpenAI — GPT-4.1, GPT-Image-1.5, GPT-5.2 и т.д.). Дашборды провайдеров показывают только агрегированные данные: видно «$500 потрачено вчера», но непонятно на что и кем.

Отсутствие бюджетного контроля

Сложно контролировать поставщиков LLM и сводить общий бюджет по расходам на их закупку. Отсутствие детализации по расходам приводит к бюджетным перерасходам.

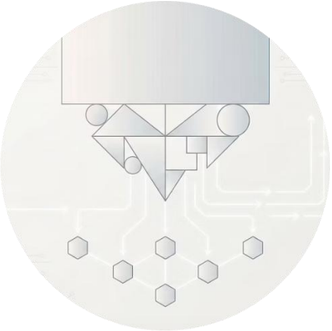
РЕШЕНИЕ #1

Система контроля закупок и доставки сервисов LLM

Система Jetstream обеспечивает полностью прозрачный процесс закупки и доставки сервисов до конечного продукта, а также полный контроль всех этапов бизнес-процесса — от закупки ключей до детального учета каждой транзакции.

Единый протокол доступа

Jetstream работает как LLM-прокси/шлюз (gateway): для клиента это обычный OpenAI-совместимый API, но точка входа (endpoint) — это URL вашего сервиса, а не провайдера. Универсальный gateway для любых внешних и внутренних LLM-систем.



Proxy-балансер

Несколько прокси-точек в разных странах (Финляндия, Польша, Эстония, Германия) обеспечивают высокую доступность. Автоматический failover при недоступности узла или региона гарантирует бесперебойность работы.

Контроль закупок ключей

Управление партнерами-поставщиками API-ключей в админ-кабинете. Подключение ключей, настройка закупочных условий, автоматический учет потребления и прозрачные взаиморасчеты без ручной сверки.

Актуальные цены

Команда Jetstream отслеживает и обновляет цены на ~30 моделей каждого провайдера. Вы задаете коэффициенты закупки/продажи один раз — система считает стоимость корректно, защищая от перерасхода.

Программный комплекс: три роли пользователей

В системе предусмотрено три основные роли, каждая с собственным личным кабинетом, набором прав и функций. Такое разделение обеспечивает безопасность, прозрачность и эффективное управление на всех уровнях организации.

Суперадминистратор

Полный доступ ко всей системе: управление данными, справочниками, биллингом, компаниями и всеми разделами. Контроль глобальных настроек и политик безопасности.

Разработчик / Компания

Управление и работа только в рамках своей компании или подразделения. Доступ к аналитике расходов, управлению токенами и настройкам проектов.

Поставщик LLM-ключей

Управление LLM-ключами и балансами, начисления средств, возможность вывода средств и контроль предоставляемых услуг.

Гранулярный биллинг: учет каждого запроса

*Гранулярный биллинг — это учет **каждого** API/LLM-запроса как отдельной транзакции с полной детализацией: кто сделал запрос, в каком проекте, какая модель использовалась, сколько токенов потрачено, какая стоимость.*

Зачем нужен?

Прозрачный контроль расходов, возможность быстро находить перерасход и оптимизировать бюджет. Гранулярность позволяет понять реальную картину использования LLM на уровне отдельных разработчиков и проектов.

Преимущества гранулярного биллинга

- ***Прозрачность затрат:*** *видимость каждого запроса с детализацией по исполнителю, проекту, модели и стоимости*
- ***Поиск аномалий:*** *выявление пиков нагрузки, зацикленных скриптов и лишних токенов*
- ***Раздельный учет:*** *разделение расходов по командам, проектам, production vs test окружениям*
- ***Оптимизация бюджета:*** *выбор оптимальных моделей и сокращение избыточных запросов*
- ***Лимиты и алерты:*** *установка квот и блокировок на уровне пользователя, команды или ключа*
- ***Аудит и прогноз:*** *детальные отчеты, сверка и точное планирование расходов*

Балансер ключей и отказоустойчивость

Для работы с любой LLM необходимы API-ключи и контроль баланса. Две критические проблемы могут остановить работу: исчерпание средств на аккаунте провайдера или временная недоступность API.

