

0.15.2 → 0.16.0

8 октября 2026 · для администратора

Каждый токен на маршруте и на счету

Пул моделей с маршрутом на холсте, классификатором и переливом с обезличиванием. Тарифы с лимитами и запасной моделью, журнал каждого вызова, отчёт расхода и живой мониторинг.



Пулы

Маршрут,
классификатор,
перелив



Тарифы

Лимиты, ставки,
назначение



Учёт

История, аналитика,
деньги



Мониторинг

Слоты, очередь,
ошибки



6

шагов мастера тарифа

3

периода лимита: день, неделя,
месяц

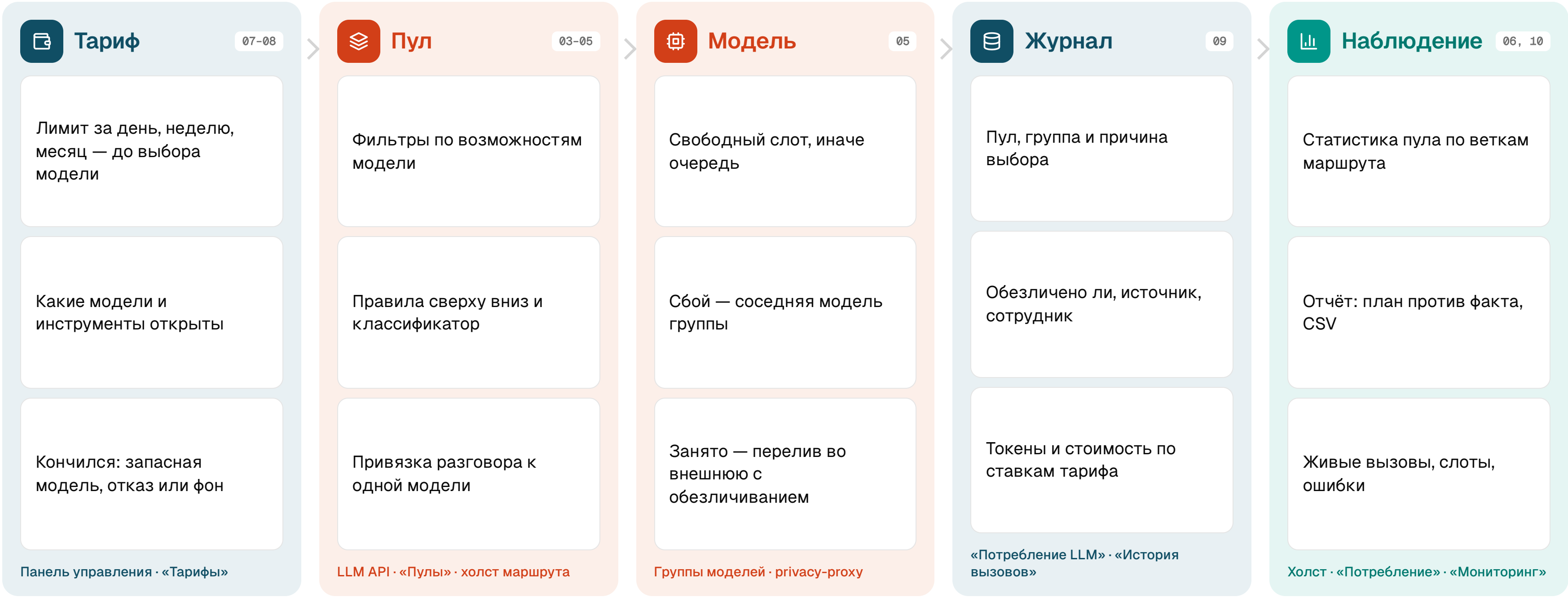
0

новых обязательных
переменных окружения

Каждый вызов модели проходит пять остановок

Тариф проверяется до выбора модели, пул выбирает модель, журнал записывает, куда и почему ушёл запрос.

откуда Чат Агент NoCode-агент Автоматизация OpenAI API Эмбединги баз знаний



● **Маршрут** — администратор LLM: пулы, группы, правила ● **Учёт** — администратор и право billing.manage: тарифы и деньги ● **Сотрудник** видит только свой остаток в токенах и вызовах

Пул — виртуальная модель среди подключений

Пул заводится в LLM API рядом с подключениями. Его выбирают как модель — в чате, у агентов и автоматизаций, в API и как эмбебдер базы знаний.

