Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

Gemini 3.x в малом бизнесе: как мы внедряем ИИ-инструменты Google

Omni, Spark и 3.5 Flash: что из майской волны анонсов мы берём в рабочие процессы клиентов

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Заказчику до 50 рабочих мест нужно снять с сотрудников рутину — разбор входящих писем и первички, выжимки из договоров, черновики ответов клиентам, — не нанимая отдельных людей и не сливая персональные данные в облако. Мы решаем это внедрением моделей Gemini 3.x через API: подбираем модель под задачу, считаем бюджет по токенам и выстраиваем контур приватности.

Почему это важно бизнесу

- Час работы бухгалтера или юриста на рутинном разборе документов стоит дороже, чем 1М токенов Flash-модели
- Gemini 3.5 Flash бесплатен в приложении Gemini — сотрудники начинают пользоваться сами, без контроля данных
- Промо-цены API на 3.6/3.7 Flash действуют до 31.12.2026 — бюджет 2027 года надо считать по двойному тарифу
- Утечка ПДн клиентов через облачный ИИ — это риск по 152-ФЗ, а не только репутационный
- Часть анонсов I/O — Spark, Omni — в статусе preview: строить на них основной процесс сейчас нельзя

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

1 048 576

входное окно Gemini 3.5 Flash в токенах; вывод — до 65 536 токенов

ai.google.dev/gemini-api/docs/models

$0.75/1M

вход Gemini 3.7 Flash (GA) по промо-тарифу до 31.12.2026; выход $3.75, с 2027 — $1.50/$7.50

ai.google.dev/gemini-api/docs/pricing

$1.50/1M

вход Gemini 3.5 Flash по API (выход $9.00); в приложении Gemini модель бесплатна

ai.google.dev/gemini-api/docs/pricing

$0.30/1M

вход Gemini 3.5 Flash-Lite (выход $2.50) — субагент для массовой классификации

ai.google.dev/gemini-api/docs/pricing

-50%

скидка Batch API на отложенную пакетную обработку — наш режим для ночного разбора архивов

ai.google.dev/gemini-api/docs/pricing

$2-4/1M

вход Gemini 3.1 Pro (preview): $2 до 200k токенов, $4 свыше; выход $12/$18

ai.google.dev/gemini-api/docs/pricing

РЕАЛИЗАЦИЯ 1 ИЗ 3

Разбор входящей первички и писем через Gemini API

Что настраиваем

Бухгалтерская фирма, 15-30 рабочих мест, поток 100-300 входящих документов в день

Как мы это делаем

1. Пилот на обезличенной выборке: 50 типовых документов прогоняем через gemini-3.5-flash-lite (классификация) и gemini-3.5-flash (извлечение полей)
2. Включаем structured outputs: JSON-схема с полями контрагента, суммы, даты, типа документа — модель возвращает строго валидный JSON
3. Ночной архив уводим в Batch API (-50% к тарифу), онлайн-поток — синхронно через gemini-3.5-flash-lite за $0.30/1M входа
4. Промежуточный сервис на нашей стороне маскирует ИНН/ФИО перед отправкой в API и восстанавливает после ответа
5. Результат складываем в очередь на подтверждение оператором — автозапись в учётную систему только после клика человека

РЕЗУЛЬТАТ

Оператор перестаёт перепечатывать документы и только подтверждает распознанное. Бюджет предсказуем: при среднем документе 3-5 тыс. токенов месячный поток в 6 000 документов укладывается в десятки долларов API даже на 3.5 Flash, на Flash-Lite — ещё дешевле.

КЛЮЧЕВОЙ НЮАНС

Каскад «дешёвая модель классифицирует — средняя извлекает» экономит в разы против прогона всего через одну модель. Structured outputs обязательны: без схемы парсинг ответов ломается.

РЕАЛИЗАЦИЯ 2 ИЗ 3

Ассистент по базе знаний: договоры и регламенты

Что настраиваем

Юридическая фирма, до 20 рабочих мест, корпус документов 2-4 тыс. страниц

Как мы это делаем

1. Собираем корпус в File Search поверх gemini-embedding-001 — модель отвечает по документам фирмы, а не из общих знаний
2. Для длинных договоров используем окно 1M токенов Gemini 3.5 Flash напрямую: PDF до сотен страниц уходит одним запросом без нарезки
3. Часто используемый корпус кладём в context caching: кешированные входные токены кратно дешевле обычных, платим только за кеш и почасовое хранение
4. Function calling подключаем к внутреннему реестру дел: модель сама дёргает поиск по номеру дела вместо галлюцинаций
5. Логируем все запросы и ответы в наш журнал — юристы проверяют выборочно, спорные ответы уходят в дообучение промптов

РЕЗУЛЬТАТ

Поиск по регламентам и типовым договорам сокращается с десятков минут до секунд, при этом ответ всегда со ссылкой на исходный фрагмент. Кеширование корпуса снижает стоимость повторных обращений на порядок против «сырой» подачи документов в каждый запрос.

