Ай-ТИ Фреш

ТЕХНИЧЕСКИЙ РАЗБОР

AI-агент для разбора и приоритизации входящей почты

Сортировка, черновики ответов и утренняя сводка для секретаря и офис-менеджера

Июль 2026

itfresh.ru · ИТ-аутсорсинг для юридических лиц

О ЧЁМ ЭТОТ РАЗБОР

Суть проблемы

Секретарь получает 100+ писем в день: запросы клиентов, счета поставщиков, рассылки, юридически значимые требования. Ручная сортировка съедает часы, важное тонет в потоке. Мы разворачиваем ИИ-агента на Claude API: он читает почту по IMAP, классифицирует письма, готовит черновики ответов и присылает утреннюю сводку в Telegram — без автоотправки, решение за человеком.

Почему это важно бизнесу

- Секретарь тратит 2-3 часа в день на ручную сортировку вместо содержательной работы
- Критичное письмо (требование, дедлайн, VIP-клиент) теряется среди рассылок и рутины
- Ответ клиенту задерживается на часы, пока письмо ждёт своей очереди в общем списке
- Руководитель не видит картину по почте без утренней сводки приоритетов
- Рост штата не масштабируется — 100+ писем требуют либо второго секретаря, либо автоматизации

ЧТО МЫ ЗАКЛАДЫВАЕМ

Ключевые параметры реализации

100+

писем в день — порог, с которого ручная сортировка перестаёт масштабироваться

наш стандарт

5-15 мин

интервал опроса почтового ящика по IMAP для near-real-time классификации

наш стандарт

$1/$5

цена за 1M токенов Haiku 4.5 (вход/выход) — модель для массовой классификации писем

Claude API, прайсинг 2026

~90%

экономия на повторных токенах системного промпта при prompt caching

по докам Claude Prompt Caching

50%

скидка Message Batches API при ночной пакетной разборке накопившегося backlog

по докам Claude Batches API

128K

макс. токенов вывода при потоковой генерации черновика ответа (streaming)

Claude API, Opus 4.8/Sonnet 5

РЕАЛИЗАЦИЯ 1 ИЗ 3

Секретариат бухгалтерской компании: требования и дедлайны

Что настраиваем

офис бухгалтерского аутсорсинга, единый почтовый ящик на 4 сотрудников

Как мы это делаем

1. Разворачиваем IMAP-раннер на Python: опрос ящика раз в 10 минут, дедуп по Message-ID в SQLite
2. Классификация через Claude Haiku 4.5 по JSON-схеме: категория, приоритет, дедлайн, requires_reply
3. Ключевые слова эскалации (ФНС, СФР, «требование», «предписание») поднимают приоритет мгновенно
4. Для писем с requires_reply=true генерируем черновик на Sonnet 5 и кладём в папку «Черновики» через IMAP APPEND
5. В 8:00 отправляем в Telegram сводку: сколько писем, что срочное, что ждёт решения человека

РЕЗУЛЬТАТ

Секретарь открывает готовую сводку вместо 120 непрочитанных писем, критичные требования не тонут среди рассылок, черновики сокращают время ответа вдвое.

КЛЮЧЕВОЙ НЮАНС

Разделение классификатора (дёшево, часто) и генератора черновиков (дороже, только для важного) держит стоимость под контролем при росте объёма почты.

РЕАЛИЗАЦИЯ 2 ИЗ 3

Отдел продаж торговой компании: VIP-клиенты и приоритизация

Что настраиваем

коммерческий отдел торговой компании, общий ящик sales@, 150+ писем/день

Как мы это делаем

1. Загружаем список VIP-доменов и контактов в системный промпт с cache_control (ttl 1h)
2. Классификация помечает письма от VIP отдельным приоритетом независимо от содержания
3. Письма от новых лидов получают короткое авто-резюме и предложенный черновик первого ответа
4. Накопившийся backlog после выходных разбираем через Message Batches API с 50% скидкой
5. Журнал обработки (категория/приоритет/время) пишем в SQLite для еженедельного разбора руководителем

РЕЗУЛЬТАТ

Менеджеры реагируют на VIP-клиентов в течение часа вместо суток, backlog после выходных разбирается без переработок в понедельник утром.

КЛЮЧЕВОЙ НЮАНС

Прайм-кэширование системного промпта с правилами и VIP-списком — обязательное условие, иначе стоимость классификации растёт кратно на каждый опрос ящика.

РЕАЛИЗАЦИЯ 3 ИЗ 3

Регистратура медклиники: жалобы против рутинных обращений

Что настраиваем

регистратура частной клиники, почтовый ящик для пациентов и поставщиков

Как мы это делаем

1. Разделяем поток на категории: запись на приём, жалоба, поставщики, регуляторные письма
2. Жалобы пациентов эскалируются отдельным правилом и не ждут общей утренней сводки
3. Для рутинных запросов на запись готовим черновик с шаблонным ответом и свободными слотами
4. Вложения (сканы направлений) прогоняем через PDF-инструмент для извлечения текста перед классификацией
5. Настраиваем always_ask-подтверждение перед отправкой автоответа по чувствительным темам

РЕЗУЛЬТАТ

Жалобы пациентов получают ответ в течение часа вместо ожидания в общей очереди писем, регистратор занимается только предметными случаями.