Автоматическая ротация ключей

При ошибке или исчерпании лимита Jetstream автоматически переключается на другой активный ключ, обеспечивая непрерывность работы приложений без вмешательства администратора.

Failover между провайдерами

Если API провайдера недоступен, система может автоматически перенаправить запросы на совместимый API другого провайдера (например, с OpenAI на DeepSeek), минимизируя простои.

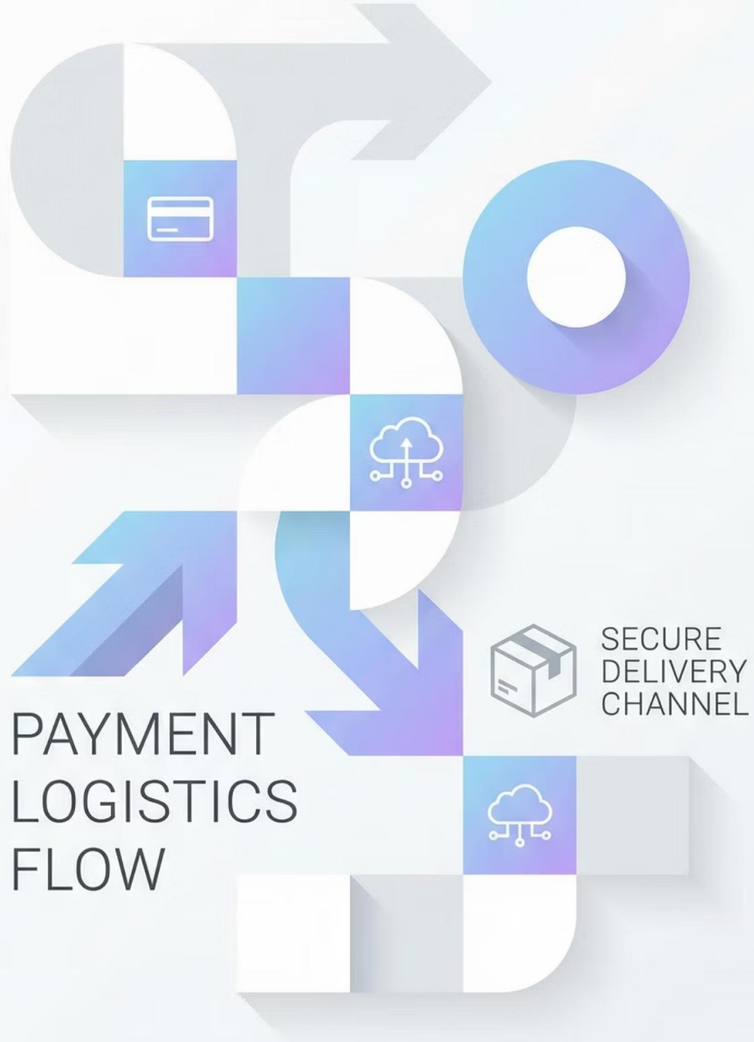
Такая архитектура обеспечивает высокую доступность (high availability) сервисов и снижает риски потерь из-за недоступности отдельных провайдеров или ключей.

Канал поставки любых LLM

> ***Хорошая новость!*** *Команда AiST Jetstream выстроила сеть поставщиков доступа к LLM и обеспечивает бесперебойный канал к самым современным текстовым и генеративным моделям в нужных объемах.*

Вы можете потреблять LLM-сервисы и API, ***рассчитываясь в рублях****, не думая о том, как оплачивать иностранные сервисы, где взять карту или счет, и как организовать регулярные платежи. Мы берем на себя всю операционную сложность международных расчетов.*

Доступ к ведущим российским и мировым провайдерам — OpenAI (GPT-4, GPT-4 Turbo, GPT-3.5), Anthropic (Claude 3 Opus, Sonnet, Haiku), Google (Gemini Pro, Ultra), Mistral AI, AlisaAi и другим — через единый интерфейс с едиными условиями оплаты и поддержки.