Панель управления · LLM API · Подключения

LLM API — Подключения

Управляйте подключениями и моделями через единый шлюз

Обновить

Подключение

Активные модели

20

из 23 всего

Запросов за 24ч

89

По всем моделям

Подключения

26

8 внешних · 18 агентских

Вызовы за 24 часа

gemma4:26b	31 (35%)
qwen3.8:27b	24 (27%)
qwen3-embedding	19 (21%)

Токены за 24 часа

qwen3.8:27b	734.2K (83%)
gemma4:26b	119.8K (14%)
baaibge-reranker-v2-m3	24.9K (3%)

Подключения и модели

Учётные данные на уровне подключения; модели группируются по connection_id

Все

Внешние

Агентские

1

Пулы моделей

Виртуальная модель: вызов уходит карточке со свободными слотами, при сбое — следующей

Корпоративная модель

corp-model

Чат

Активен

3

3 модели · 3 группы · 1 правило · привязка треда · за 24 ч: 31 вызов / 2 ошибки

4

Доступ

Public

5

Открыть маршрут

6

Создать пул

1 Вкладка «Пулы»
Рядом с внешними и агентскими подключениями

2 «Создать пул»
Имя для людей и идентификатор для API

3 Карточка пула
Тип «Чат», статус, сколько моделей, групп и правил, привязка треда

4 Нагрузка за сутки
Вызовы и ошибки прямо в списке

5 Доступ
Всем или выбранным — как у обычной модели

6 «Открыть маршрут»
Холст с группами, правилами и статистикой

7 Счёт — по реальной модели
Вызов пула учитывается у модели, которая ответила

В OpenAI-совместимом API — по идентификатору

```
POST /api/v1/chat/completions
{ "model": "corp-model", "messages": [ ... ] }
```

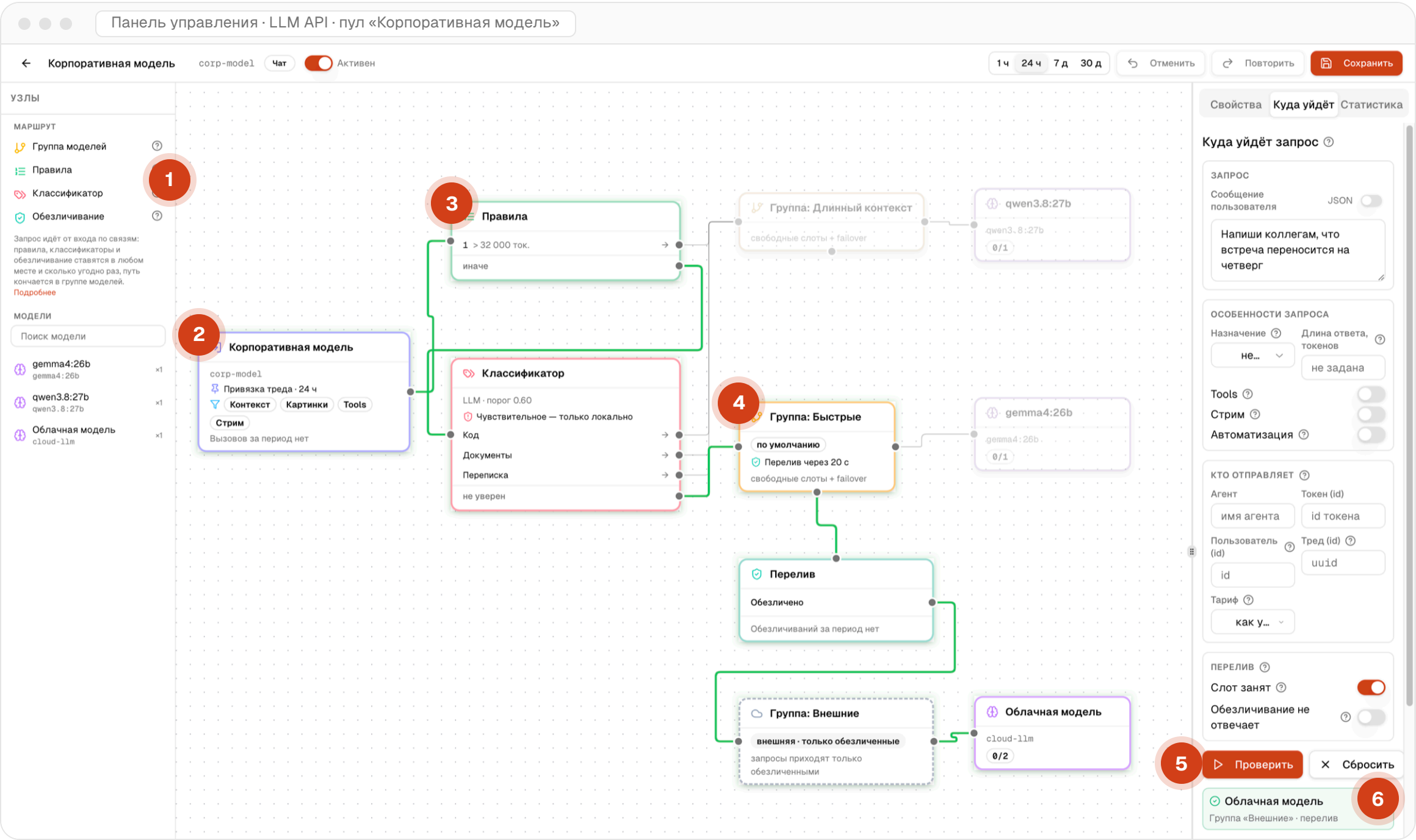
Клиенту не нужно знать, какие модели внутри и где они стоят

Панель управления · LLM API · «Пулы моделей»

Скриншот со стенда, адреса замаскированы

Маршрут рисуется на холсте

Запрос уходит в модель со свободным слотом — по приоритету, среди равных по очереди; при сбое вызов повторяется на соседней.



