runetlex . academy × Яндекс

ПРИКЛАДНАЯ
КОНФЕРЕНЦИЯ
ПО ИИ ОТ ПРАКТИКОВ

ДЕЛАЙ RAG: как «обучить» нейросеть вашими данными (и где предел, когда проще заплатить)

Екатерина Якуненко

Эксперт-юрист команды разработки Нейроюриста, Яндекс

RAG: технология обогащения контекста

RAG
Retrieval Augmented Generation

RETRIEVAL: поиск по базе знаний – внешней и заранее подготовленной

AUGMENTATION: обогащение промпта данными из базы

GENERATION: генерация нейросетью ответа с учётом данных из базы

Модель не обучается на вашей базе.

LLM не знает ваших документов,
она получает их прямо сейчас.

И доставляет их RAG.

Ценность технологии

Без RAG – знания из параметрической памяти
Представления, осевшие в весах моделей во время настоящего обучения (training)

Актуальность
в базе актуальные документы в момент ответа

Достоверность
модель обязана отвечать с опорой на текст из базы

Верифицируемость
модель цитирует источник

Юридические задачи, которые с RAG решаются лучше

- доступ чат-ботов-консультантов к НПА и практике
- драфтинг с учётом вашей собственной библиотеки примеров
- работа с локальными актами и политиками

RAG хорошо работает там,
где **конкретный массив** данных
важнее общей эрудиции LLM

Можно ли сделать RAG самому?

Нужно!

Правило 0: вайб-кодинг в помощь!

Правило 1
Зачем вам это?

Сначала нужно чётко определить задачу, чтобы понять, как проектировать базу

У хорошей задачи:

- понятные отраслевые и функциональные границы
- качество результатов измеримо
- решается с помощью набора однородных документов

Создаваемая вами система не должна обслуживать созданную вами базу!
База в синергии с остальными компонентами системы.

Правило 2 (самое важное)
Подготовка данных

Единого рецепта создания RAG-системы не существует!

Для каждого массива данных нужна своя логика обработки данных.

Законы
Нарезка по статьям
Сохранение иерархических связей

Судебная практика
Нарезка по смысловым блокам
Суммаризация

Локальные акты
Сохранение связей между актами одной тематики

Правило 3
Итеративные тесты

Система готова, когда даёт устраивающий результат на запросах, где вы знаете правильный ответ.

Запуск → Проверка на реальных запросах → Корректировка параметров → Повтор

Параметры для итераций:

- техника поиска (семантический, ключевые слова, гибридные техники)
- размер фрагментов текстов
- количество возвращаемых фрагментов

Пределы DIY RAG

Когда задача становится инженерной

Массив данных
неоднородный, исчисляется сотнями тысяч документов

Точность поиска
критична высокая релевантность и скорость поиска

Актуальность
данные обновляются регулярно и иногда внезапно

В таком случае логично обратиться к LegalTech-провайдерам,
заявляющим, что у них есть RAG

Чек-лист вопросов вендорам

- что именно есть в базе?
- как часто она обновляется? удаляются ли утратившие силу документы?
- сервис в ответах цитирует источник?
- бонусный вопрос: *откуда данные?*

RAG – это не модель и не обучение.
Это архитектура подачи нужного документа в нужный момент.

Качество архитектуры = качество данных.

СПАСИБО ЗА ВНИМАНИЕ

Сайт

ТГ-канал

База знаний