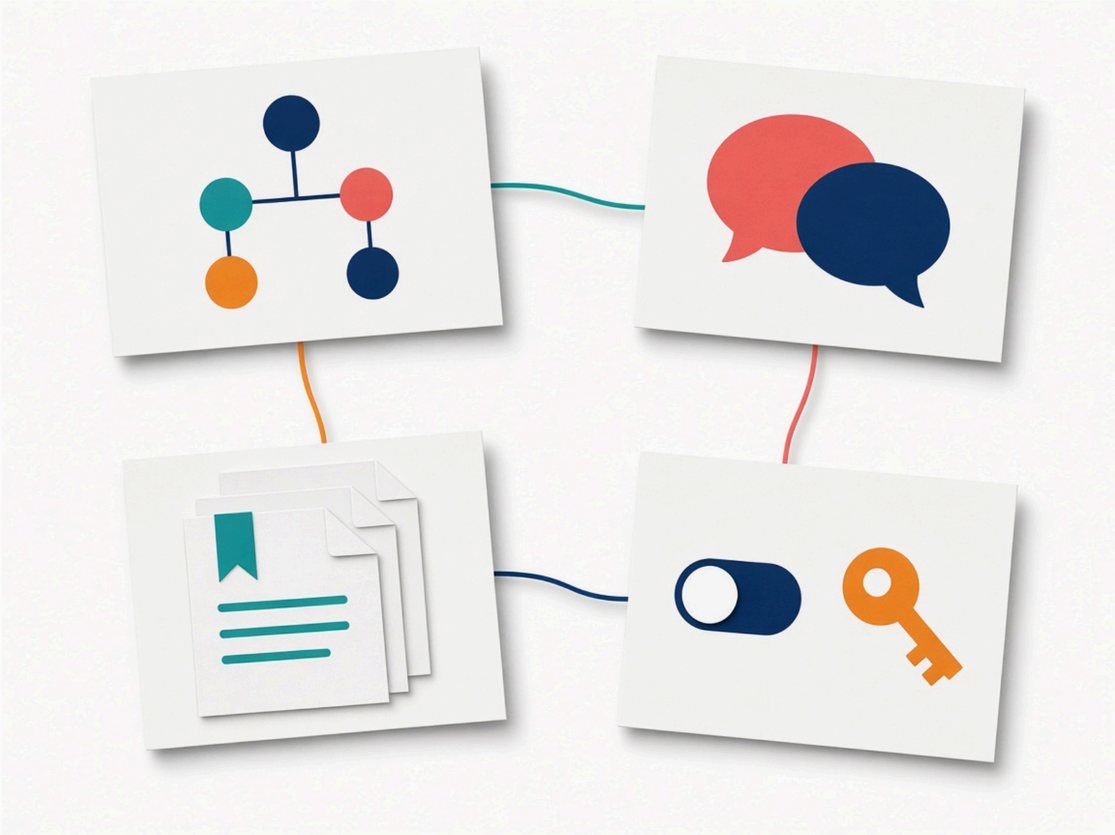
Cognitum AI Platform

версия 0.16.0

Платформа AI-агентов внутри вашего контура

Сотрудники работают в своём интерфейсе. Знания, агенты, модели и права остаются под контролем компании.



Рабочее место
Чат с историей, файлами и базами знаний

Знания
Ответ со ссылкой на документ

Агенты
В конструкторе и на Python SDK

Контроль
Права, тарифы, аудит и расход

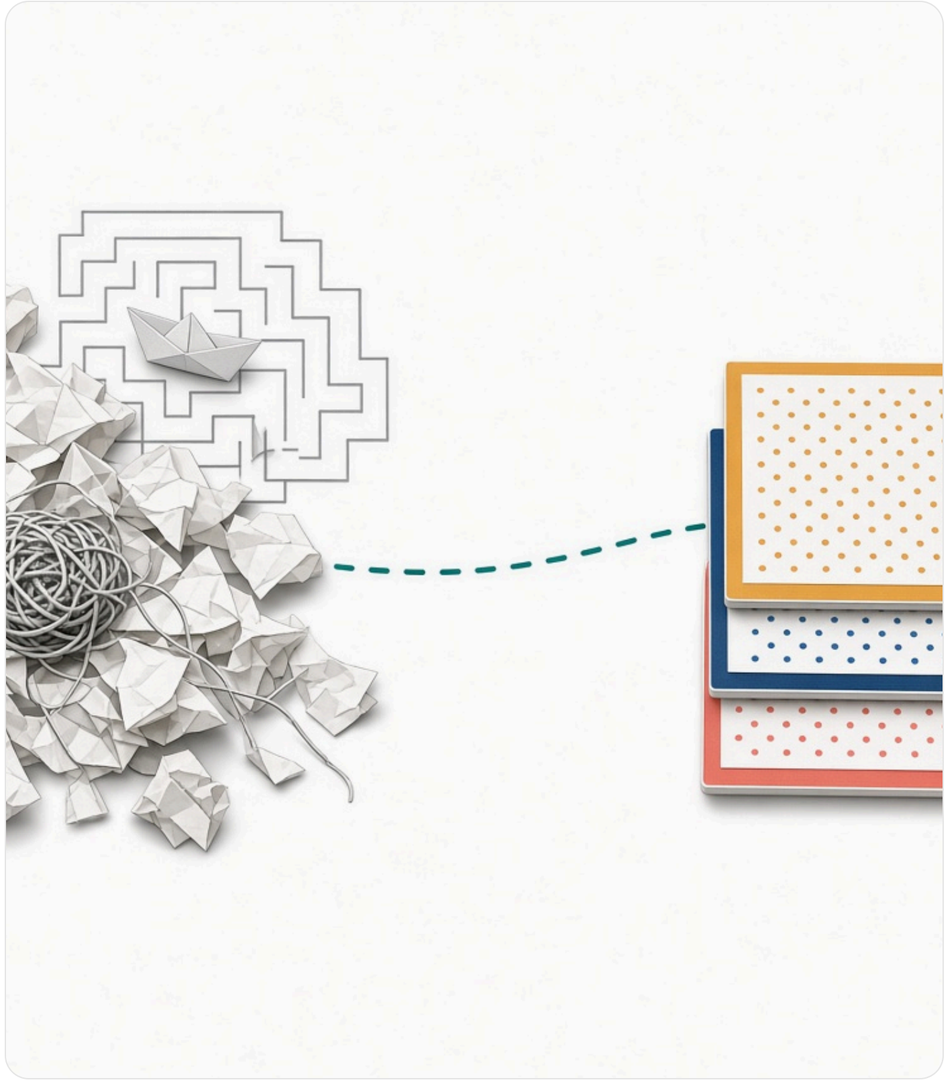
on-prem
данные не покидают периметр

1 API
для своих и внешних моделей

дни
до первого агента, не месяцы

Пилот застревает там, где нет контура и знаний

Генеративный ИИ в компаниях уже пробовали. В рабочие процессы он чаще всего так и не встроился.



89% проектов так и остаются пилотами

Демо впечатляет, а массовому процессу нужны качество, права и источник ответа.

73% компаний упираются в данные

Регламенты лежат в файлах, вики и переписке. Агенту не на что опереться.

6–18 мес уходит на свою инфраструктуру

Очереди, права, учёт токенов и интерфейс пишут заново под каждый проект.

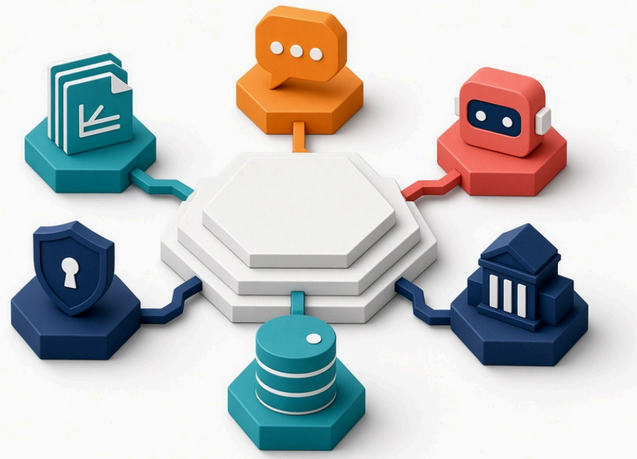
Cognitum закрывает эту обвязку

Рабочее место, знания, агенты и контроль уже в продукте. Команда занимается задачей, а не платформой.

Оценки российского рынка генеративного ИИ, 2026

Одна платформа между людьми, моделями и системами

Всё, что нужно агенту в компании, собрано в одном месте и подчиняется одним правилам.



Сотрудник видит чат. Администратор — какая модель ответила, по каким документам и за чей счёт.

КТО ПРИХОДИТ

- **Сотрудники** в чате
- **Руководители** расход и доступ
- **Telegram и Rocket.Chat**
- **Свои приложения** OpenAI API

COGNITUM

- Рабочее место
- Агенты
- Базы знаний
- Сценарии
- Автоматизации
- Пулы моделей
- Шлюз MCP
- Права и аудит
- Тарифы и расход

К ЧЕМУ ПОДКЛЮЧАЕТСЯ

- **Локальные модели** vLLM, Ollama
- **Внешние API**
- **Jira и Confluence**
- **S3, SMB, WebDAV**
- **LDAP / AD**
- **MCP-серверы компании**

Четыре роли — и у каждой свой результат

Один продукт отвечает на разные вопросы. Ниже то, что человек получает, а не список модулей.