- Узлы маршрута**
Группа моделей, правила, классификатор и обезличивание — перетаскиваются на холст
- Вход пула**
Фильтры убирают модели, которым не подходит запрос: длинный контекст, картинки, инструменты, стрим
- Правила сверху вниз**
«В группу», «не в группу», «только локально» — в том числе по тарифу сотрудника
- Группы моделей**
Локальные и внешние, по приоритету или по очереди. Сбойная модель на время выводится из оборота
- «Куда уйдёт»**
Проверка маршрута без вызова модели: текст, агент, сотрудник, тариф
- Путь подсвечен**
Видно, по какой ветке прошёл запрос. «Статистика» — то же за период

Привязка чата

Разговор остаётся на одной модели, чтобы не терять кэш промпта. Каждый ход заново проходит правила — если история перестала помещаться, чат переезжает в группу с большим окном.

Классификатор знает тему, обезличивание — имена

Чувствительный запрос остаётся в локальных моделях, а во внешнюю уходит только текст с метками.

СВОЙСТВА

Классификатор

Название

Классификатор

Тип

1

LLM

Порог уверенности

0,6

Таймаут, мс

1500

Текст, символов

4000

Модель

gemma4:26b

КАТЕГОРИИ

2

1 Код → Длинный кон...

2 Документы → Длин...

3 Переписка → Быстр...