ПРОБЛЕМА #2

Контроль доступа к LLM-инфраструктуре

По мере того как AI становится частью инфраструктуры любой компании, четкое распределение доступов и их контроль становится наивысшим приоритетом. Отсутствие контроля приводит к критическим проблемам безопасности и финансовым рискам.

1

Риски при работе с API-ключами

Команде выдается один мастер-ключ без лимитов по бюджету и без ограничений по моделям. Это создает угрозу перерасхода и нерационального использования дорогих моделей типа GPT-4 или Claude Opus.

2

Отсутствие персональной ответственности

При работе с LLM транзакции могут быть небольшими, но частыми и неконтролируемыми. Общий ключ делает расходы «без владельца» — это как раздать команде корпоративные карты без имен и лимитов.

3

Проблемы масштабирования

С ростом числа разработчиков и микросервисов резко увеличивается количество интеграций и точек потребления. Без централизованного контроля растут «теневые» расходы, а перерасход обнаруживается постфактум.

РЕШЕНИЕ #2

Система управления LLM, доступами и токенами

Более 3 лет мы плотно занимаемся AI-решениями и более 15 лет — разработкой ПО для автоматизации бизнеса. Все описанные проблемы мы прочувствовали на себе при запуске и масштабировании AI-продуктов.

*Ключевая боль — **полный финансовый контроль:** прозрачный учет расходов, понимание эффективности использования AI-сервисов, а также контроль доступа и персональная ответственность за использование моделей и API.*

Управление компаниями и подразделениями

Все начинается с создания компании (или подразделения) в системе. Для каждой компании можно выдать персональный доступ в личный кабинет, где руководители и администраторы контролируют ключевые показатели и управляют настройками.

Создание компаний

Быстрое создание и настройка компаний или подразделений в системе с гибкими настройками доступа.

Персональные токены

Выпуск неограниченного числа токенов с индивидуальными лимитами и правилами доступа.

Контроль пиков

Автоматическое выявление аномалий нагрузки и предотвращение перерасхода.

Защита ключей

Исходные LLM-ключи зашифрованы и никогда не попадают в открытый доступ, защищая от утечек.

Детальная аналитика

Полный журнал транзакций с детализацией по компаниям, токенам и моделям.

Управление балансами

Автоматическое списание с баланса компании при использовании API.

Персональные токены: гибкий контроль доступа

Для каждого отдела, команды или продукта можно выпустить отдельный токен с индивидуальными настройками. Количество токенов внутри компании не ограничено, что позволяет гибко организовать доступ на любом уровне детализации.

Лимиты по расходам

- *Фиксированный общий лимит на токен*
- *Недельные лимиты для контроля текущего периода*
- *Месячные лимиты для бюджетного планирования*
- *Автоматическая блокировка при превышении*

Ограничения по моделям

- *White-list разрешенных моделей*
- *Black-list запрещенных моделей*
- *Разные правила для разных команд*
- *Например, junior-разработчикам только GPT-4.0 mini или Claude Haiku 4.1*

Защита от утечек

Персональные токены защищают от случайных утечек ключей в Git или при компрометации. При использовании персональных токенов риск критических потерь минимален: доступ ограничен лимитами, и при подозрительной активности токен можно мгновенно отключить.

Гибкость управления

Создавайте отдельные токены для production и test окружений, разных микросервисов, команд или даже отдельных разработчиков — система поддерживает любую модель организации доступа.

