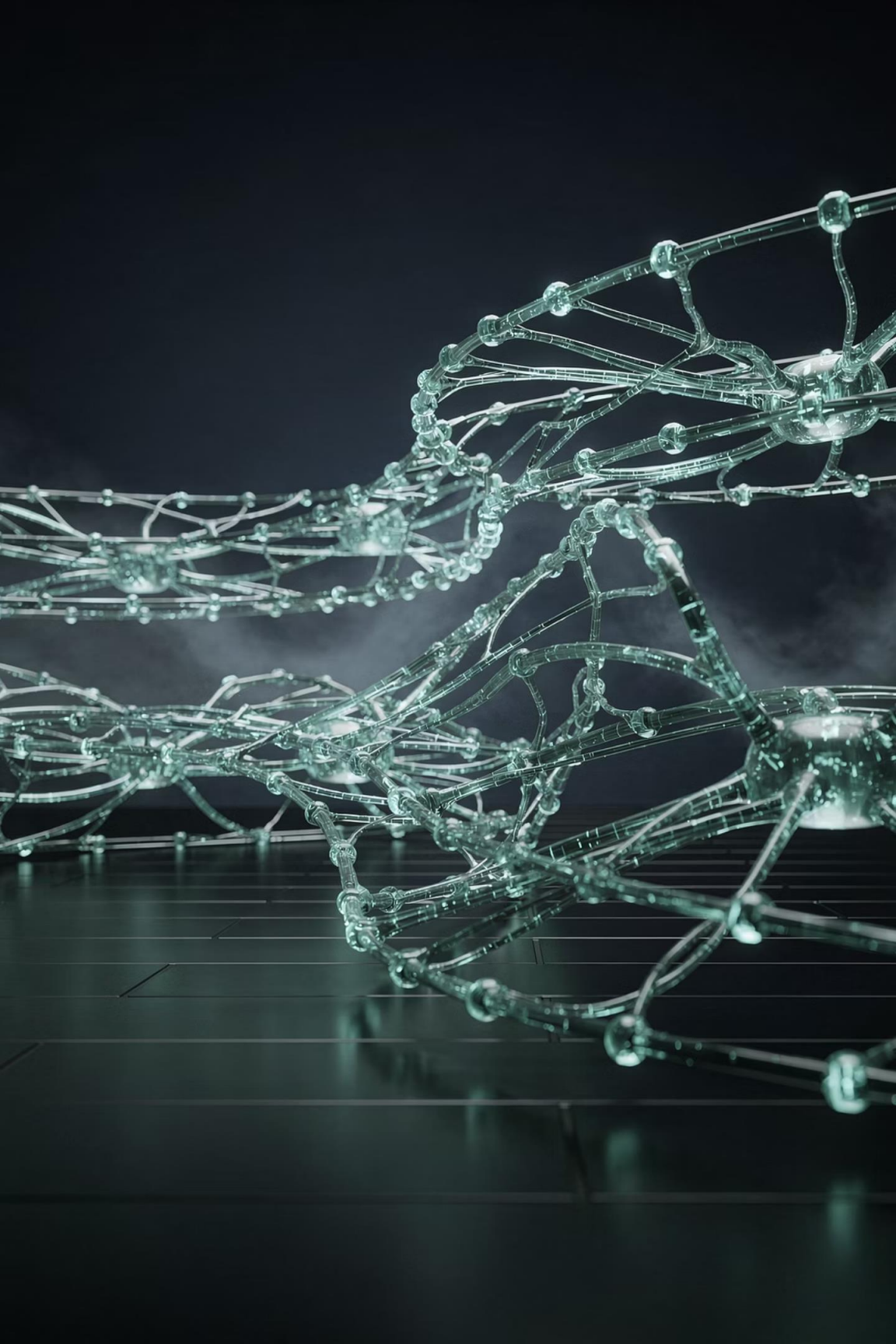




КЛИЕНТСКИЙ КЕЙС

НЕОФЛЕКС × СБЕР GIGACHAT

Корпоративный ИИ-ассистент на базе GigaChat и Neoflex Neon

Как за **52 дня** запустить R&D-лабораторию ИИ-гипотез для **1 500** пользователей крупного холдинга

Неофлекс × Сбер GigaChat

Компания Неофлекс

- **20+ лет** на рынке ИТ-трансформации
- Финтех, ритейл, промышленность
- **Neoflex Neon** — AgentOps-платформа
- Управление ИИ-агентами предприятия

Neoflex Neon + GigaChat

- Единый шлюз к GigaChat API
- Мониторинг токенов и расходов
- Безопасность и compliance
- Запуск пилотов — **от 5 дней**

Вызов клиента

Крупный холдинг · 700+ сотрудников · выручка 50+ млрд ₽/год



Масштаб

1 500 пилотных →
30 000 клиентов-
пользователей



Сроки

52 дня на полный
запуск



Безопасность

Контроль данных,
логирование,
compliance



Партнёр

Нужна лаборатория для проверки ИИ-гипотез и платформа эксплуатации ИИ



3

РЕШЕНИЕ

Решение: Neoflex Neon + GigaChat



R&D-лаборатория

Конвейер ИИ-гипотез, пошаговый процесс, low-code прототипы



Neoflex Neon

AgentOps-платформа, мониторинг токенов, безопасность данных



GigaChat API

Генерация, суммаризация, классификация, мультимодальность

Пример ИИ-гипотезы: от идеи до метрик



Гипотеза

Если запустить ИИ-агента для подбора и генерации иллюстраций по тексту и бренд-гайду, то время дизайнера сократится в 2 раза