+ Добавить категорию

ЧУВСТВИТЕЛЬНЫЕ СВЕДЕНИЯ

3

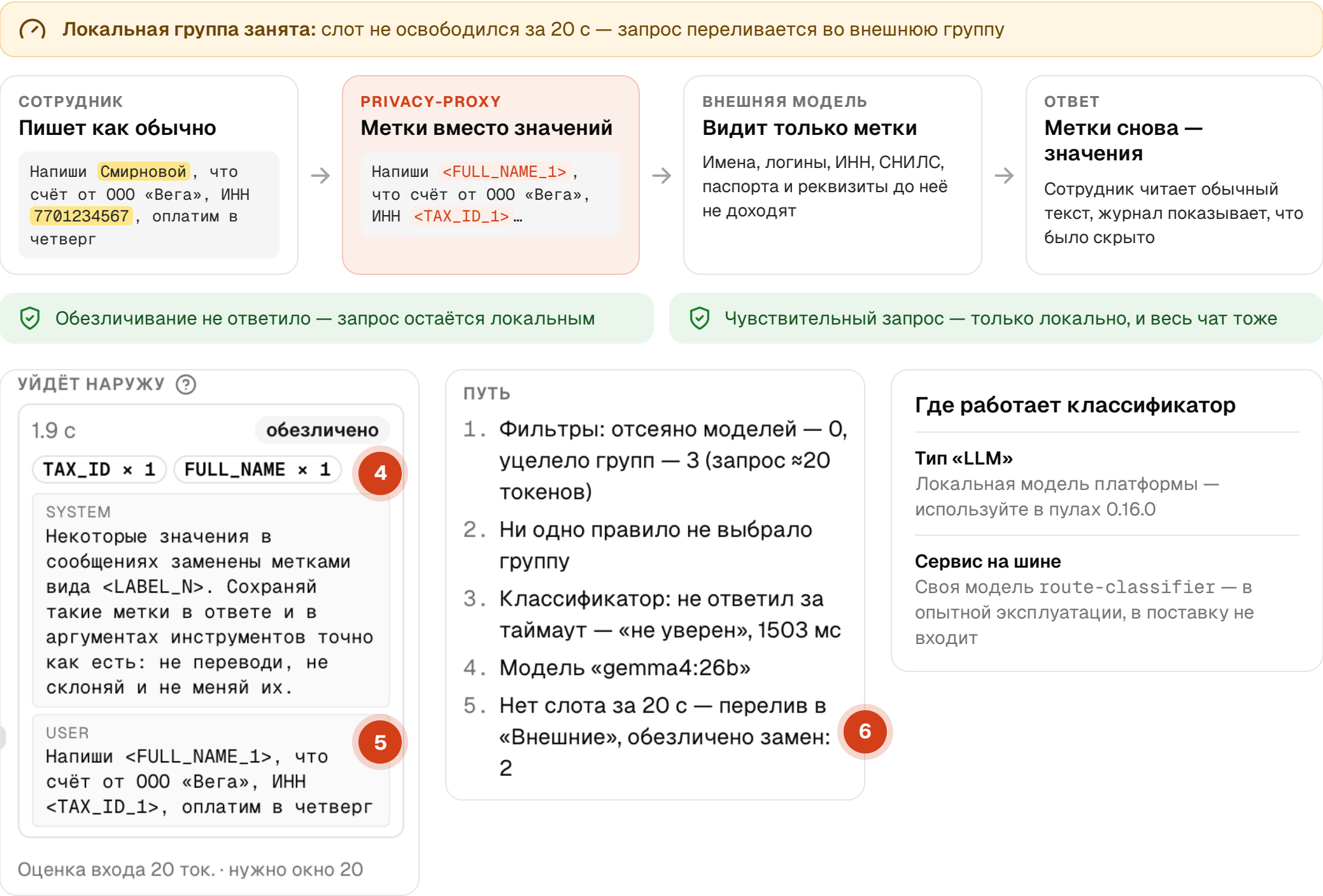
Описание

Персональные данные сотрудников и клиентов, реквизиты договоров, коммерческая тайна

Примеры, по одному в строке

Проверь паспортные данные Иванова

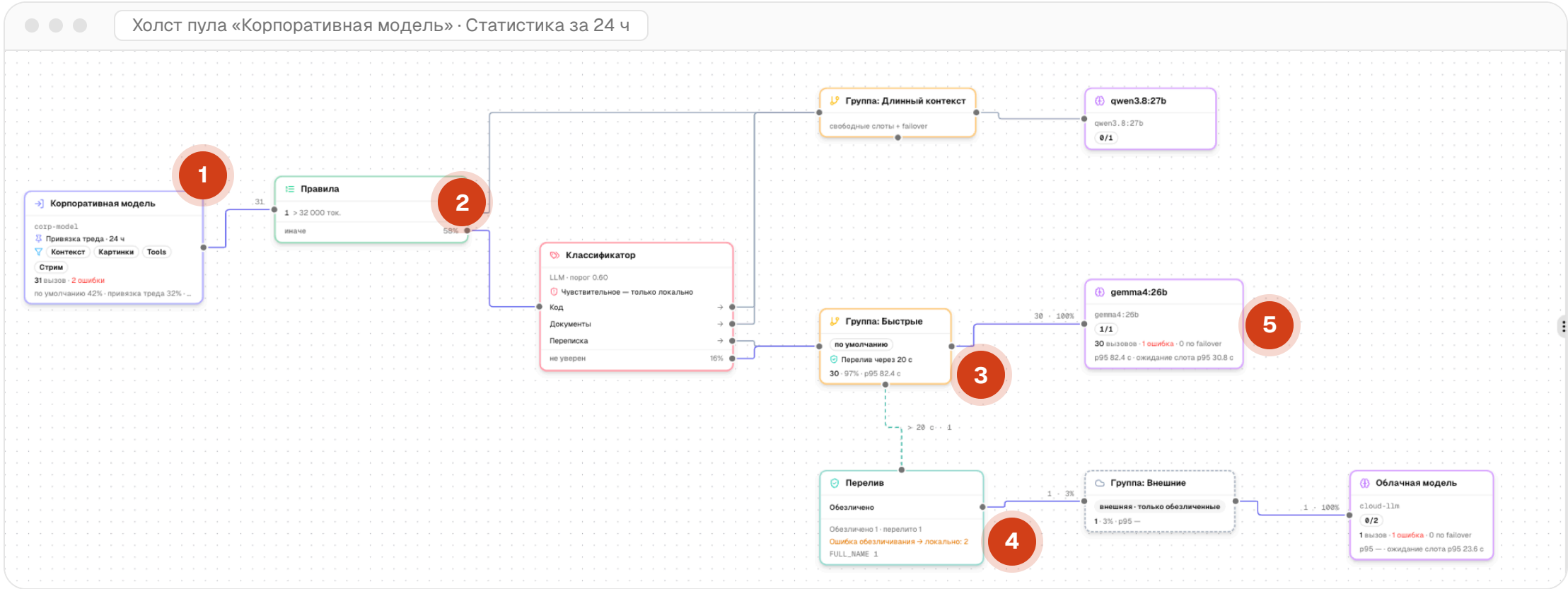
Сравни условия договора с ООО «Ромашка»



- 1 Тип и порог уверенности
- Ниже порога или нет ответа за таймаут — «не уверен», группа по умолчанию
- 2 Категории → группы
- «Код», «Документы», «Переписка» — с примерами и целевой группой моделей
- 3 Чувствительные сведения
- Описание и примеры: персональные данные, реквизиты, коммерческая тайна
- 4 Что скрыто
- «Куда уйдёт» показывает метки: TAX_ID × 1, FULL_NAME × 1
- 5 Что уйдёт наружу
- Текст с метками и инструкция модели сохранять их как есть
- 6 Почему так
- Путь по шагам: фильтры, правила, классификатор, перелив

Видно, по какой ветке идут запросы

Вкладка «Статистика» накладывает реальные вызовы на холст: доли по веткам, задержки, ошибки и причины выбора.



