Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Дайджест ИИ: Opus 4.8, Алиса на LLM, Android из AI Studio

Как мы оцениваем и внедряем новые ИИ-инструменты у клиентов: пилот, безопасность, замер цены

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Новые ИИ-инструменты выходят каждую неделю, и в компаниях их внедряют стихийно: сотрудники грузят договоры в чужие облака, код от нейросети уходит в прод без аудита, бюджет тратится на модель, не подходящую под задачу. Мы ставим процесс: закрытый пилот на реальных данных, проверка хостинга и происхождения данных, замер стоимости — и только потом перевод потока в прод.

Почему это важно бизнесу

- Утечка ПДн клиентов в зарубежное ИИ-облако — штрафы по 152-ФЗ и репутационные потери
- Код от нейросети без аудита — SQL-инъекции и XSS уезжают в прод вместе с фичей
- Зарубежные API из РФ нестабильны: процесс без резервного маршрута встаёт на день
- Неверно выбранная модель — переплата ×3–5 за токены на потоковых задачах
- Прототип за вечер вместо недель разработки — быстрее проверка гипотез и решений

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

$5/$25

Тариф Opus 4.8 за 1 млн токенов вход/выход — тот же, что у 4.7; бюджет пилота не пересчитываем

Anthropic API, claude-opus-4-8

до ×2.5

Скорость вывода в fast-режиме Opus 4.8 (speed:"fast", beta fast-mode-2026-02-01) — точечно на...

Anthropic API, fast mode (beta)

3 слоя

security-guidance: паттерн-предупреждения на правках, diff-ревью в чистом контексте, агентное...

anthropics/claude-code, security-guidance

1 млн

Контекст Opus 4.8 при выводе до 128K токенов — кодовая база клиента целиком в одном аудите

Anthropic API

×5

Заявленная вендором экономия Alice AI LLM Flash (почти в 5 раз) — проверяем на пуле из 100+ ре...

Yandex AI Studio

100

Лимит тестеров internal-трека Play Console для прототипа из AI Studio — без публикации в стор

Google Play Console

РЕАЛИЗАЦИЯ 1 ИЗ 3

ИИ-разработка: Opus 4.8 + аудитор security-guidance

Что настраиваем

Команды разработки и 1С-доработок: Claude Code на рабочих станциях, репозитории клиента в Git

Как мы это делаем

1. Ставим Claude Code на claude-opus-4-8: на агентных задачах effort high/xhigh, thinking adaptive задаём явно (по умолчанию на 4.8 thinking выключен); независимые под…
2. Подключаем плагин security-guidance командой /plugin из официального маркетплейса anthropics/claude-code — бесплатен на всех тарифах
3. Слой 1 — regex-предупреждения на каждом Edit/Write без вызова модели, ~25 опасных паттернов: pickle.load/yaml.load на недоверенных данных, torch.load(weights_only=F…
4. Слой 2 в конце хода ревьюит git diff отдельной моделью в чистом контексте; слой 3 при git commit — агентное кросс-файловое ревью (Read/Grep/Glob): IDOR, обход автор…
5. Fast-режим (speed:"fast") включаем точечно на интерактивных правках — у него отдельные лимиты и premium-тариф

РЕЗУЛЬТАТ

Разработчик получает параллельного «инженера по безопасности»: типовые дыры (инъекции, XSS, небезопасная десериализация) отсеиваются до PR, а не на пентесте. Доля security-замечаний на ревью падает кратно, стоимость решения — только токены: сам плагин бесплатный.

КЛЮЧЕВОЙ НЮАНС

Ключевой принцип из доки: ревьюер никогда не тот же инстанс, что писал код, — свежий контекст без привязанности к своему решению. Этот принцип «генератор и проверяющий в разных контекстах» мы переносим на все…

РЕАЛИЗАЦИЯ 2 ИЗ 3

Обработка обращений: Alice AI LLM Flash через API

Что настраиваем

Клиентские сервисы до 50 РМ: классификация писем и заявок, база знаний, суммаризация — контур данных в РФ