Сотрудник

- Спрашивает модель или агента в своём чате, с историей и файлами
- Получает ответ со ссылкой на документ, а не пересказ «из воздуха»
- Просит модель найти в базе, свести в таблицу и сохранить туда же
- Собирает ассистента отдела и ставит сводку по расписанию

Ответ с источником в одном окне — Рабочее место, Базы знаний

Руководитель

- Доступ к знаниям выдаётся подразделению, а не списку логинов
- Отделу назначается тариф с лимитами, расход сверяется с планом
- Перевод сотрудника не требует собирать группу заново

Видно, кому открыто и сколько стоит — Оргструктура, Тарифы

Администратор и ИБ

- Модели, инструменты и веб-поиск включаются явно
- Наружу уходит только обезличенное, чувствительное — в локальные модели
- Аудит действий и журнал вызовов моделей остаются в контуре

Ничего не включается само — Права и роли, Пулы, Аудит

Разработчик и интегратор

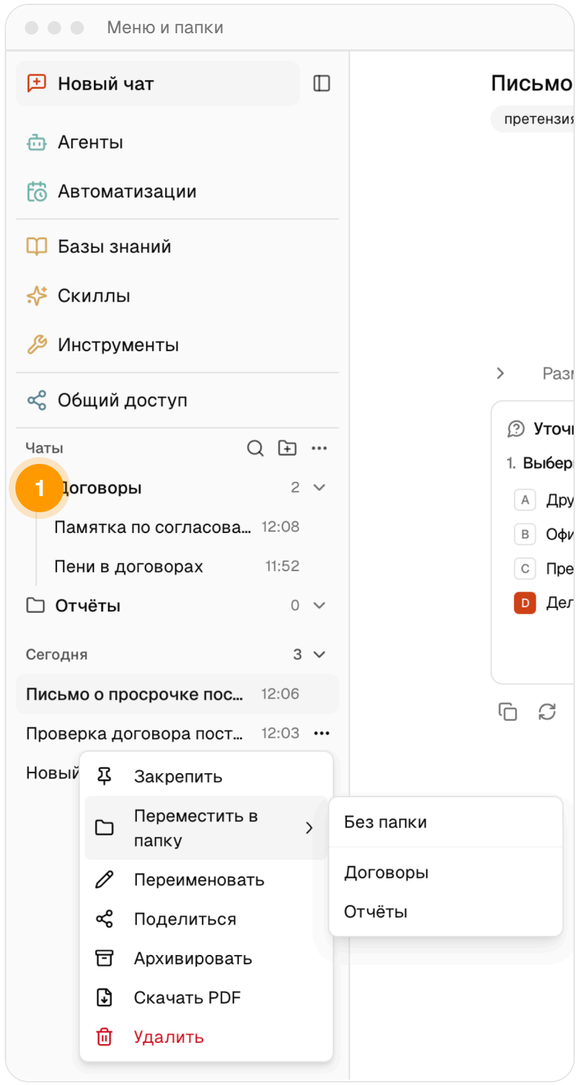
- Python SDK: задачи, чат и инструменты — без своей шины и учёта
- Инструменты баз знаний — для своей модели внутри агента
- Свой блок сценария добавляется без пересборки платформы

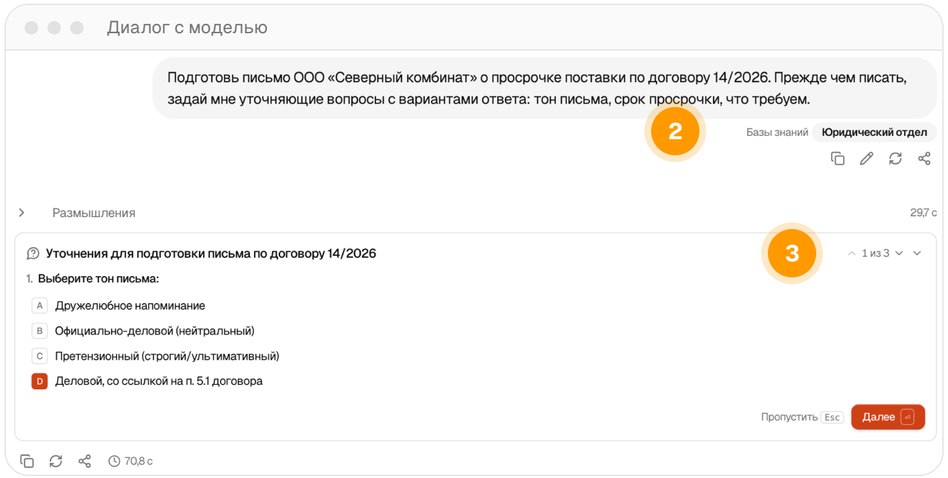
Код про задачу, а не про обвязку — SDK, Сценарии, OpenAI API

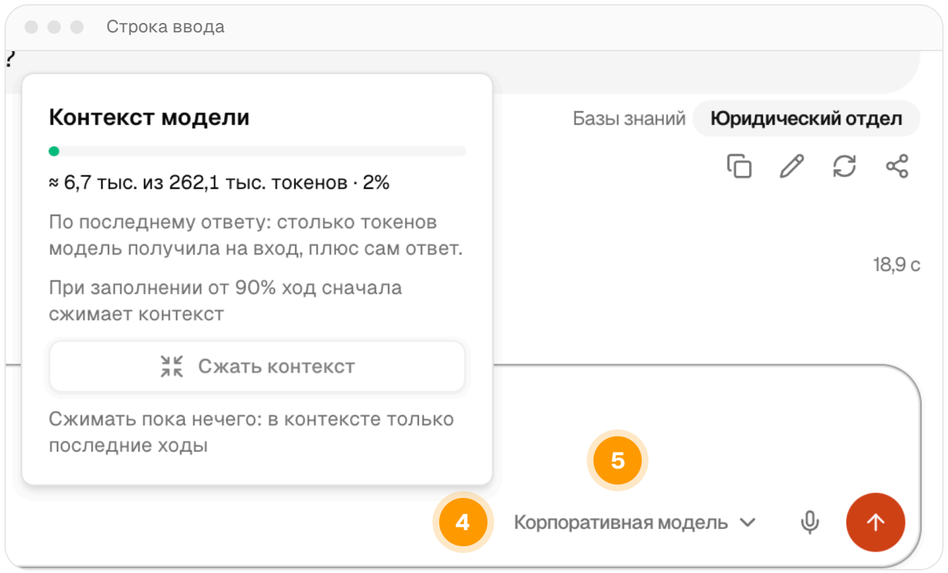
Вход через корпоративный IdP (OIDC/SAML) — в плане. Сейчас локальные учётки и LDAP

Сотрудник работает в своём чате, а не в чужом

История, файлы, базы знаний и настройки ответа живут в Cognitum. Сторонний чат для этого больше не нужен.







1 Одно меню и свои папки

Чаты раскладываются по папкам, меню сворачивается в рейку. Поиск, архив, PDF всего чата.

2 Файлы, базы и справка

База подключается к диалогу кнопкой. На вопрос о платформе модель отвечает по справке этой версии.

3 Сначала уточнит

Вопросы по одному, с вариантами. Всегда можно ответить своими словами.

4 Длинный разговор не обрывается

Кольцо показывает заполнение окна. Начало разговора сжимается в конспект.

5 Одна понятная модель

«Корпоративная модель» сама выбирает, кто ответит. Остаток лимита — в шапке.

База отвечает со ссылкой на документ

Если ничего не прошло порог релевантности, поиск честно возвращает пусто и не подмешивает слабые совпадения.

01 · ИСТОЧНИКИ
Откуда берутся
Файлы, S3, SMB, WebDAV, проект Jira, пространство Confluence