Аналитика запросов и журнал транзакций

Полный контроль всех обращений к LLM с детальной статистикой по компаниям, токенам и моделям. Журнал транзакций включает как успешные запросы, так и ошибки, что позволяет быстро выявлять проблемы и оптимизировать использование.

100%

Покрытие запросов

Каждый запрос логируется и доступен для анализа

<1s

Задержка логирования

Данные доступны практически в реальном времени

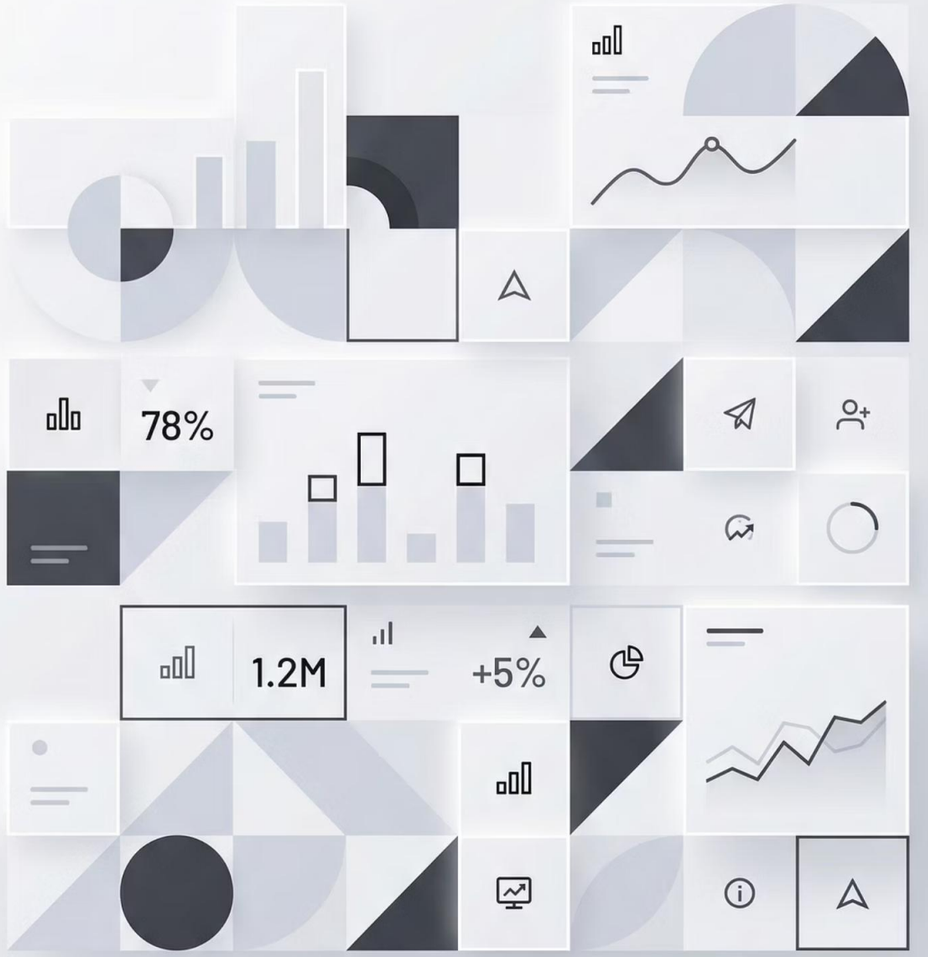
90+

Дней хранения

История транзакций для глубокого анализа

В журнале транзакций отображаются: временная метка, пользователь/токен, модель, количество токенов (input/output), стоимость, статус выполнения, время отклика и детали ошибок. Все данные можно фильтровать и экспортировать для дальнейшего анализа.

APEX
ANALYTICS



Контроль пиковых нагрузок и аномалий

Система автоматически анализирует трафик и помогает предотвращать перерасход, выявляя аномальный рост потребления в режиме реального времени. Это критически важно для защиты от случайных зацикливаний скриптов или ошибок в коде.