КЛЮЧЕВОЙ НЮАНС

Для медицинского контекста автоотправка недопустима даже для рутинных писем — черновик и подтверждение человеком обязательны из-за PII и репутационных рисков.

ЧТО МЫ ВЫУЧИЛИ НА ПРАКТИКЕ

Подводные камни

- ✕ **Автоотправка без модерации**
  Черновики агент кладёт в папку «Черновики», финальную отправку всегда делает человек — иначе одна ошибка классификации улетает клиенту.

- ✕ **Один промпт без кэширования**
  Без cache_control системный промпт с правилами и VIP-списком пересчитывается каждый опрос ящика — стоимость растёт в разы без выгоды.

- ✕ **Нет дедупликации по Message-ID**
  Повторный IMAP-опрос без хранения обработанных ID заново классифицирует и переотправляет уже обработанные письма.

- ✕ **Классификация и черновик одним вызовом**
  Дорогая модель на каждое письмо вместо дешёвого классификатора для всех и дорогого генератора только для важных — лишние расходы.

- ✕ **Хранение вложений в памяти агента**
  Персональные данные и вложения, записанные в постоянную память, реплицируются в будущие сессии — риск утечки и нарушения 152-ФЗ.

- ✕ **Игнор stop_reason refusal/max_tokens**
  Без проверки stop_reason часть черновика тихо обрезается или отклоняется классификатором безопасности — ответ выглядит завершённым, но неполон.

- ✕ **Batch API для срочной почты**
  Message Batches экономит 50%, но результат приходит до часа — для срочных писем нужен обычный Messages API, Batches только для backlog.

- ✕ **Нет эскалации по ключевым словам**
  Без списка триггеров (ФНС, дедлайн, штраф) критичное письмо получает обычный приоритет и ждёт своей очереди в общей сводке.

УРОВНИ РЕАЛИЗАЦИИ

Как правильно

МИНИМУМ

- IMAP-опрос раз в 5-15 минут с дедупом по Message-ID в SQLite
- Классификация на Haiku 4.5 по JSON-схеме: категория, приоритет, дедлайн
- Утренняя сводка приоритетов в Telegram в фиксированное время
- Черновики только в папку «Черновики», без автоотправки

НОРМАЛЬНО

- Кэширование системного промпта (cache_control, ttl 1h) для экономии токенов
- Список VIP-отправителей и ключевых слов эскалации в реальном времени
- Структурированный вывод (output_config.format) для надёжного парсинга приоритета
- Журнал обработки писем в SQLite с метриками по категориям для аудита

ХОРОШО

- Пакетная разборка backlog через Message Batches API по ночам и выходным
- Managed Agents сессия с always_ask-подтверждением для нестандартных действий
- Мониторинг ошибок API (429/5xx) с алертом в Telegram при сбоях
- Версионирование системного промпта и правил приоритизации с возможностью отката

ПРОВЕРЬТЕ У СЕБЯ

Чек-лист самопроверки

- [ ] Настроен дедуп по Message-ID перед классификацией писем?
- [ ] Системный промпт с правилами приоритизации закэширован через cache_control?
- [ ] Есть список VIP-отправителей и ключевых слов эскалации?
- [ ] Черновики ответов пишутся только в папку «Черновики», без автоотправки?
- [ ] Утренняя сводка формируется и доставляется в фиксированное время?
- [ ] Ведётся журнал обработки писем (категория, приоритет, время) для аудита?
- [ ] Настроен алерт при ошибках API (429/5xx/сетевые сбои)?
- [ ] Учтён режим хранения данных — вложения и PII не хранятся дольше необходимого?
- [ ] Есть сценарий разбора backlog из 100+ писем через Batch API?
- [ ] Проверена обработка stop_reason=refusal и stop_reason=max_tokens?

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).

- Разворачиваем IMAP-триаж с классификацией на Haiku 4.5 и черновиками на Sonnet 5 под ключ
- Настраиваем эскалацию по ключевым словам и VIP-клиентам под специфику вашего бизнеса
- Внедряем утреннюю Telegram-сводку и журнал обработки писем для аудита
- Переводим систему на Batch API или Managed Agents при росте объёма почты
- Обеспечиваем безопасность: секреты вне промпта, минимизация хранения вложений и PII

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 Messages API — Tool Use и структурированные ответы (platform.claude.com — 2026)

02 Prompt Caching — кэширование системного промпта (platform.claude.com — 2026)

03 Message Batches API — пакетная обработка (platform.claude.com — 2026)

04 Managed Agents — Overview, Outcomes, Vaults (platform.claude.com — 2026)

05 Models Overview — Haiku 4.5, Sonnet 5, Opus 4.8 (platform.claude.com — 2026)

06 Наш стандарт почтовых раннеров (IMAP/Sbis/mailcow) (itfresh.ru — 2026)

Основано на официальной документации продуктов и нашей практике внедрения.