- 1 **Вход пула**
Вызовы, ошибки и главные причины выбора
- 3 **Группа**
Вызовы, доля и p95 задержки
- 5 **Модель**
Ошибки, failover и ожидание слота p95
- 7 **Группы**
Доля, токены, p50/p95, слот p95
- 9 **Классификатор**
Сколько раз спрошен, «не уверен», время

- 2 **Доля по ветке**
Сколько запросов прошло по каждому правилу
- 4 **Перелив**
Сколько обезличено и сколько осталось локально из-за ошибки
- 6 **Период**
24 часа, 7 и 30 дней или свой
- 8 **«Почему туда»**
Правило, привязка треда, «не уверен», перелив
- 10 **Модели**
Кто реально отвечал и как часто сбоил

СвойстваКуда уйдёСтатистика

6 24 ч 7 д 30 д Свой

32 Вызовов 2 Ошибок 122.8K Токенов

ПО ДНЯМ И ГРУППАМ

32 24 16 8 0 09.10

Быстрые Внешние

ГРУППЫ

Быстрые 31 · 97%
ошибок 3% · токены 104.6K / 17.3K · p50/p95 18.9 с / 81.9 с · слот p95 29.6 с

Внешние 1 · 3%
ошибок 100% · токены 941 / 0 · p50/p95 — / — · слот p95 23.6 с

ПОЧЕМУ ТУДА

по умолчанию 8 14 · 44%
привязка треда 10 · 31%
не уверен 5 · 16%
ошибка обезличивания → локально 2 · 6%
перелив 1 · 3%

КЛАССИФИКАТОР

Спрошен 9 15 из 29
Не уверен: нет ответа 33% · низкая уверенность 0%

Чувствительных 0 · 0%
Время p50 / p95 1.5 с / 1.5 с

МОДЕЛИ

gemma4:26b · Быстр... 31 · 97% 10
ошибок 1 · failover 0 · p95 81.9 с · слот p95 29.6 с
Облачная модель · Вне... 1 · 3%
ошибок 1 · failover 0 · p95 — · слот p95 23.6 с

Тариф собирается мастером из шести шагов

Какие модели и инструменты доступны, сколько можно потратить и что будет, когда лимит кончится.

Тарифы · новый тариф · шаг 2

Новый тариф

Какие модели входят в тариф, их лимиты и ставки

Тариф

2 Модели и пулы

3 Лимит кончился

4 Инструменты

5 Где доступен

6 Проверка

Все модели и пулы

Включая будущее — без лимитов и по нулевым ставкам

Выбрать из каталога

Лимит и ставки для каждой модели и пула

В тарифе: 2

Добавить модель

Корпоративная модель

Пул · Чат

Лимит токенов

С лимитом

в день200000

в неделю800000

в месяц2500000

200 тыс. / день · 800 тыс. / нед · 2,5 млн / мес токенов

Вход, за 1 млн

30₽

Из кэша, за 1 млн

3₽

Выход, за 1 млн

90

qwen3.8:27b

Модель · Чат

Лимит токенов

Без лимита

Вход, за 1 млн45₽

Из кэша, за 1 млн

как вход₽

Выход, за 1 млн135₽

Назад

Шаг 2 из 6

Далее

Когда лимит кончился

Корпоративная модель

200 тыс. / день · 800 тыс. / нед · 2,5 млн / мес

Другая модель

Ответит запасная модель этого тарифа того же типа

Отказ

Запросы отклоняются до следующего периода

Фоновый режим

Запросы выполняются, но в фоне и медленнее

Запасная модель

qwen3.8:27b

Сотрудник увидит: Ответит «qwen3.8:27b»; кончится и её лимит — отказ до следующего периода

Лимиты вызовов: 1

web-search

web_search

в день50

в неделю—

в месяц500

50 / день · 500 / мес вызовов

Добавить лимит

- 1

Все модели или каталог

Из каталога — лимит и ставки у каждой модели и пула
- 2

День, неделя, месяц

Лимит в токенах; без лимита — только ставки
- 3

Ставки и абонентская плата

За 1 млн токенов: вход, из кэша, выход
- 4

Когда лимит кончился

Запасная модель тарифа, отказ или фоновый режим
- 5

Лимиты MCP-инструментов

Число вызовов у инструмента или всего сервера
- 6

«Сотрудник увидит»