1

Обнаружение

Система выявляет резкий рост числа запросов или затрат по токену или компании

2

Уведомление

Администраторы получают мгновенное уведомление о потенциальной проблеме

3

Блокировка

Автоматическое отключение токена или модели при критическом превышении

4

Проверка

Администратор анализирует причину и принимает решение о разблокировке

Настраиваемые пороги аномалий позволяют адаптировать систему под специфику вашей нагрузки: от консервативных настроек для строгого контроля до гибких для динамичных проектов.

Личные кабинеты и API интеграция

Личные кабинеты

Вы можете выдавать сотрудникам и подразделениям доступ в персональный кабинет, задавать лимиты и балансы, а также разрешить самостоятельное пополнение через бухгалтерию.

Подразделения смогут самостоятельно отслеживать баланс, контролировать расходы и нагрузку, используя все преимущества сервиса без постоянного участия центрального IT.

API сервиса

У административного кабинета и кабинета компании есть полноценный API, который позволяет интегрировать Jetstream с вашими внутренними системами.

Возможности API:

- *Получение данных по транзакциям и расходам*
- *Программное управление компаниями и токенами*
- *Настройка лимитов и ограничений через код*
- *Автоматизация начислений и списаний*
- *Интеграция с корпоративными биллинг-системами*

Работа с локальными LLM

Для среднего и крупного бизнеса on-premise решения часто являются приоритетом: критически важную информацию нельзя отправлять во внешние LLM, а стабильность и бесперебойность работы — в числе главных требований. Построить собственную инфраструктуру и необходимую «обвязку» обычно сложно и дорого.

Масштабирование

Компании часто начинают с 1-2 локальных машин для LLM, затем расширяют инфраструктуру. Jetstream подключает неограниченное число узлов и работает как балансировщик, распределяя запросы на машины со свободными ресурсами, обеспечивая оптимальную утилизацию оборудования.

Амортизация и контроль стоимости

Локальные LLM тоже стоят денег — серверы, электричество, обслуживание. Без учета стоимости запросов легко получить перерасход. Jetstream предоставляет прозрачную аналитику по расходам, позволяя считать реальную окупаемость on-premise решения и принимать обоснованные решения об инвестициях.

Перепродажа и монетизация

Если у вас есть сильная отраслевая LLM-модель или сервис, Jetstream поможет организовать безопасный канал выдачи доступа и биллинг для внешних клиентов. На фоне высокого спроса на LLM можно монетизировать простаивающие ресурсы, заполняя их внешней нагрузкой и окупая инфраструктуру.

AiST Jetstream — ваш надежный партнер

Система помогает точно контролировать ресурсы, оптимизировать запросы и понимать себестоимость поддержки и развития AI-сервисов. Jetstream можно сравнить с «банком для LLM»: он учитывает потребление, управляет лимитами и обеспечивает персональную ответственность по доступам.

Безопасность

Полный контроль данных и доступов с возможностью модерации запросов

Экономия

Прозрачный учет расходов и оптимизация использования моделей

Масштабируемость

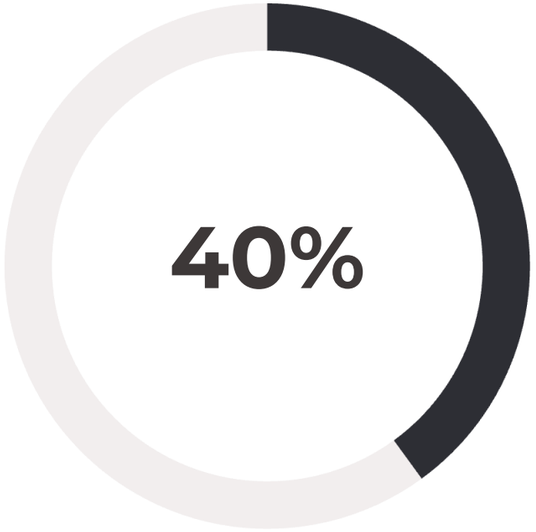
От нескольких разработчиков до сотен микросервисов

При этом вы можете безопасно открыть доступ к своим LLM-мощностям и моделям для партнеров и других компаний — и продавать свои решения на быстро растущем рынке LLM, превращая инфраструктуру в источник дохода.