→

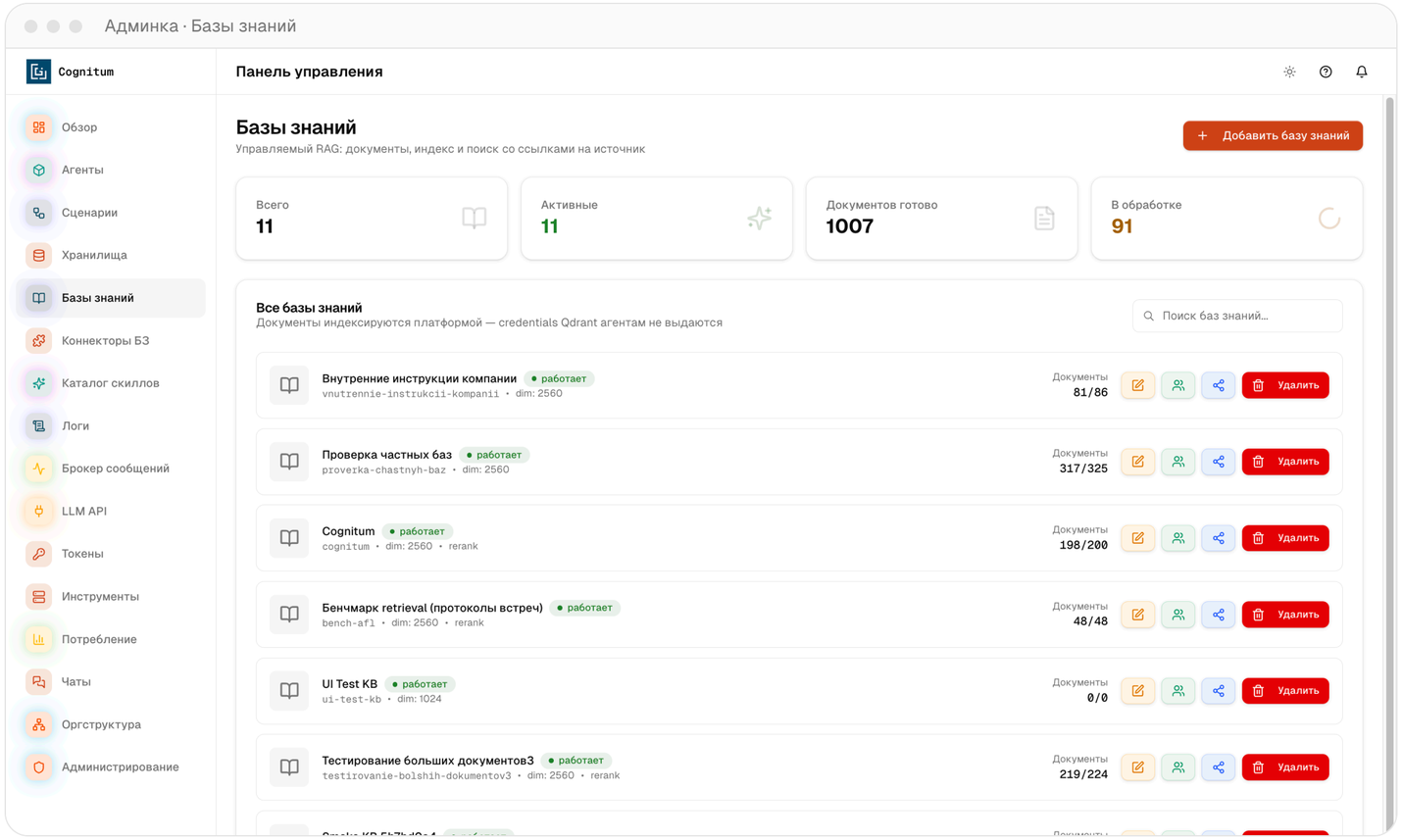
02 · РАЗБОР
Как читаются
Страницы, таблицы, OCR. Тяжёлое можно увести в ночное окно

→

03 · ПОИСК
Как находятся
Вектор и полнотекст, реранк, теги и порог

→

04 · ОТВЕТ
Кто спрашивает
Сотрудник в чате, агент и внешний API — одни базы



КАК ВЫГЛЯДИТ ОТВЕТ

Сколько дней согласуется договор поставки?

Договор поставки согласуется за три рабочих дня: юрист проверяет его в течение двух [1], финансовый контроль — в течение одного [2].

1	**Регламент закупок.pdf**	стр. 4
2	**Confluence · Финконтроль**	раздел 2

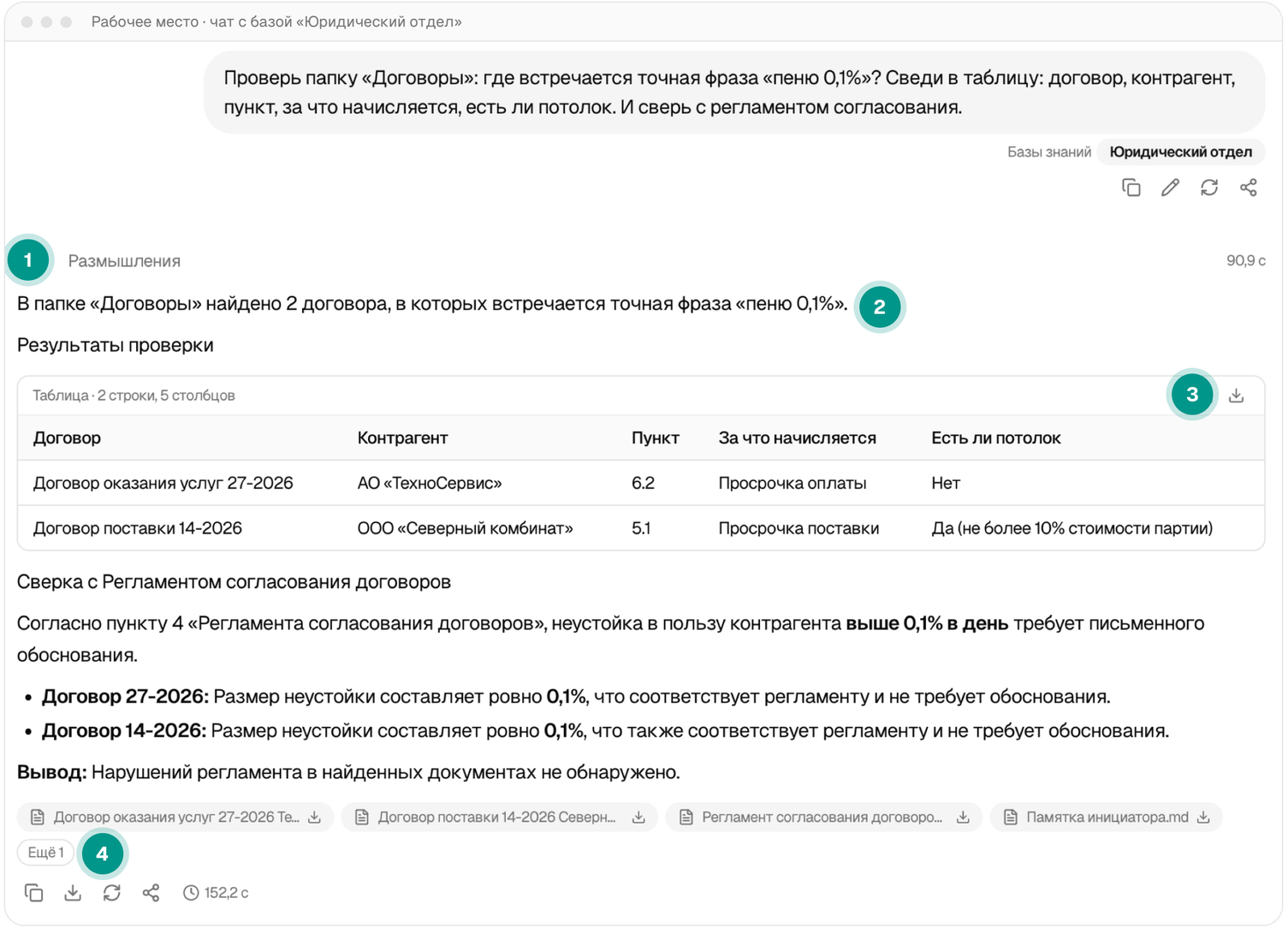
КАК ИСКАЛИ: вектор + полнотекст · реранк · порог

Пустая выдача — тоже ответ. Если совпадений выше порога нет, модель не придумывает источник.

Скриншот со стенда · пример ответа условный · смена модели эмбеддингов не останавливает поиск

Модель сама ищет, читает и пишет в базе

Не только найденные фрагменты: модель листает папки, ищет точную фразу, читает документ целиком и сохраняет итог.



1 Шаги на виду

Какой инструмент, в какой базе и с какой фразой — в «Размышлениях».

2 Точная фраза, а не похожая

Поиск по строке находит «пеню 0,1%» во всех договорах папки.

3 Таблица — сразу в Excel

Копирование, CSV или Markdown прямо из ответа.

4 Итог — обратно в базу

Памятку модель кладёт в нужную папку. Запись — по правам, удаление — только с подтверждением.

ВОСЕМЬ ИНСТРУМЕНТОВ

по смыслу · список · точная строка · сведения · целиком · **добавить** · **заменить** · **удалить**

Скриншот со стенда, договоры вымышлены · тот же набор — у агентов без кода, в SDK и по MCP

Ассистента отдела собирают без разработчика

Типовой агент — это модель, инструкция и то, чем он оснащён. С ним сразу можно говорить в чате.

Модель + Инструкция + Базы знаний + Инструменты + Скиллы + Память

1 Скилл — приём, а не простыня

Короткая инструкция под задачу. Полный текст агент подгружает, только когда он нужен.