Как мы это делаем

1. Разворачиваем доступ в Yandex AI Studio (Foundation Models API); Alice AI LLM Flash заточена под скорость и поток однотипных задач: классификация, диалог, извлечени…
2. Собираем эталонный пул из 100–200 реальных обращений клиента и гоняем A/B против текущего решения: точность, латентность, цена за 1000 обращений
3. Включаем function calling: модель сама дёргает наши инструменты — поиск по базе знаний, постановку задачи в CRM, эскалацию на инженера
4. ПДн маскируем на нашем шлюзе до отправки в API; договор и хостинг данных — РФ, требования 152-ФЗ закрываются штатно, VPN и валютная оплата не нужны

РЕЗУЛЬТАТ

Первичная линия работает 24/7: типовые обращения классифицируются и получают черновик ответа за секунды, инженер подключается только на эскалациях. Стоимость потока кратно ниже зарубежных аналогов — и мы подтверждаем это собственным замером, а не пресс-релизом вендора.

КЛЮЧЕВОЙ НЮАНС

Цифры производителя («почти в 5 раз дешевле») — маркетинг, пока не воспроизведены на задачах клиента. Решение принимаем только по своему эталонному пулу с метриками до/после.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Прототипы Android-приложений из Google AI Studio

Что настраиваем

Проверка мобильных гипотез заказчика: нативный Android-прототип без бюджета на команду разработки

Как мы это делаем

1. В AI Studio открываем Build mode, в переключателе платформ выбираем Android — агент генерирует полный проект на Kotlin + Jetpack Compose (Material 3, build.gradle.k...
2. Прогоняем прототип в браузерном Android-эмуляторе, ставим на физическое устройство; баги агент чинит сам по описанию, фичи добавляем промптами
3. Там, где нужны ИИ-функции (чат, распознавание), приложение подключает Gemini API прямо из сгенерированного кода
4. Публикуем во внутренний трек Play Console — до 100 тестеров на стороне заказчика без релиза в стор
5. Для боевого продукта экспортируем проект ZIP-архивом в Android Studio: серверная логика, secrets и Firebase в AI Studio недоступны (client-side only) — дорабатывает...

РЕЗУЛЬТАТ

Гипотеза проверяется за 1–2 дня вместо 4–6 недель разработки MVP: заказчик показывает кликабельный нативный прототип инвестору или сотрудникам до того, как потратил бюджет. Решение о полноценной разработке принимается на данных теста, а не на слайдах.

КЛЮЧЕВОЙ НЮАНС

Жёсткая граница из документации: приложения из AI Studio — client-side only. Всё, что требует серверной части (Firebase, управление секретами, мультиплеер), в прототип не попадёт — закладываем это в скоуп тест...

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- ✕ **ПДн улетают в зарубежное ИИ-облако**
  Сотрудники грузят договоры в чат-боты. Ставим шлюз с маскированием ПДн и политику: клиентские данные — только в модели с хостингом в РФ (152-ФЗ).

- ✕ **Fast-режим включён на всём подряд**
  speed:"fast" даёт до ×2.5 к скорости вывода, но по premium-тарифу и с отдельными лимитами; фоновые и батчевые задачи держим на стандартном режиме той…

- ✕ **Плагин «заменил» ревью человека**
  security-guidance режет типовые дыры, но бизнес-логику и права доступа проверяет инженер: авто-аудит — это фильтр перед ревью, а не приёмка.

- ✕ **Вера цифрам вендора без замера**
  «Почти в 5 раз дешевле» — меряем сами: эталонный пул обращений клиента, A/B и цена за 1000 запросов до перевода боевого потока.

- ✕ **Прототип из AI Studio ушёл в прод**
  Client-side only: нет secrets, Firebase и серверной логики; ключ Gemini в APK — утечка. Боевую версию собирает инженер в Android Studio.

- ✕ **Нет резерва на зарубежный API**
  Доступ к зарубежным ИИ-API из РФ нестабилен. Для критичных потоков закладываем резервный маршрут и локальную альтернативу (Alice AI LLM Flash).