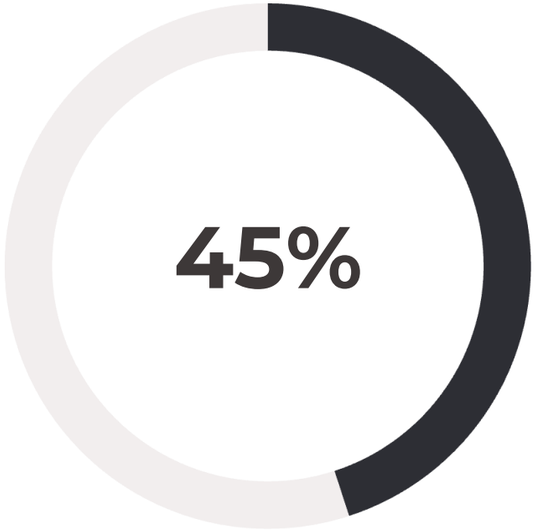
ПРОБЛЕМА #3

Безопасность и контроль данных

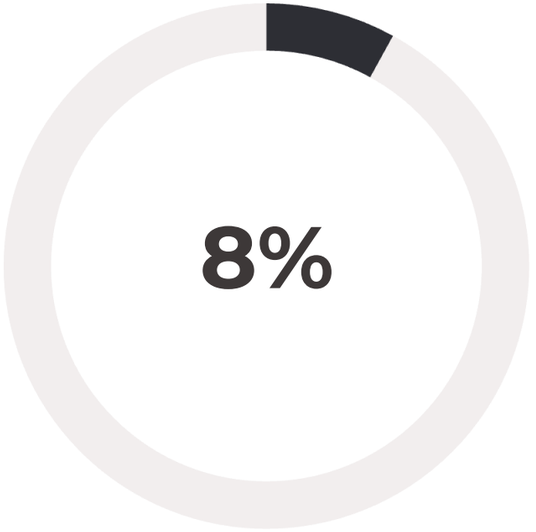
*В 2025 году через ИИ-сервисы (ChatGPT и Gemini) из российских компаний утекло **в 30 раз больше** конфиденциальных данных, чем годом ранее. Главная причина — сотрудники массово загружают рабочие документы в чат-боты для анализа и обработки.**



российских IT-компаний внедрили генеративный ИИ



компаний создали отдельные ИИ-подразделения



компаний не используют ИИ совсем

Анализ трафика 150 компаний показал, что публичные нейросети стали новым «теневым IT» и почти незаметным каналом утечек. В ИИ загружают презентации, стратегические планы, отчеты, таблицы с бизнес-данными, фрагменты исходного кода, переписку и техническую документацию.

**Источник https://www.cnews.ru/news/top/2026-02-04_sotrudniki_rossijskih_kompanij*

Что сотрудники загружают в публичные LLM

Доля использования ИИ в компаниях ежегодно растет — преимущества и уровень автоматизации, которые дает ИИ, трудно игнорировать. При этом отсутствие правил и инструментов контроля повышает риски утечек данных, нарушений требований ИБ и комплаенса.