2 Память между диалогами

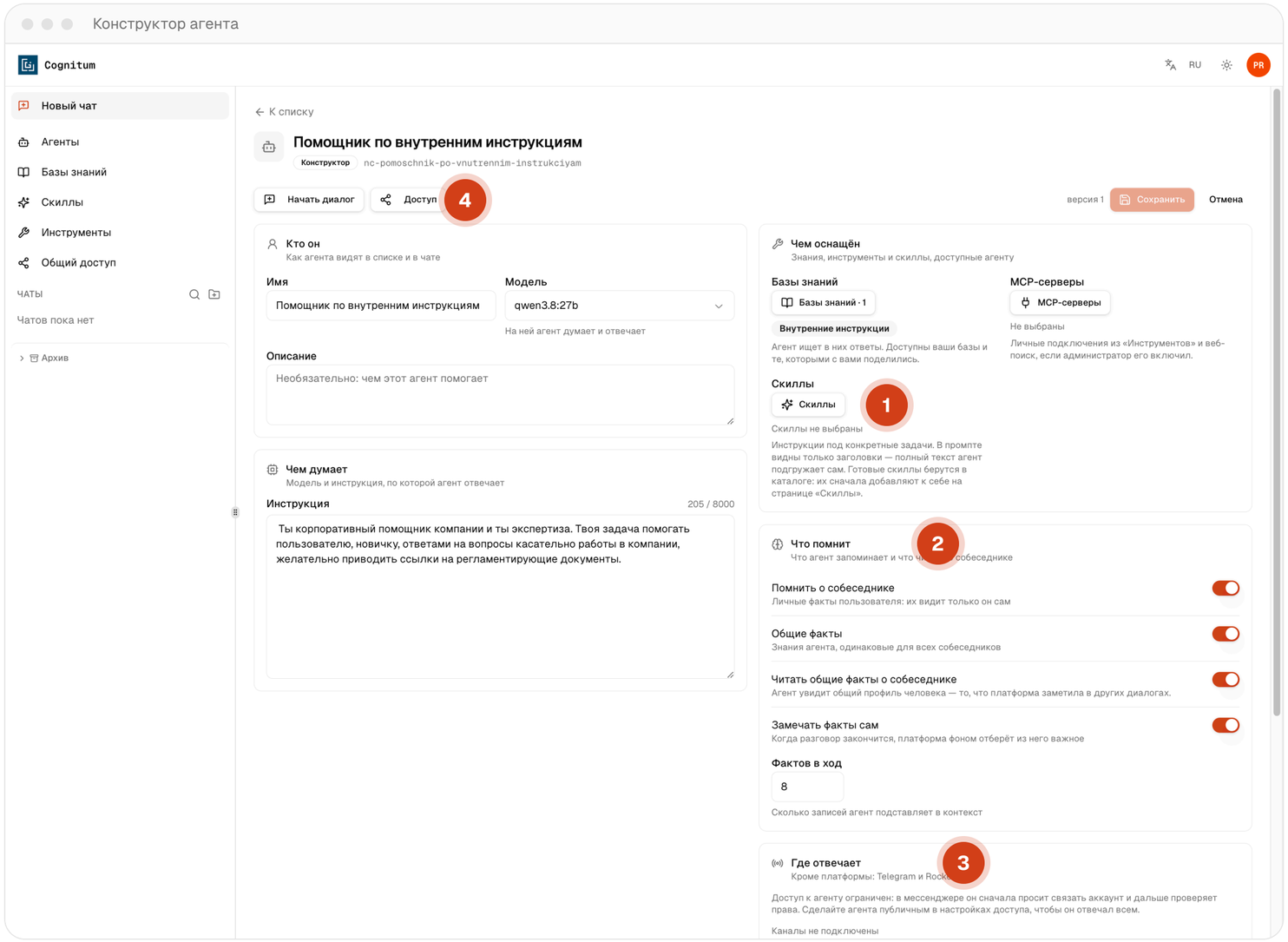
Личное, «обо мне» и общее у агента. Сотрудник видит свои факты и может их удалить.

3 Там, где уже пишут

Тот же агент с теми же базами отвечает в Telegram и Rocket.Chat.

4 Поделился — и коллега работает

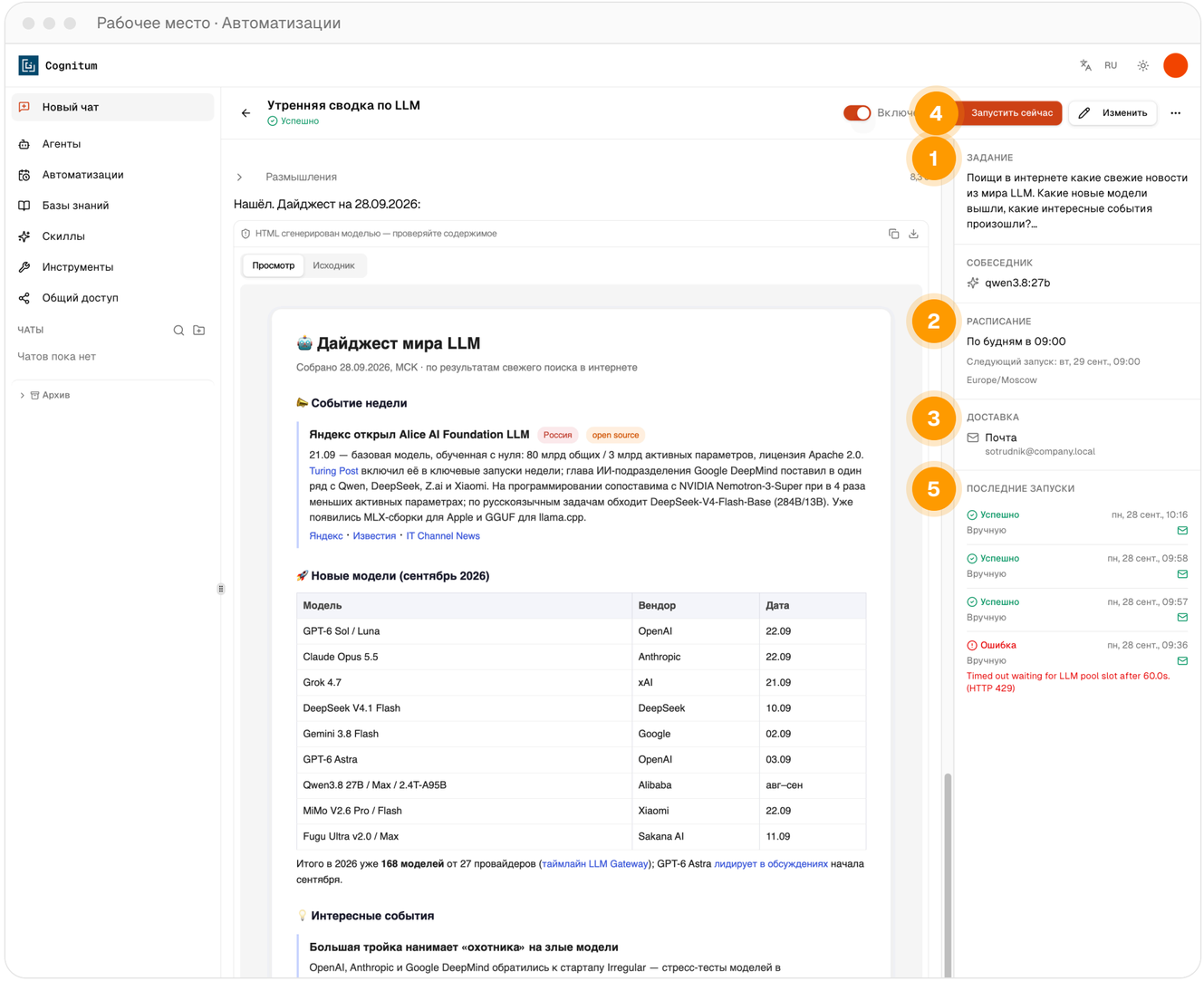
Скиллы агента доступны сразу. Нужна личная учётка к сервису — агент покажет «Подключить».



Скриншот со стенда · права на базу и модель агента при этом не расширяются

Регулярную работу модель делает сама

Задание ставится один раз. Дальше платформа сама отправляет его модели или агенту и присылает ответ.



НАЧАТЬ С ШАБЛОНА

Утренняя сводка | Итоги недели | Регулярная проверка

1 Задание — как в чат

Сотрудник выбирает модель или агента и пишет задание своими словами.

2 Любое расписание

От одного раза до каждых N минут, в своём часовом поясе. Ближайшие запуски видны сразу.

3 Ответ туда, где читают

Каждый ответ остаётся в треде, а по желанию приходит в Telegram, Rocket.Chat или на почту.

4 Проверить, не дожидаясь

«Создать и проверить» и «Запустить сейчас» показывают результат сразу.

5 История и контроль

Видно, когда был запуск и куда доставлен. Упавший запуск придёт в колокольчик и повторится.

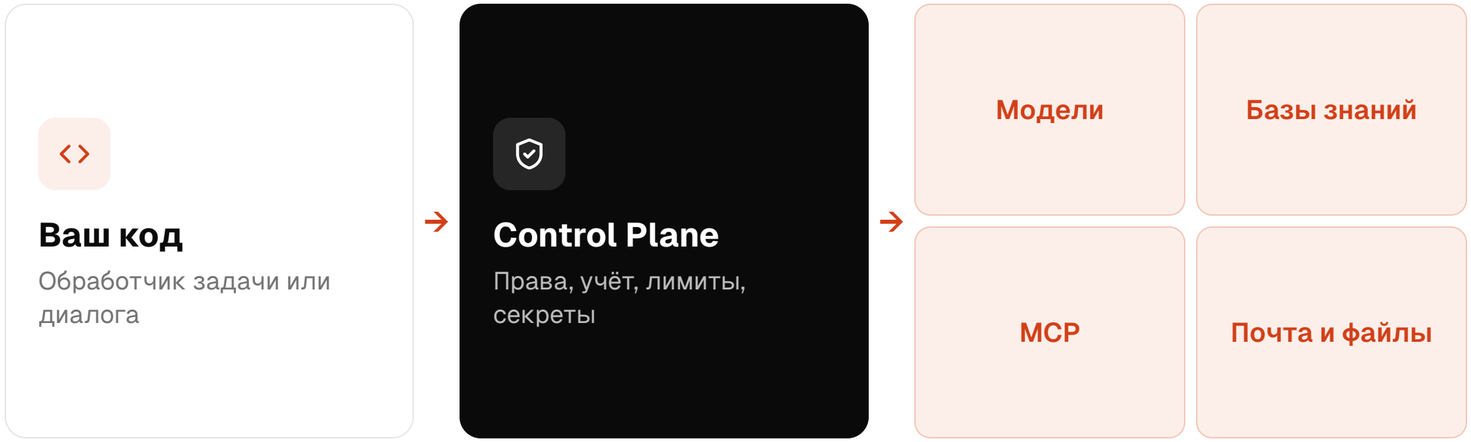
Скриншот со стенда · администратор видит все автоматизации и может остановить любую

Свой агент на Python — без обвязки вокруг

SDK cognitum_agent берёт на себя шину, регистрацию, очереди и вызов моделей. В коде остаётся задача.

```python
# упрощённый пример · SDK 0.16
agent = Agent(AgentConfig(name="review",
                          job_types=["review-contract"]))

@agent.on_job
async def on_job(job):
    q = job.body["question"]
    found = await agent.search_knowledge("contracts", q)
    answer = await agent.llm_chat(
        [{"role": "user", "content": q}])
    return {"answer": answer, "sources": found}
```



1 Задачи и чат

Очередь с гарантией доставки и диалог в рабочем месте — в одном процессе.