- ✕ **Коммиты мимо агентного аудита**
  Кросс-файловый слой плагина стартует при git commit из сессии Claude; коммиты мимо неё не проверяются — дублируем ревью GitHub Action claude-code-sec…

- ✕ **ИИ-потоки без владельца и бюджета**
  Токены списываются, качество не меряется. На каждый поток назначаем владельца, лимит расходов и метрики: cost/1000 запросов, точность, латентность.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- Письменная политика: какие данные можно отправлять в облачный ИИ; ПДн — нельзя
- Плагин security-guidance у всех, кто пишет или заказывает код с ИИ
- Прототипы — в AI Studio, боевой код — только после аудита инженером

НОРМАЛЬНО

- Шлюз с маскированием ПДн перед любым внешним ИИ-API
- Пилот Alice AI LLM Flash на 100+ реальных обращениях с замером цены и качества
- Claude Code: effort high/xhigh на агентных задачах, fast-режим — точечно
- CI-ревью безопасности (GitHub Action claude-code-security-review) как дубль локально...

ХОРОШО

- Двухконтурная схема: данные клиентов — контур РФ, кодовые задачи — Claude с резервом
- Метрики по каждому ИИ-потоку: cost/1000 запросов, точность, латентность
- Разделение «генератор/ревьюер» в чистых контекстах во всех ИИ-пайплайнах
- Ежеквартальный пересмотр стека моделей по свежим релизам и ценам

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Есть ли письменная политика, какие данные сотрудникам можно отправлять в облачные ИИ-сервисы?
- [ ] Стоит ли security-guidance или аналогичный авто-аудит у всех, кто пишет код с помощью ИИ?
- [ ] Проверяется ли сгенерированный код вторым контуром — чистым контекстом или CI-ревью?
- [ ] Замеряли ли вы цену за 1000 запросов на своих задачах, а не по прайс-листу вендора?
- [ ] Есть ли резервный маршрут на случай недоступности зарубежного ИИ-API?
- [ ] Обрабатываются ли клиентские ПДн только в моделях с хостингом данных в РФ?
- [ ] Отделены ли прототипы (AI Studio) от боевых приложений формальной приёмкой?
- [ ] Назначен ли владелец, бюджет и метрики для каждого ИИ-потока в компании?

Если хотя бы на два вопроса ответ «нет» или «не знаю» — тема требует внимания.

ЧТО МЫ ДЕЛАЕМ

Как поможет ITFresh

ITFresh — ИТ-аутсорсинг для юридических лиц до 50 рабочих мест в Москве и области. 15+ лет практики, собственная инфраструктура в дата-центре МТС (8 серверов Dell Xeon Platinum).

- Подбор ИИ-стека под задачи, бюджет и требования к данным: пилот на ваших кейсах с метриками до/после
- Внедрение Claude Code с security-guidance и CI-ревью в процессы разработки и 1С-доработок
- Интеграция Alice AI LLM Flash: обработка обращений, база знаний, function calling в вашу CRM
- Android-прототип из AI Studio за 1–2 дня для проверки гипотезы перед инвестициями в разработку
- Шлюз маскирования ПДн и политика использования ИИ в компании под 152-ФЗ

15+
лет в ИТ-поддержке

50
рабочих мест — наш профиль

МТС
дата-центр, Москва

КОНТАКТЫ

Обсудить вашу задачу

Сайт	**itfresh.ru**
Телефон	**+7 903 729-62-41**
Telegram	**@ITfresh_Boss**

Бесплатно посмотрим вашу инфраструктуру по этому чек-листу и скажем, где тонко — без обязательств.

itfresh.ru

ДОКУМЕНТАЦИЯ И НАШ СТЕК

Техническая база

01 Anthropic — Models overview: Claude Opus 4.8 (цены, контекст 1M) (platform.claude.com — 2026)

02 Anthropic — Fast mode (speed, beta fast-mode-2026-02-01) (platform.claude.com — 2026)

03 anthropics/claude-code — plugins/security-guidance (3-слойное ревью) (github.com — 2026)

04 Yandex AI Studio — обзор AI-моделей: Alice AI LLM Flash (yandex.cloud — 2026)

05 Gemini API — Build Android Apps in Google AI Studio (ai.google.dev — 2026)

06 Play Console Help — внутреннее тестирование (до 100 тестеров) (support.google.com — 2026)

07 Шаблон ITfresh: пилот ИИ-инструмента (эталонный пул, A/B-замер) (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.