Стратегические документы

Презентации для инвесторов, стратегические планы развития, финансовые модели, бизнес-планы

Конфиденциальные данные клиентов

Таблицы с персональными данными, договоры, коммерческие предложения, отчеты по проектам

Техническая информация

Фрагменты исходного кода, API-ключи и токены доступа, архитектурные решения, техническая документация

Внутренняя переписка

Email-переписка с коллегами и партнерами, внутренние обсуждения проектов, служебные записки

Юридические риски использования публичных LLM

1

152-ФЗ о персональных данных

Загрузка в публичные ИИ персональных данных клиентов или сотрудников = риск нарушения требований к обработке и защите ПДн

2

Трансграничная передача

ChatGPT/Gemini означают передачу данных за рубеж, что требует отдельного правового основания и соблюдения условий 152-ФЗ

3

98-ФЗ о коммерческой тайне

Выгрузка стратегий, финмоделей, планов, техдоков может считаться разглашением КТ

1

Договорные риски

Нарушение NDA и контрактов с клиентами (запрет передачи третьим лицам) → штрафы и расторжения

2

Утечка кода и доступов

Отправка исходников, логов, ключей/токенов повышает риск компрометации и может нарушать лицензии

3

Необходимость контроля

Нужны правила + обучение + технический контроль (DLP/прокси) + безопасный корпоративный контур

БЕЗОПАСНОСТЬ

Безопасность на первом месте

Ваши данные и AI-инфраструктура под надежной защитой. Jetstream построен с учетом самых строгих требований безопасности и compliance для регулируемых индустрий. Мы понимаем, что для корпоративных клиентов безопасность — это не опция, а базовое требование.

Шифрование

- *AES-256 для хранения данных*
- *TLS 1.3 для передачи*
- *Zero-knowledge архитектура*

Compliance

- *SOC 2 Type II (в процессе)*
- *GDPR compliant из коробки*

Контроль доступа

- *SSO и SAML 2.0 интеграция*
- *Обязательная MFA*
- *IP whitelisting*

Аудит

- *Immutable logs*
- *Экспорт для внешнего аудита*
- *SIEM интеграция*

РЕШЕНИЕ #3

Система модерации и деперсонализации данных

Jetstream — инструмент для администраторов и служб безопасности, который позволяет настраивать политики модерации и деперсонализации данных, отправляемых сотрудниками во внешние LLM-сервисы. Система анализирует текстовые запросы, изображения и документы и ***блокирует передачу*** *во внешние системы, если содержимое не соответствует установленным правилам.*

ВАЖНО: *Система модерации и деперсонализации работает на базе локальной LLM, размещенной в контуре заказчика либо у доверенного российского облачного провайдера. Обработка и хранение данных не выходят за пределы РФ — модель и инфраструктура расположены на территории Российской Федерации, данные не передаются и не сохраняются за рубежом.*

Сервис модерации запросов

Так как через протокол Jetstream проходит поток данных к различным LLM, мы добавили дополнительный уровень защиты — сервис модерации. Это отдельная LLM-система, которая работает локально на сервере в контуре заказчика или у доверенного провайдера в РФ.

01

Запрос пользователя

Пользователь отправляет запрос в LLM через Jetstream gateway

02

Проверка модерацией

Jetstream направляет запрос сначала в локальный сервис модерации

03

Анализ по политикам

Сервис проверяет содержимое по заданным политикам безопасности и деперсонализации

04

Решение о передаче

Запрос либо разрешается и отправляется во внешнюю LLM, либо блокируется с алертом администратору

Внутри кабинета есть раздел отчетов, где можно просмотреть все запросы и причины отклонения для анализа попыток нарушения политик безопасности.

Сервис деперсонализации

Если мы хотим допустить возможность доступа к внешним LLM, но при этом удалить конфиденциальную информацию, мы включаем сервис деперсонализации. Запросы проходят через локальную LLM, которая согласно правилам убирает лишние данные, и только затем передает очищенный запрос в целевую LLM.