2 Модель только через платформу

Учёт, лимиты и политики не обходятся прямым адресом GPU.

3 Свои инструменты

Декоратор публикует инструмент агента в общий MCP-шлюз.

4 Ответ не только текстом

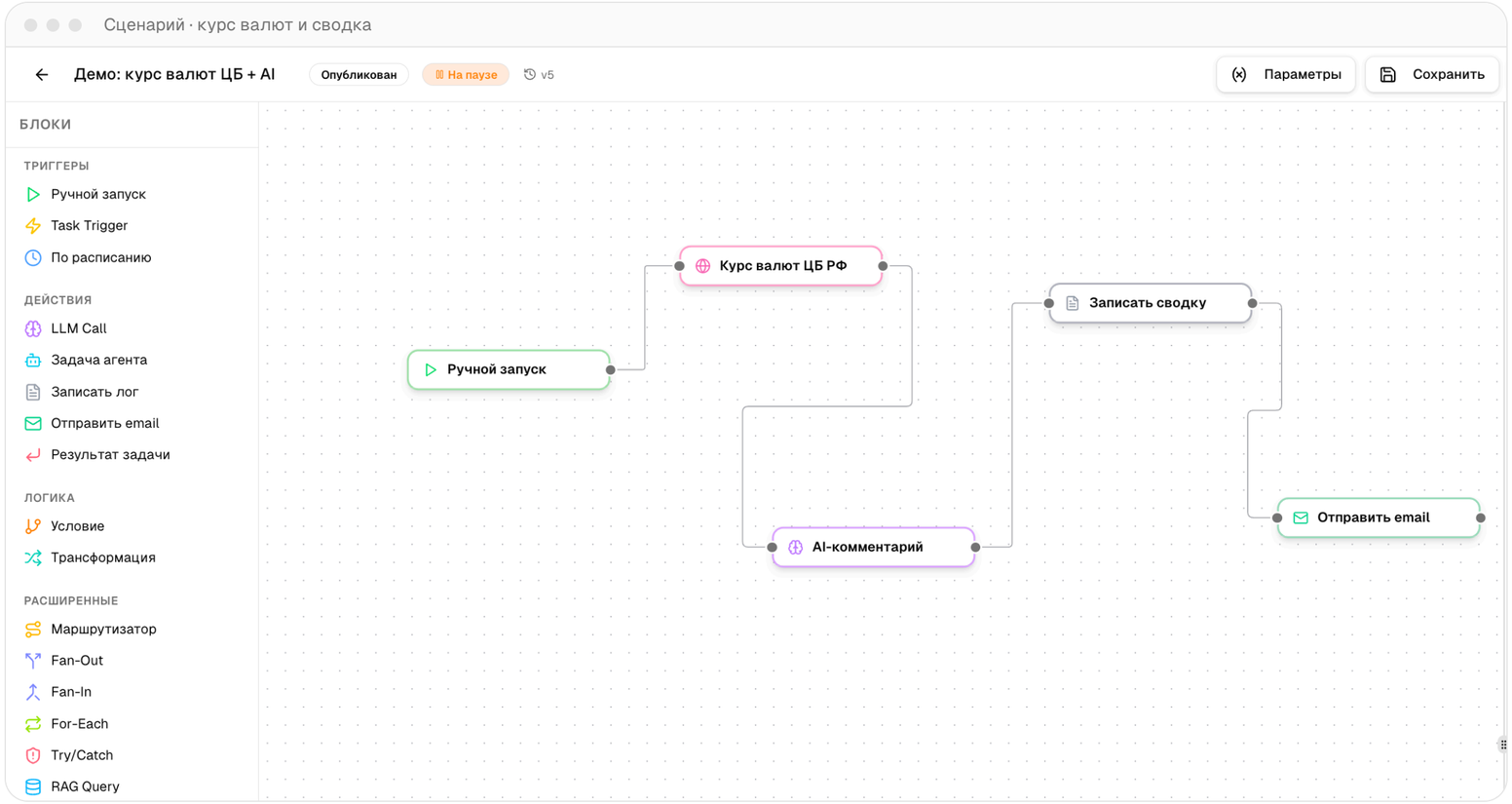
Таблица, график или форма рисуются прямо в диалоге.

5 Секреты не в коде

Хранилища и почту выдаёт платформа, а не строка в образе.

Процесс рисуют блоками, а не заказывают релиз

Конструктор связывает модель, агента и внешний вызов. Свой тип блока добавляется без пересборки образа.



Запуск по делу

Вручную, по расписанию или из задачи агента. История хранит запуски и статусы.

Свой кубик

Свой тип блока добавляется манифестом, без пересборки образа. Код проверяется белым списком импортов.

Версия закреплена

Новая публикация кубика не меняет уже разложенные сценарии. Заливка кода — отдельное право.

ТРИГГЕРЫ

Вручную
По расписанию
Из задачи агента

ДЕЙСТВИЯ

Вызов модели
Задача агента
Письмо, лог

ЛОГИКА

Условие
Трансформация
Маршрутизатор

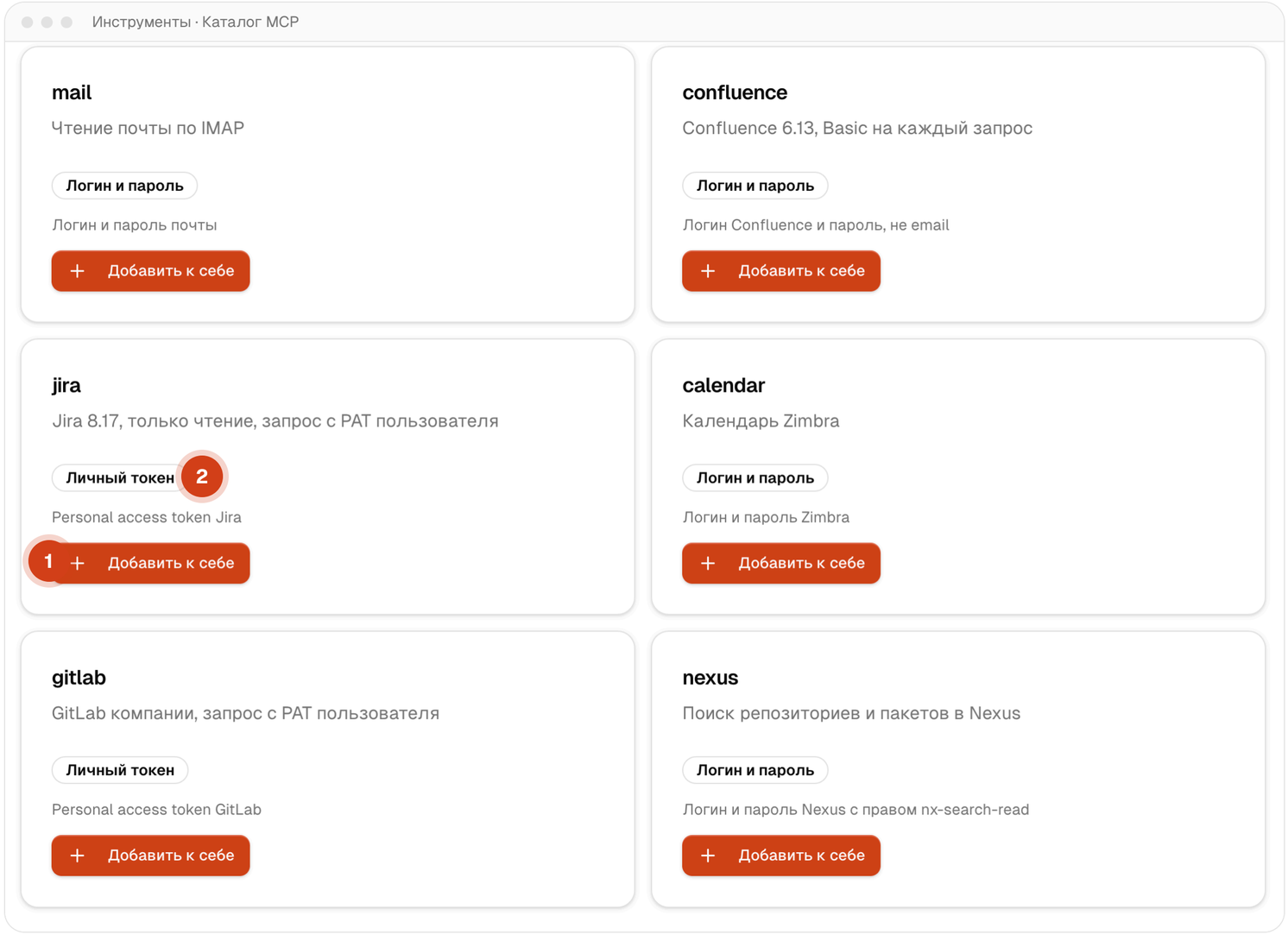
РАСШИРЕННЫЕ

Fan-out и цикл
Try/Catch, HTTP
Поиск по базе

Скриншот со стенда · типовой процесс можно начать с шаблона, а не с пустого холста

В корпоративные системы — от имени сотрудника

Платформа — шлюз MCP в обе стороны: к серверам компании и к инструментам собственных агентов.



