



Барьеры и проблемы внедрения решений на базе ИИ:

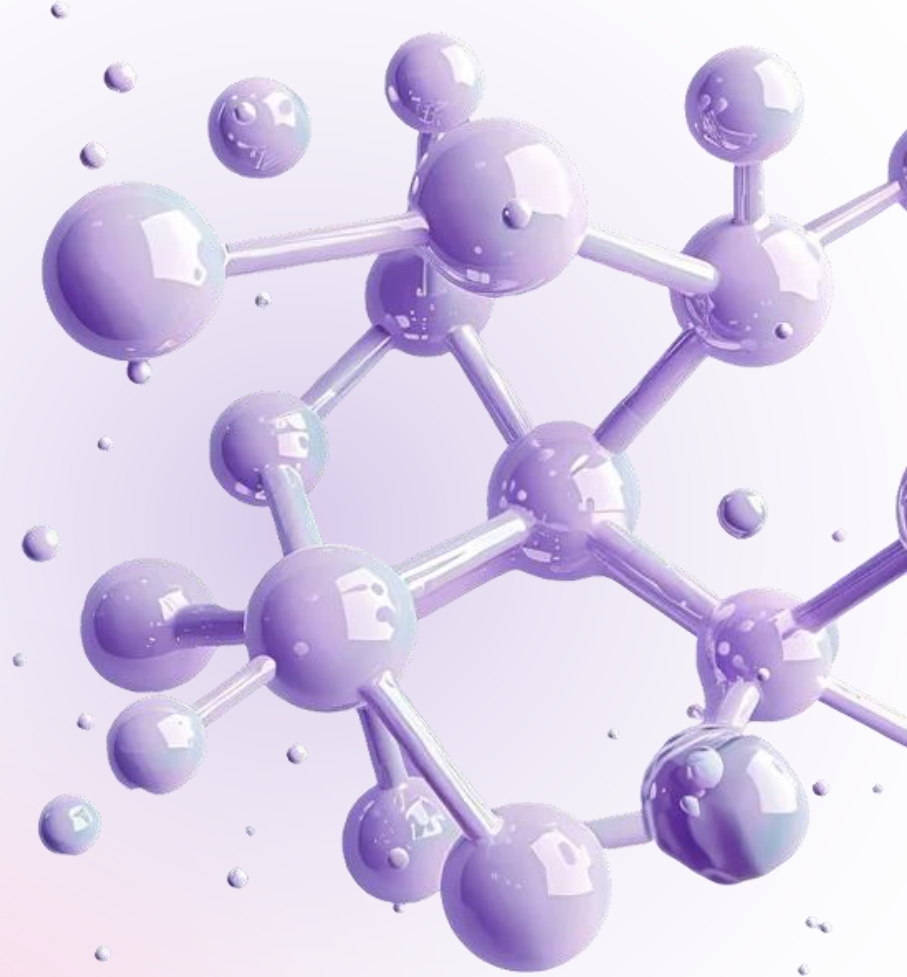
от проверки гипотез до промышленной
эксплуатации

Федор Надейкин

Бизнес-лидер направления ИИ, Neoflex



ИЦК ХИМИЯ
И ФАРМАЦЕВТИКА



Качество данных

Одна из главных и важных проблем корректности работы ИИ-агентов – качество данных

Разрозненность

Данные распределены по десяткам legacy-систем без единой модели данных

Отсутствие структуры

Данные не структурированы и не пригодны для использования

Качество и консистентность

Ошибки, дубли, пропуски – стандартная ситуация в производственных данных

Решение – инкрементальный подход:

- 1 Начните с задач, где данных достаточно
- 2 Параллельно выстраивайте процесс работы с данными
- 3 Постепенно расширяйте количество источников и объем данных

Стратегия индексирования документов для RAG

Типичные проблемы

Потеря контекста

- Разрушение логических связей между разделами документа
- Потеря иерархии и структуры при чанкинге
- Смешение информации из разных контекстов

Проблемы retrieval

- Неточный подбор релевантных чанков
- Семантическая путаница при схожих эмбедингах
- Избыточный шум в результатах поиска

Ограничения контекста

- Переполнение контекстного окна LLM
- Неспособность обработать сложные взаимосвязи между документами
- Потеря важных деталей при агрегации

Качество ответов

- Галлюцинации даже при наличии корректных ответов
- Неполные ответы на сложные вопросы
- Отсутствие проверки фактов и валидации

Решение - для разных документов используй разные подходы:

RAG

Для простых документов (статьи, FAQ), когда нужны прямые ответы. Пример запроса: *Температура синтеза аммиака по регламенту №12?*

Graph RAG

Когда важны связи между сущностями и нужно отслеживать влияния и зависимости. Пример запроса: *Какие регламенты затрагивают изменение вещества X?*

Agentic RAG

Для многошаговых задач, требующих синтеза данных из разных источников и проверки ответа. Пример запроса: *Проверь актуальность всех ссылок на ГОСТы в базе знаний и предложи замены*

Page Index

Для сложных структурированных документов, где важно сохранить целостность данных. Пример запроса: *«СИЗ при работе с хлором по инструкции ОТ-047?»*

Модели для проверки гипотез

отсутствует железа • не достаточно текущего железа • нет опыта развертывания и мониторинга моделей

Cloud

Эксперимент за дни

Не месяцы, а дни до первого результата.

Отсутствует CAPEX

Оплата по факту использования.

Разные модели под разные задачи

Подбор оптимальной модели для каждого сценария.

On-premise

Полный контроль данных

Соответствие требованиям ИБ.

Нет внешних зависимостей

Стабильность и предсказуемость.

Возможность дообучения адаптеров

Дообучение адаптеров для улучшения качества ответов.

А как же ИБ? 1. Используйте синтетические или обезличенные данные на этапе экспериментов – это снимает основное препятствие согласования. 2. Используйте AI-Gateway для контроля доступа к облачным моделям.

Метрики

субъективная оценка • споры «работает / не работает» • невозможно сравнить результаты • деградация незаметна • ROI не рассчитывался • бесконечные доработки

Решение - определите метрики до старта работ:

Бизнес

- Экономия времени (часы, дни, месяцы)
- Снижение ошибок (%)
- ROI (срок окупаемости)
- NPS пользователей (удовлетворенность)

Технические

- Точность ответов (precision / recall)
- Латентность (мс на запрос)
- Галлюцинации (% отказов)
- Контроль ИБ (промпт атаки, токсичность и т.д.)

Тестирование решений на базе ИИ

ПРОБЛЕМЫ

ИИ-агенты работают в многошаговых сценариях – ошибка на одном шаге каскадно ломает весь результат

Не возможно отслеживать результаты тестирования и влияния на результат

Ручная проверка каждого сценария занимает недели и требует глубокой экспертизы

РЕШЕНИЕ

Фреймворки тестирования

Эталонные датасеты, метрики качества, автоматическое сравнение ответов

LLM-судья

Модель оценивает ответы модели

CI/CD для ИИ-агентов

Автоматические регрессионные тесты



Спасибо за внимание!
Вопросы?



Федор Надейкин
@f_nadeykin
8-929-554-47-05



Консультант по
созданию AI-
помощников