Тариф глазами сотрудника ещё до сохранения

Сотрудник увидит

Ваш тариф

Стандарт

Корпоративная модель и веб-поиск с дневным лимитом — для большинства сотрудников

Модели и пулы

Корпоративная модель

200 тыс. / день · 800 тыс. / нед · 2,5 млн / мес

Лимит кончится: Ответит «qwen3.8:27b»; кончится и её лимит — отказ до следующего периода

qwen3.8:27b

без лимита

Инструменты

web-search · web_search

50 / день · 500 / мес

Назад

Шаг 6 из 6

Создать тариф

Кому достаётся тариф

Организации

Тариф открыт её сотрудникам; есть тариф по умолчанию

Подразделению

Вместе с поддеревом

Сотруднику

Действует последнее назначение

Тарификация включена

до включения лимиты не действуют

Панель управления · «Тарифы»

Скриншоты со стенда, тариф «Стандарт» и ставки условные

Сотрудник видит остаток лимита

Кольца в шапке Workspace — расход за день, неделю и месяц; подробности и выбор тарифа — в «Потреблении».

1

Тариф «Стандарт»

День

сброс через 7 ч

Корпоративная модель

2

72%

144 тыс. из 200 тыс. токенов

web-search · web_search

46%

23 из 50 вызовов

Неделя

сброс пн, 12.10, 00:00

Корпоративная модель

35%

280 тыс. из 800 тыс. токенов

Месяц

сброс вс, 01.11, 00:00

Корпоративная модель

30%

750 тыс. из 2,5 млн токенов

web-search · web_search

37%

187 из 500 вызовов

Весь расход →

3

Workspace · Настройки · Потребление

6

1 Кольца в шапке

День, неделя и месяц — по самому загруженному лимиту

4 Выбор тарифа

Из открытых организации сотрудника

2 Предупреждение заранее

Цвет меняется по мере расхода; рядом — когда сброс

5 Лимиты инструментов

Вызовы web_search рядом с токенами моделей

3 «Весь расход»

Переход в «Настройки» → «Потребление»

6 Что будет дальше

«При исчерпании: дальше — qwen3.8:27b»

Workspace · шапка и «Потребление» · денег сотрудник не видит, только токены и вызовы

Скриншоты со стенда, расход условный

В журнале — куда и почему ушёл каждый вызов

История вызовов показывает пул, группу и причину выбора; аналитика и отчёт сводят расход по людям, моделям и тарифам.