ГОТОВЫЕ СЕРВЕРЫ

GitLab · Jira · Confluence · Zimbra · Nexus · Nextcloud

1 Добавить к себе

Сотрудник сам берёт из каталога нужные серверы.

2 Своя учётка и свои поля

Свой токен и поля, пароль в чат не вводится.

Файл из чата — в сервер

Вложение уходит в Nextcloud или GitLab по временной ссылке, без ключей хранилища.

Допуск — по желанию

Проверку реестра включают для строгого контроля, иначе серверы доступны сразу.

Скриншот со стенда · вызовы видны в ленте · лимиты вызовов инструментов задаются в тарифе

Любая модель — через один шлюз

Локальный vLLM или Ollama и внешний API выглядят для агента одинаково. Прямого адреса GPU у агента нет.

Чат сотрудника
Рабочее место и мессенджеры

Агенты
Конструктор и процессы на SDK

Сценарии
Блок «Вызов модели»

Внешние клиенты
Обычный OpenAI SDK по токену

Шлюз моделей Cognitum

Одна точка, через которую проходит каждый вызов

Пулы
Одна модель для сотрудников, маршрут и повтор при сбое

Тарифы
Лимиты на день, неделю и месяц — до вызова модели

Учёт
Кто, какая модель, сколько токенов, сколько ждал

Приоритет
Чат не стоит в очереди за ночной индексацией

vLLM
Свой инференс в контуре

Ollama
Лёгкие локальные модели

Внешние API
Ключ в подключении, не в агенте

Эмбеддинги и реранк
Для баз знаний, тем же путём

OpenAI-совместимый API

Чат, модели, эмбеддинги и реранк. Уже написанные клиенты подключаются без нового протокола.

Нагрузка вживую

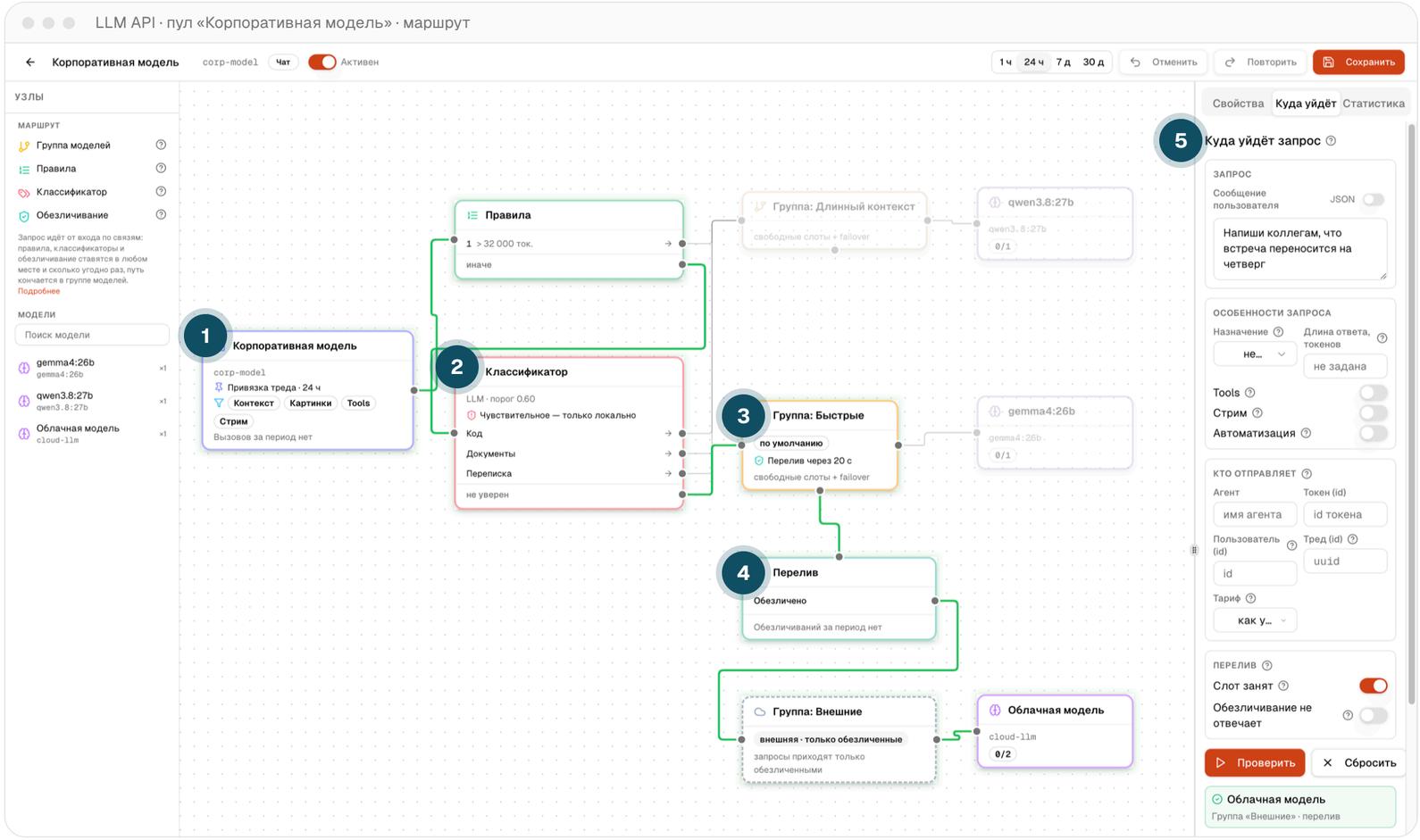
«Мониторинг»: активные вызовы, слоты и очередь, GPU и VRAM по каждой карте, режим для ТВ.

Включается явно

Модель и подключение можно выключить. Ограниченная модель выдаётся списком, а не «всем, кто вошёл».

Одна корпоративная модель вместо нескольких

Администратор собирает пул и рисует маршрут. Сотрудник выбирает одну модель, а ответит подходящая.



ПУЛ ВЫБИРАЮТ КАК ОБЫЧНУЮ МОДЕЛЬ

Чат · Агенты · Автоматизации · OpenAI API · Эмбеддер базы знаний

1 Одна модель для всех

Состав пула меняется без участия сотрудников.

2 Правила и классификатор

Длинный документ — в модель с большим окном. Чувствительное — только локально.

3 Сбой — не ошибка в чате

Запрос идёт в свободный слот, при сбое повторяется на соседней модели.

4 Перелив с обезличиванием

В пик — во внешнюю модель, но с метками вместо имён, ИНН и реквизитов.

5 Проверка до запуска

«Куда уйдёт» — без вызова модели.
«Статистика» — реальные запросы по веткам.

Скриншот со стенда · классификатор работает на локальной модели платформы

Данные остаются в периметре компании

Платформа ставится у заказчика. Внешний выход — только туда, куда администратор его явно открыл.



On-premise

Работает без интернета, если модели локальные. Веб-поиск в таком контуре не включается.

LDAP / Active Directory

Локальные учётки и домен. Группы каталога сопоставляются с ролями платформы.

Роли и права

Права на разделы, доступ к базе, модели и агенту выдаются по отдельности.

Аудит

Действия пользователей и обращения к API. Журнал вызовов моделей — для разбора, не только для счёта.