КЛЮЧЕВОЙ НЮАНС

Длинный контекст не отменяет RAG: подавать 1M токенов в каждый запрос дорого. Кеш плюс эмбеддинги — рабочая экономика, прямая подача — только для разовых разборов.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Контент для карточек товара: Omni и Nano Banana

Что настраиваем

Торговая компания, интернет-каталог на 1-2 тыс. позиций, без штатного дизайнера

Как мы это делаем

1. Фото товаров дорабатываем через gemini-3.1-flash-image (Nano Banana 2): фон, свет, ретушь — по текстовому промпту, без графредактора
2. Короткие ролики для карточек и соцсетей пробуем на gemini-omni-1.1-flash: вход — фото товара плюс текст, выход — видео с корректной физикой
3. Omni держим в статусе управляемого пилота: модель в preview, результаты проходят ручную премодерацию до публикации
4. Метаданные и описания генерирует gemini-3.5-flash-lite пакетно через Batch API — 2 тыс. карточек за один ночной прогон

РЕЗУЛЬТАТ

Карточки товара получают единый визуальный стиль и видео без съёмочной группы. Текстовая часть каталога полностью автоматизирована, стоимость генерации описаний на весь каталог — единицы долларов в Batch-режиме.

КЛЮЧЕВОЙ НЮАНС

Генеративные медиа — это про маркетинг, а не про учётные процессы: сюда их и внедряем. Всё сгенерированное до публикации смотрит человек — правило без исключений.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

× **Процесс на бесплатном приложении**
Бесплатный 3.5 Flash в приложении Gemini — без SLA, API и контроля данных. Для бизнес-процесса — только API с ключом организации и логированием.

× **Бюджет по промо-ценам**
Тарифы 3.6/3.7 Flash ($0.75/$3.75) действуют до 31.12.2026, с 2027 — вдвое выше. Годовой бюджет считаем сразу по $1.50/$7.50.

× **ПДн и тайна в облако без маскировки**
Всё отправленное в облачный API уходит на сторону. Мы маскируем ИНН, ФИО и реквизиты на своём сервисе до запроса — иначе риск по 152-ФЗ.

× **Preview-модели в проде**
3.1 Pro, Omni, Spark — preview или свежие релизы: поведение и цены меняются. В прод берём только GA-модели, preview — в пилоты.

× **Модель без пина версии**
Алиас «latest» молча меняет поведение промптов после апдейта. Фиксируем конкретный ID (gemini-3.5-flash) и прогоняем регресс-набор при смене.

× **Считать только входные токены**
Выход у 3.5 Flash дороже входа в 6 раз ($9 против $1.50). Ограничиваем max_output_tokens и режем болтливость промптом — иначе смета плывёт.

× **Spark с широкими доступами**
Фоновый агент, действующий от имени сотрудника круглосуточно, — это доступ к почте и файлам 24/7. Даём только песочницу и отдельный аккаунт.

× **Один запрос — вся задача**
Прогон всего потока через старшую модель сжигает бюджет. Каскад Flash-Lite -> Flash -> Pro по сложности снижает стоимость в разы.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- Пилот на API-ключе организации: gemini-3.5-flash, обезличенные данные
- Structured outputs с JSON-схемой для любых извлечений полей
- Запрет загрузки ПДн и коммерческой тайны в публичные ИИ-чаты — политикой
- Пин конкретного ID модели, никаких алиасов latest

НОРМАЛЬНО

- Каскад моделей: Flash-Lite на классификацию, Flash на извлечение и тексты
- Batch API (-50%) для всего, что не требует ответа онлайн
- Маскировка реквизитов на своём прокси-сервисе до отправки в API
- Логирование запросов/ответов и выборочная проверка человеком

ХОРОШО

- Vertex AI с проектной изоляцией и управлением данными вместо голого API
- Context caching для постоянного корпуса документов + File Search (RAG)
- Function calling к учётным системам вместо копипаста данных в промпт
- Регресс-набор промптов и метрики качества перед каждой сменой модели

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Определён перечень данных, которые запрещено отправлять в облачный ИИ (ПДн, тайна, реквизиты)
- [ ] API-ключ выпущен на организацию, доступ сотрудников — через наш сервис, а не личные аккаунты
- [ ] Выбрана модель под задачу: Flash-Lite ($0.30/1M) для потока, Flash для документов, Pro — точечно
- [ ] Включены structured outputs и задан max_output_tokens в каждом интеграционном вызове
- [ ] Массовые задачи переведены на Batch API со скидкой 50%
- [ ] Посчитан бюджет 2027 года по полным тарифам, а не по промо до 31.12.2026
- [ ] Версия модели зафиксирована по ID, есть регресс-набор для проверки при обновлении
- [ ] Ответы модели попадают в учётную систему только после подтверждения человеком
- [ ] Логи запросов и ответов хранятся у нас и выборочно проверяются
- [ ] Preview-функции (Omni, Spark) — только в изолированных пилотах

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).

- Проведём пилот на ваших обезличенных документах и покажем точность и стоимость до внедрения
- Построим контур приватности: маскировка реквизитов, свой прокси, логирование — под требования 152-ФЗ
- Организуем легальный доступ к Gemini API через партнёрские схемы оплаты для юрлиц РФ
- Интегрируем модели с 1С и почтой через function calling и наш промежуточный сервис
- Возьмём на сопровождение: мониторинг лимитов, смену версий моделей, контроль бюджета токенов

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 Gemini API — Models (актуальная линейка и ID моделей) (ai.google.dev — 2026)

02 Gemini API — Pricing (тарифы, Batch, context caching) (ai.google.dev — 2026)

03 Gemini 3.5 Flash — карточка модели (вход 1 048 576, вывод 65 536) (ai.google.dev — 2026)

04 Gemini API — Release notes (релизы 3.6/3.7 Flash, Omni, embedding) (ai.google.dev — 2026)

05 Gemini API — Batch API и Context caching (режимы экономии) (ai.google.dev — 2026)

06 Vertex AI — Generative AI on Google Cloud (governance) (cloud.google.com — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.