Потребление LLM · История вызовов

История вызовов

1

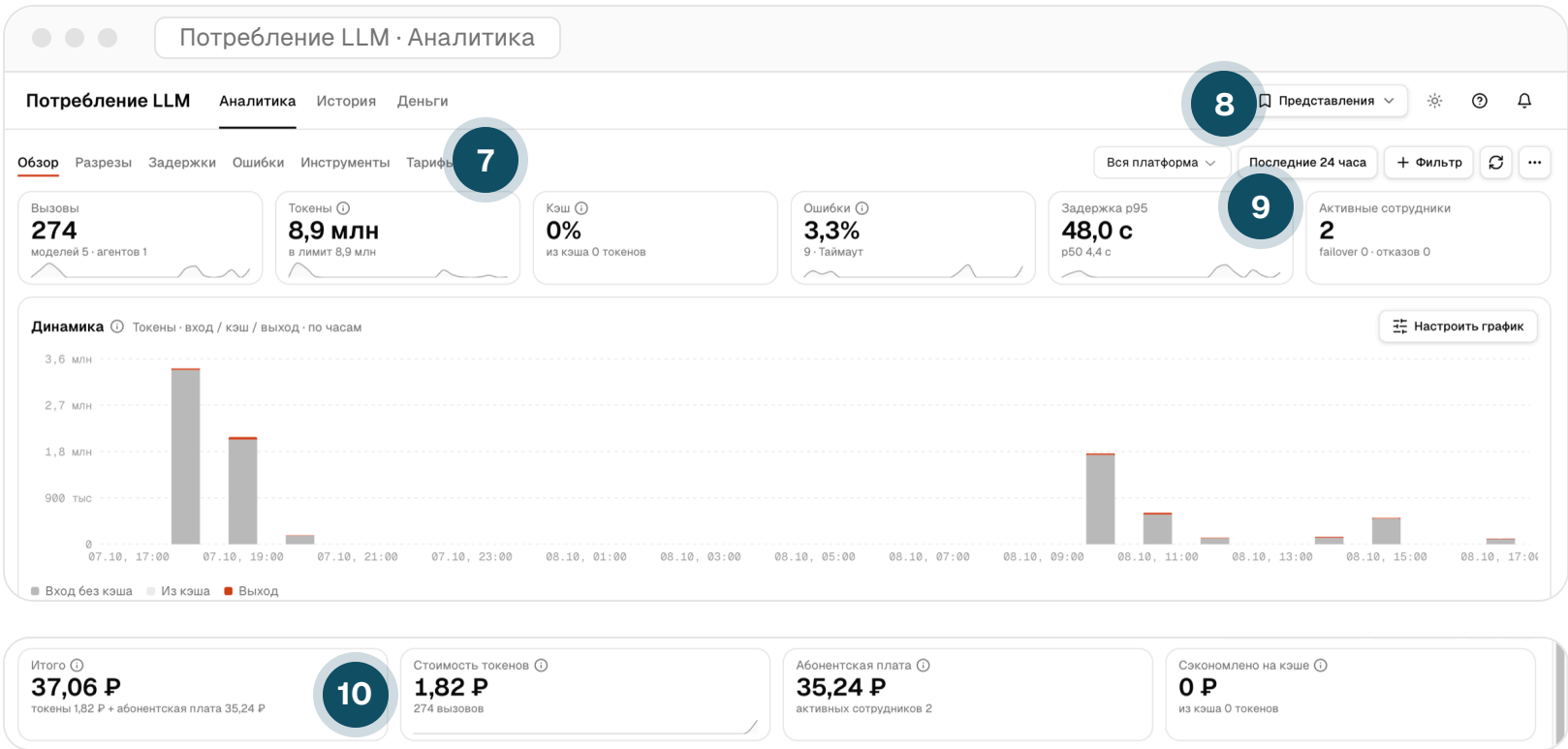
Колонки

2

CSV

1–50 из 90 · стр. 1 из 2

ID	Время (точное)	Провайдер	2	Модель	Пул	Группа моделей	Причина	Обезличивание	Источник	Пользователь	База знаний
49315	09.10 12:19:32	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			nc-yurist-po-dogovoram	Ирина Соколова	—
49314	09.10 12:19:12	OpenAI @ llm.company.local/api		qwen3-embedding					API	Ирина Соколова	—
49313	09.10 12:19:11	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49312	09.10 12:13:26	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49311	09.10 12:08:08	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49310	09.10 12:07:58	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49309	09.10 12:07:49	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49308	09.10 12:06:16	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49307	09.10 12:05:35	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			API	Ирина Соколова	—
49306	09.10 12:05:29	cloud-llm		Облачная модель	Корпоративная модель	Внешние	перелив	обезличен	API	Ирина Соколова	—
49304	09.10 12:03:15	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			nc-yurist-po-dogovoram	Ирина Соколова	—
49303	09.10 12:03:03	OpenAI @ llm.company.local/api		gemma4:26b	Корпоративная модель	Быстрые			nc-yurist-po-dogovoram	Ирина Соколова	—
49302	09.10 12:02:50	OpenAI @ llm.company.local/api		qwen3-embedding					nc-yurist-po-dogovoram	Ирина Соколова	—