Секреты не в агенте

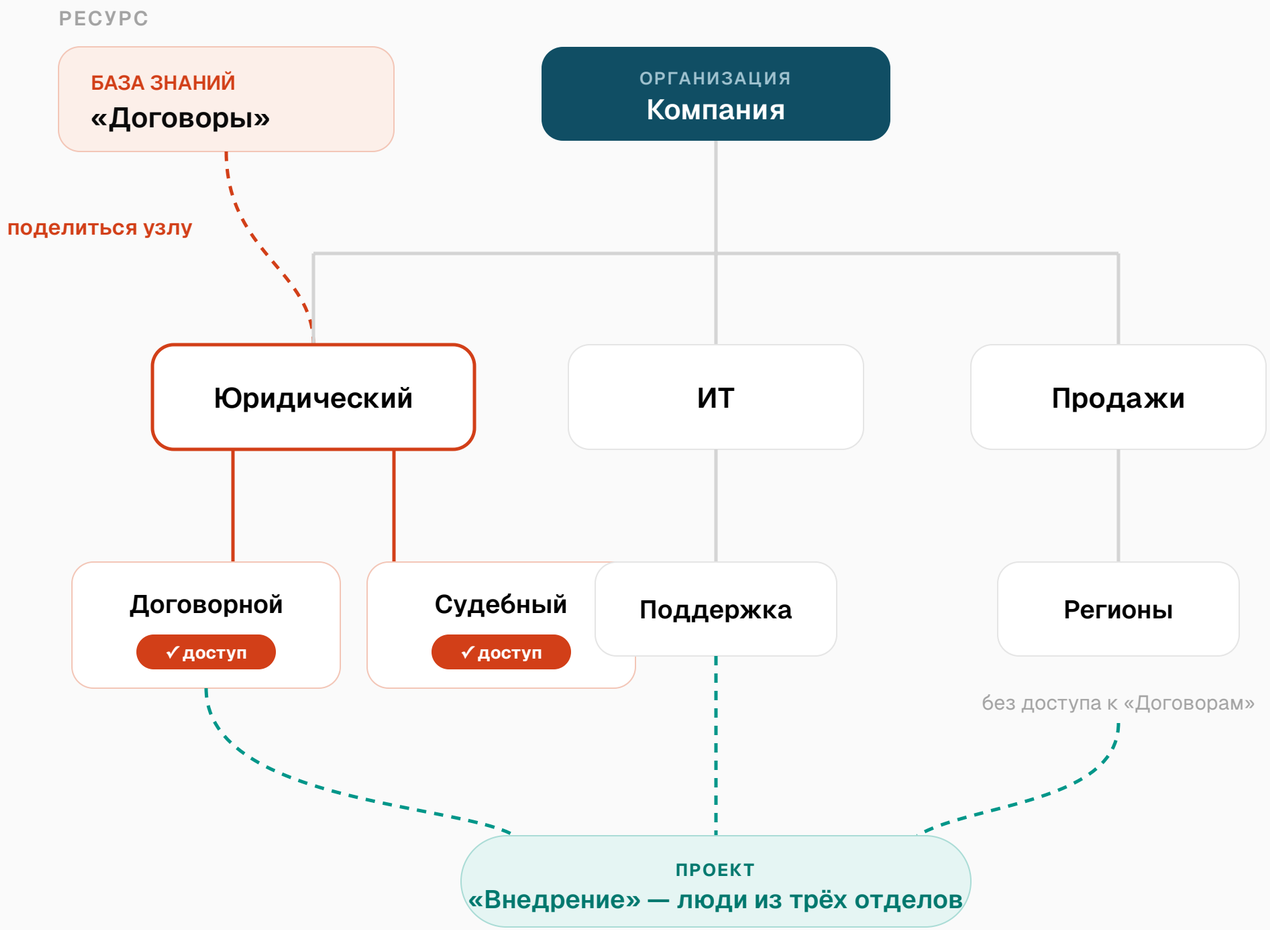
Ключи хранилищ, почты и моделей выдаёт платформа. В код и образ агента они не попадают.

Обезличивание данных

Во внешнюю модель уходят метки вместо имён, ИНН, СНИЛС и реквизитов, в ответе — снова значения. Чувствительное остаётся локально.

Доступ выдаётся узлу структуры и идёт вниз

Подразделение и проект — кому можно поделиться базой и в каком разрезе смотреть расход. Не техническая группа.



Наследование вниз

Право на «Юридический» получают его отделы. Соседние подразделения — нет.

Проект поперёк дерева

Временная команда из разных отделов получает доступ, не ломая оргструктуру.

Перевод сотрудника

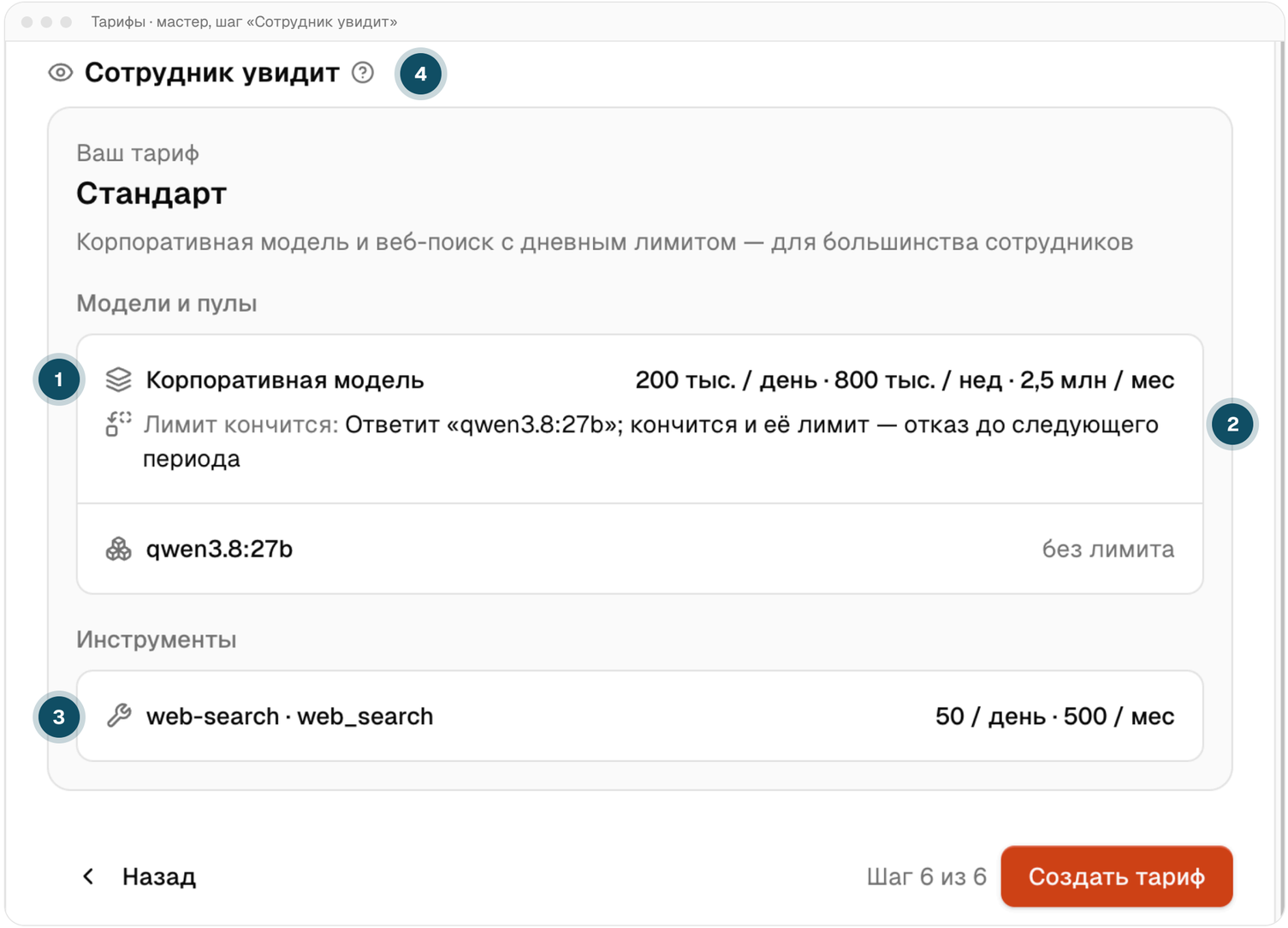
Доступ переезжает вместе с узлом. Группу под человека собирать не нужно.

Схема условная. Дерево подразделений у каждого заказчика своё.

Импорт дерева из файла — ещё не в продукте, структура заводится в интерфейсе

Лимит срабатывает до вызова, а не в счёте

Тариф задаёт модели, лимиты, ставки и запасной вариант. Шлюз проверяет его до выбора модели.



1 Модели и лимиты

На день, неделю и месяц. Модели вне тарифа недоступны.

2 Лимит кончился — работа идёт

Запасная модель, фоновый режим или отказ — решает тариф.

3 Инструменты тоже считаются

Число вызовов MCP-инструмента или сервера, например веб-поиска.

4 Сотрудник видит заранее

Остаток — кольцами в шапке, свой тариф — в «Потреблении».