Как работает деперсонализация

- *Автоматическое удаление имен, адресов, номеров телефонов*
- *Замена персональных данных на обезличенные токены*
- *Удаление коммерчески чувствительной информации*
- *Фильтрация технических данных (API-ключи, пароли)*
- *Сохранение смысла запроса для качественного ответа LLM*

Обработка ошибок

*Ошибки, которые возвращает Jetstream, **обязательно обрабатывайте** на стороне вашего сервиса. Иначе система может зацикливать повторные запросы или зря расходовать токены на повторную обработку.*

***Рекомендации:** различайте модерационные блокировки и технические ошибки, не делайте автоматические ретраи при модерации, используйте идемпотентный request_id.*

Преимущества решения Jetstream

Полная безопасность

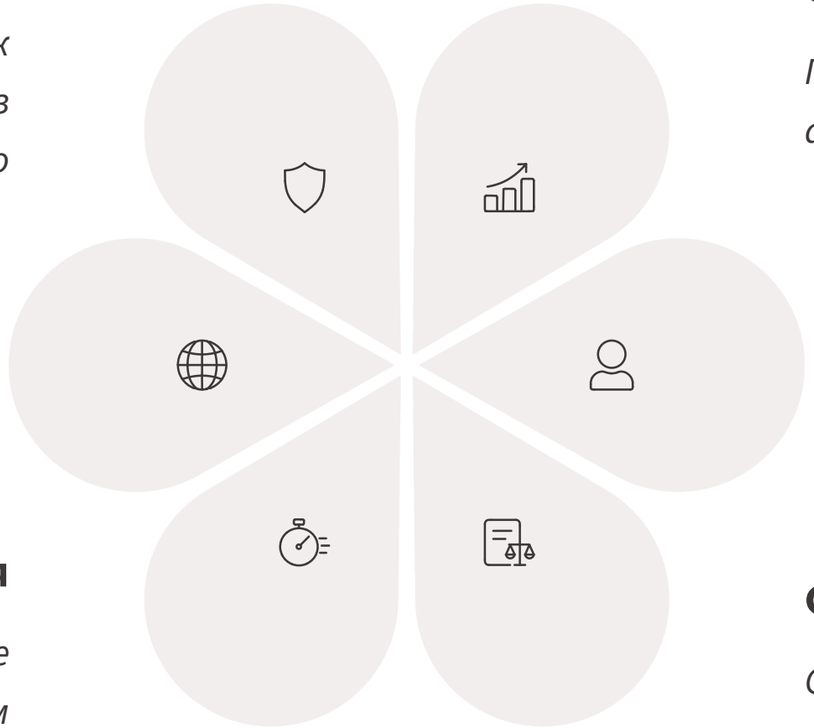
Предотвращение утечек конфиденциальной информации через модерацию и деперсонализацию

Канал поставки

Доступ ко всем ведущим LLM с оплатой в рублях через единый интерфейс

Быстрая интеграция

Подключение за 5-15 минут — изменение двух строк кода в существующем приложении



Финансовый контроль

Гранулярный биллинг и прозрачная аналитика расходов по каждому запросу

Управление доступом

Персональные токены с лимитами и правами для каждой команды или разработчика

Compliance

Соответствие 152-ФЗ, GDPR и внутренним политикам безопасности компании

Компания получает современный способ использовать преимущества генеративного ИИ, не теряя контроль над данными, расходами и соблюдением требований регуляторов.

Предлагаем провести вам демо

ИИ уже становится частью инфраструктуры любой компании. Jetstream — это «умный счетчик» и защитный слой, который помогает одновременно контролировать ***безопасность*** *и* ***расходы*** *при использовании внешних LLM.*

Без такого решения компании оказываются в одной из двух ситуаций: либо тратят время и бюджет на разработку собственного контура, либо недооценивают масштаб задач и риски, которые проявятся позже — утечки, проверки регуляторов, штрафы, договорные инциденты с клиентами.

Мы будем рады показать Jetstream в работе и помочь вам использовать все преимущества современных LLM — без потери контроля и без лишних рисков.

Email: *info@deasoft.ru* | ***Website:*** *aist-ai.com* | ***Telegram:*** *@TriggerBase*