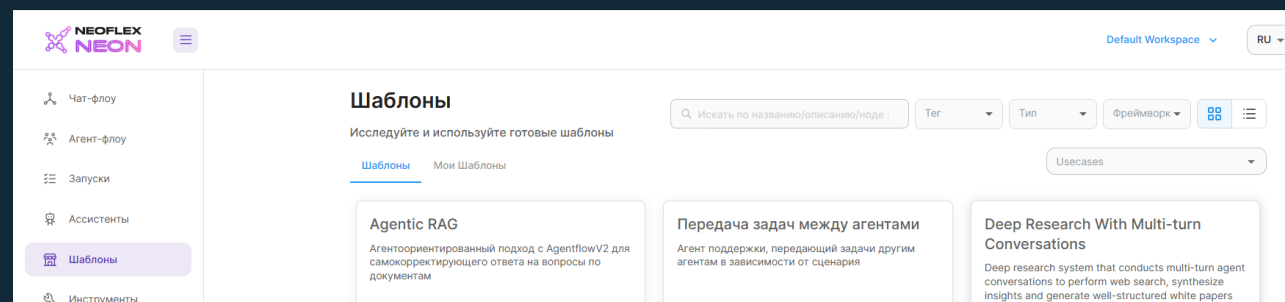


Сбор требований

Интервью с дизайнерами, метрики текущего процесса создания иллюстраций, технические ограничения, интеграции



Сборка прототипа



Интерфейс агента: загрузка текста материала, автоподбор стиля из бренд-гайда, генерация 3–5 вариантов иллюстраций



Метрики и бюджет



Время на материал: 120 -> 60 минут,

качество $\geq 4/5$,

итерации ≤ 2 .

Токены: 2–3 млн токенов в месяц

Бюджет: порядка 2 000 ₽ в месяц на команду дизайнеров

52 дня до запуска

Неделя 1–2

Развёртывание

Облачный стенд Neon, подключение GigaChat API,
корп. аутентификация

Неделя 5–7

Масштабирование

1 500 пользователей, AI AgentOps-мониторинг,
оптимизация промптов

1

2

3

Неделя 3–4

R&D-гипотезы

Сценарии суммаризации, генерации, классификации.
Прототипы ИИ-агентов

AI AgentOps: контроль ИИ-агентов

Зачем?

- «Чёрный ящик» ИИ-агентов
- Неконтролируемые расходы
- Риски ИБ и утечки данных

Neoflex Neon

- **Observability** — учет всех взаимодействий
- **Evaluation** — качество ответов
- **Analytics** — метрики и тренды
- **Compliance** — логирование, аудит



Сценарии на базе GigaChat



Корпоративный ассистент

RAG + GigaChat: ответы по внутренним документам и базам знаний



Суммаризация

Автоматические резюме документов, протоколов встреч, отчётов



Классификация

Маршрутизация обращений, приоритизация, анализ тональности



Генерация контента

Черновики текстов, описания, шаблоны писем и документов

GigaChat SaaS vs On-premise

Потребление: 2–4 млн токенов/мес на 150 пользователей

GigaChat SaaS (API)

- Старт за 1–2 недели
- Нет CAPEX на инфраструктуру
- Идеально для R&D и пилотов

⚠ Pay-per-use: рост при масштабе

On-premise / Гибрид

- Контроль данных 100%
- Fine-tuning под домен
- Фиксированная стоимость

⚠ CAPEX, команда, поддержка

Результаты проекта

52 дня

запуск

вместо 120+

1 500

пользователей

пилотная группа

на 60%

дешевле

стоимость эксперимента

-40%

Time-to-market

ускорение вывода ИИ-инициатив

R&D-
лаборатория

Готовый конвейер
ИИ-гипотез на базе
GigaChat

AI AgentOps

Полная
наблюдаемость и
контроль расходов

Стратегия

GigaChat SaaS для
R&D → гибрид при
масштабировании

Рекомендации

Начните с GigaChat SaaS

Используйте SaaS-версию для R&D и пилотных проектов, это обеспечивает быстрый запуск, низкие риски и оплату по факту использования.

Платформа AgentOps

Внедрите платформу для мониторинга, безопасности и контроля затрат с самого начала эксплуатации ИИ-агентов.

Создайте R&D-лабораторию

Разработайте поэтапный процесс для проверки гипотез и быстрого проведения экспериментов с ИИ-агентами.

Гибридная стратегия

Планируйте использование GigaChat SaaS для экспериментов и рассмотрите переход на on-premise/гибрид при масштабировании в продакшен.

Практики AI AgentOps

Применяйте передовые практики AgentOps: наблюдаемость, оценка качества, комплаенс и аналитика для полного контроля.

Фокус на ROI

Сосредоточьтесь на бизнес-сценариях с чётким возвратом инвестиций, таких как суммаризация, классификация, корпоративные ассистенты и генерация контента.