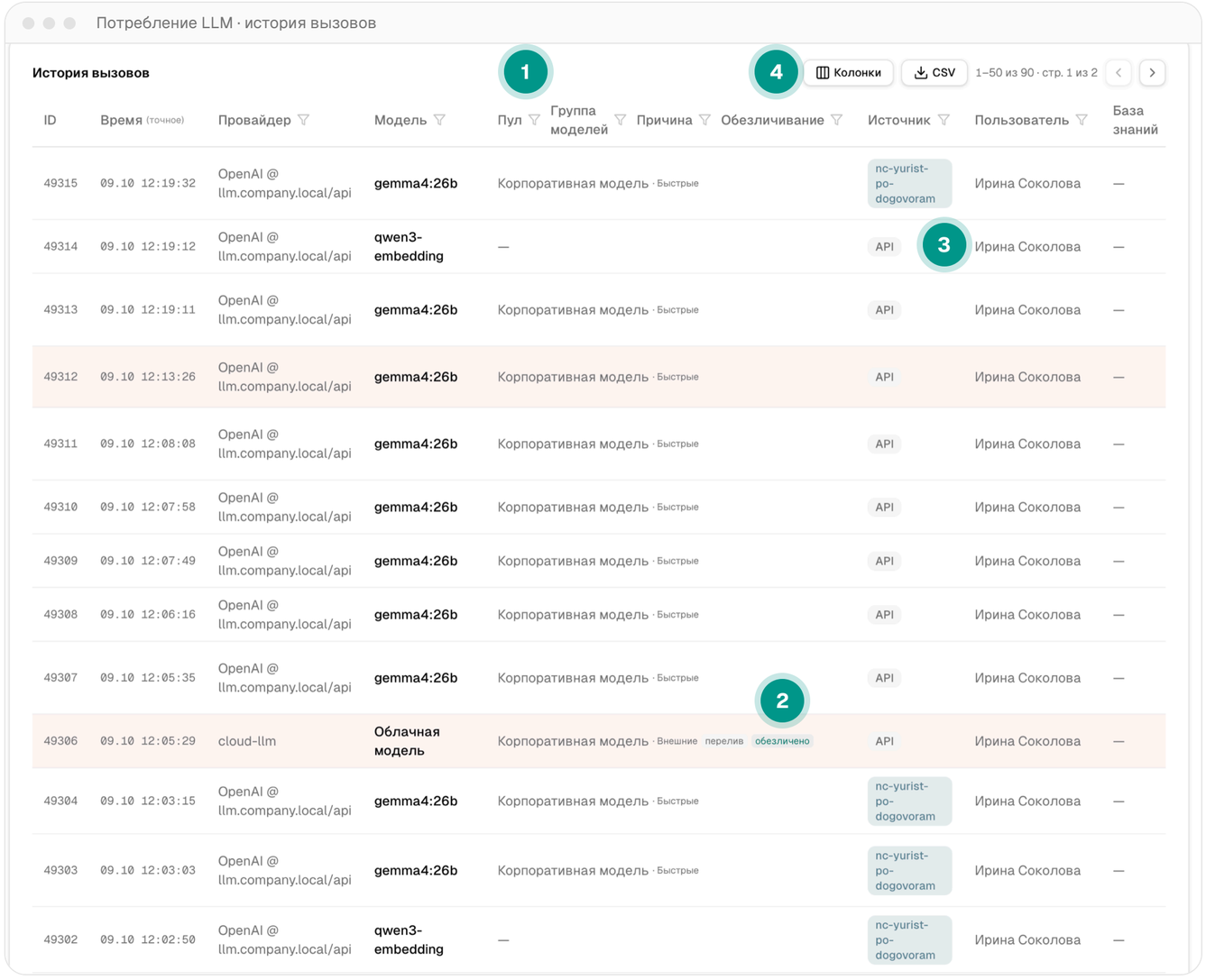
КОМУ НАЗНАЧАЕТСЯ

организации · подразделению с поддеревом · сотруднику

Скриншот со стенда · ставки за 1 млн токенов и абонплата · деньгами управляет право billing.manage

Видно, кто потратил токены и на что

Каждый вызов — с моделью, пулом и причиной выбора. Отчёт сводит расход с планом тарифа.



1 Почему эта модель

Пул, группа и причина выбора — у каждого вызова.

2 Что ушло наружу

Перелив во внешнюю модель отмечен: запрос обезличен.

3 Кто и откуда

Чат сотрудника, агент и API не падают в одну кучу. База знаний — отдельной строкой.

4 Свои колонки и CSV

Набор столбцов, представления и выгрузка для финансов.

АНАЛИТИКА И ОТЧЁТ

обзор · разрезы · задержки · ошибки · инструменты · тарифы · **план и факт** · **деньги**

Скриншот со стенда, данные условные · сотрудник видит свой расход в настройках, деньги — только финансы

Из чего состоит платформа

Агент не ходит в базу и в модель напрямую. Всё управление проходит через Control Plane и шину.

Интерфейсы
кто и как входит

Рабочее место | Админка | OpenAI API | MCP API | Мессенджеры

Control Plane
одна точка управления

Регистрация агентов | Шлюз и пулы моделей | Тарифы | Шлюз MCP | Права | Задачи | Файлы | Расписание

NATS JetStream
шина

`команды — сразу` | `задачи — очередью` | `метрики агентов` | `гарантия доставки`

Хранилища
состояние и данные

PostgreSQL | Qdrant | Loki | S3

Агенты и сервисы
отдельные контейнеры

Агенты на SDK | Агенты без кода | Обезличивание

Одни правила

Модель, база и инструмент вызываются только через платформу. Учёт и права нельзя обойти прямым адресом.

Сбой изолирован

Упавший агент не останавливает платформу. Его задачи ждут в очереди с гарантией доставки.

Масштаб копиями

Несколько экземпляров одного агента разбирают одну очередь задач.

Neo4j в поставке опционален и для работы платформы не требуется

Ставится в контур и обновляется по частям

Docker Compose, образы под amd64 и arm64. Миграции базы применяются сами при старте.

1

Установка

Один compose: платформа, база, шина, логи, разбор документов.

compose · PostgreSQL · NATS · Loki

Платформа запущена

2

Модели

Подключение к своему инференсу или внешнему API, лимиты и видимость.

vLLM · Ollama · внешний API

Чат отвечает

3

Знания

Первая база из файлов, каталога хранилища, Jira или Confluence.

файлы · S3, SMB · Jira · Confluence

Ответы с цитатой

4

Агенты

Конструктор для типовых задач, контейнер на SDK — для своих.

конструктор · SDK · сценарии

Ассистенты отделов

5

Масштаб

Оргструктура, пулы моделей, тарифы и отчёт о расходе.

оргструктура · пулы · тарифы

Работает на компанию

Обновление без правок агентов

```
docker compose up -d --build --force-recreate \
  app echo-agent nocode-harness
```

Агенты на SDK 0.15 работают без пересборки.

Закрытый контур

Офлайн-дистрибутив включает разбор документов, исполнитель агентов конструктора и справку /docs. Модели — локальные.

Эксплуатация

«Мониторинг» с живыми вызовами, GPU и VRAM по каждой карте. Логи с поиском и потоком прямо в интерфейсе.

Что собирают на платформе

Это не каталог готовых агентов из коробки, а задачи, которые закрываются конструктором, базой и SDK.

конструктор + MCP

Первая линия поддержки

Ответ по инструкциям и прошлым заявкам, со ссылкой на источник.

1. База заявок и инструкций
2. Агент в конструкторе
3. Тикет в трекер через MCP

конструктор

Регламенты и FAQ отдела

Ассистент со своими базами, инструкцией и кругом людей.

1. База регламентов
2. Доступ подразделению
3. Тот же агент в мессенджере

SDK + файлы

Разбор документов

Договор или комплект файлов читается платформой, агент сверяет с базой норм.

1. Файлы в диалог или задачу
2. Разбор документа платформой
3. Сверка агентом на SDK

агент на SDK

Протокол встречи

Транскрибация и разбор встречи — отдельным процессным агентом.

1. Запись в рабочее место
2. Транскрибация агентом
3. Протокол в диалоге

автоматизация

Регулярный отчёт

Сводка по расписанию вместо ручного копирования.

1. Задание модели или агенту
2. Расписание в своём часовом поясе
3. Ответ в Telegram или на почту

MCP + аудит

Разбор для ИБ

Агент читает журналы через допущенный инструмент и оставляет след.

1. Допуск MCP-сервера
2. Агент с узким набором
3. След в аудите платформы

Штатного набора отраслевых агентов в поставке нет — его собирает команда внедрения

Что уже есть — и что дальше

Четыре последних релиза и очередь, по которой идёт продукт. Сроки называем, когда работа взята в поставку.

0.14

Агенты без кода

- Конструктор агентов
- Telegram и Rocket.Chat
- Скиллы и память

Ассистент отдела без разработчика

0.15

Внедрение в компанию

- Оргструктура и доступ узлу
- Jira и Confluence в базах
- Личные учётки MCP, веб-поиск

Доступ по оргструктуре

0.15.1

Автоматизации

- Задания по расписанию
- Доставка в мессенджер и почту
- HTML и PDF в чате
- Готовые MCP-серверы

Регулярная работа без запроса

0.16

Модели под контролем

- Пулы и маршрут на холсте
- Перелив с обезличиванием
- Тарифы и лимиты
- Модель работает с базой

Одна модель, понятный расход

Дальше, без дат

Свою LLM не строим. Автономность «агент решает сам» не обещаем

Обезличивание везде

Не только при переливе пула: каждый внешний вызов проходит обезличивание, обойти нельзя.

Корпоративный вход

OIDC и SAML через IdP заказчика без ручной синхронизации учёток.

Человек в контуре

Перед важным действием агент спрашивает разрешения. Единый след агента, инструмента и модели.

Парк агентов

Каталог, версии и откат, чтобы агентов становилось больше без ручной сборки.

Cognitum AI Platform

Мы всегда на связи

Покажем платформу на живом стенде и обсудим, как запустить первый сценарий в вашем контуре.



САЙТ
cognitum.it-expertise.ru

ПОЧТА
info@it-expertise.ru

ТЕЛЕФОН
+7 499 450 28 86

ОФИС
119435, Москва,
ул. Малая Пироговская, 16

IT EXPERTISE

Сайт платформы