- 1

Колонки и CSV

Свой набор столбцов, выгрузка
- 2

Фильтр в заголовке

По провайдеру, модели, пулу, причине
- 3

Пул и группа

Через какой пул и группу прошёл вызов
- 4

Прямой вызов

Модель выбрана без пула — колонка пустая
- 5

Причина и обезличивание

«перелив», «обезличено» — прямо в строке
- 6

Источник

Агент, чат или API — и сотрудник
- 7

Шесть видов аналитики

Обзор, разрезы, задержки, ошибки, инструменты, тарифы
- 8

Представления

Фильтры и период — одним выбором
- 9

Отчёт

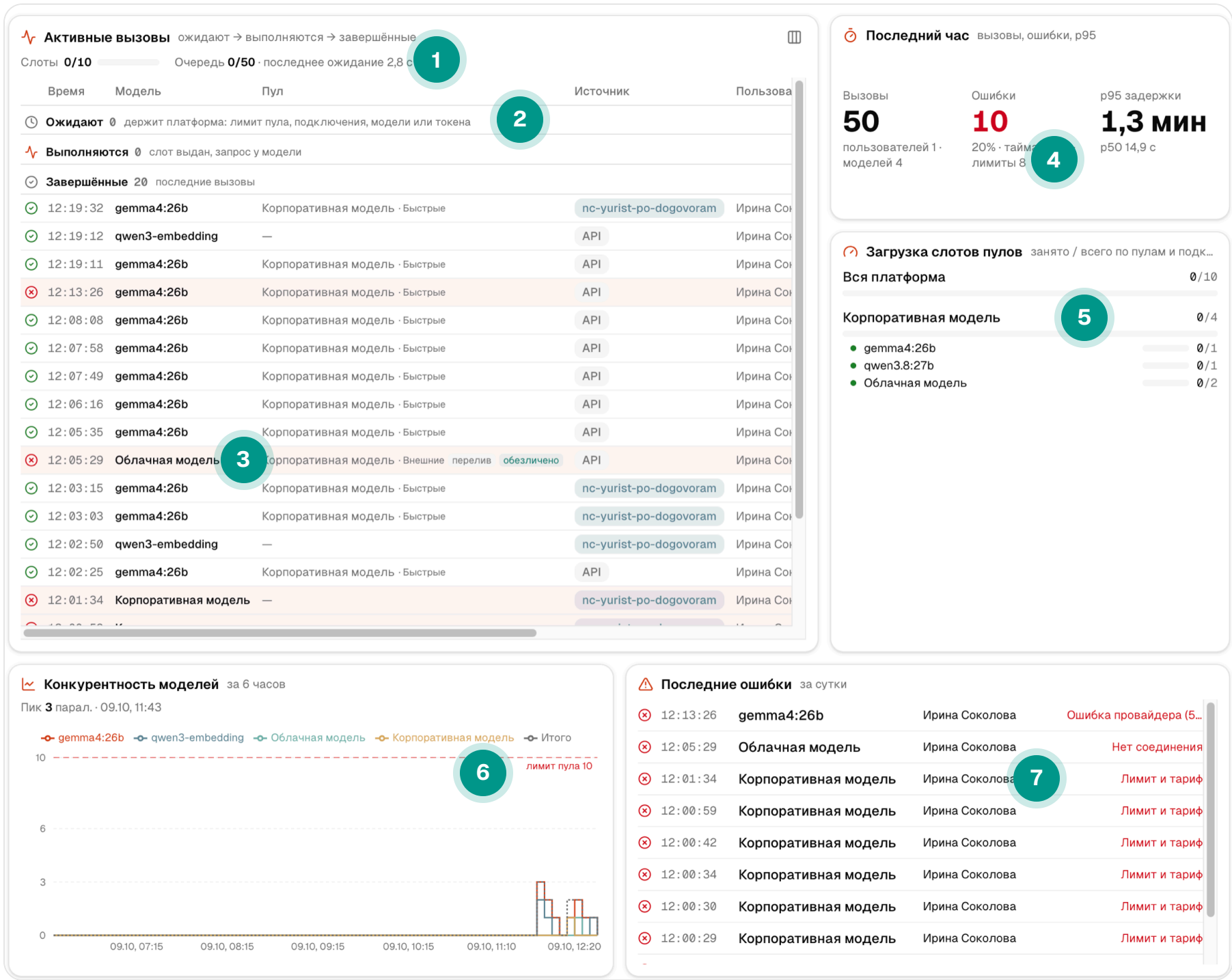
План против факта по организациям, подразделениям, людям; CSV
- 10

«Деньги»

Токены, абонплата, экономия на кэше — только с billing.manage

Что с моделями прямо сейчас

Живые вызовы с пулом и сотрудником, загрузка слотов, конкурентность и свежие ошибки — в одном дашборде.



- 1 Слоты и очередь**
Сколько вызовов идёт сейчас и сколько ждёт; последнее ожидание
- 2 Ожидает → выполняется → завершён**
Ждущий держит платформа: лимит пула, подключения, модели или тариф. У ждущего первого токена — идущий таймер
- 3 С пулом и группой**
Перелив и обезличивание видны в строке вызова
- 4 Последний час**
Вызовы, ошибки по причинам, p95 и p50 задержки
- 5 Загрузка слотов пулов**
Занято из доступного — по пулу и по каждой модели
- 6 Конкурентность моделей**
Пик параллельных вызовов против лимита пула
- 7 Последние ошибки**
«Лимит и тариф» — сотрудник упёрся в тариф; рядом ошибки провайдера

Свой экран

Плашки переставляются; раскладки личные и общие, есть режим для ТВ. Графики GPU и VRAM — по каждой карте сервера

Обновиться и включить по шагам

После обновления тарификация выключена, а пулов нет: всё заводится вручную и проверяется до того, как заработают лимиты.

ПРИ ОБНОВЛЕНИИ

**Резервная копия базы**
Миграция безвозвратно удаляет служебные записи агентного чата из журнала

**Дождаться миграций**
Индекс поиска строки на большой базе строится долго — ждать Alembic migrations applied

**privacy-проху — по желанию**
Профиль privacy, каталог config/privacy/, переменные PRIVACY_*

**Обязательных переменных нет**
Порог сжатия и таймаут обезличивания — во вкладке «Платформа»

1

Собрать пул
LLM API → «Пулы» → «Создать пул». Группы: быстрые локальные, с длинным контекстом, внешние

 Холст маршрута

2

Настроить маршрут
Фильтры, правила сверху вниз, классификатор типа «LLM» с категориями и чувствительными сведениями

 Узлы холста

3

Проверить «Куда уйдёт»
Текст, агент, сотрудник, тариф — путь без вызова модели. Связь с обезличиванием: «Платформа» → «Проверить»

 Без него перелива нет

4

Завести тарифы
Мастер из шести шагов: модели и пулы, лимиты, ставки, что делать при исчерпании, инструменты

 «Сотрудник увидит»

5

Назначить и включить
Организации, подразделению с поддеревом или сотруднику — и переключатель тарификации

 До включения лимиты не действуют

6

Наблюдать
Статистика пула, история вызовов, отчёт «план и факт», «Мониторинг» на ТВ

 Ошибки «Лимит и тариф»