

RAG и генеративный ИИ

Создаем собственные RAG-пайплайны с помощью LlamaIndex, Deep Lake и Pinecone

Дэнис Ротман

SPRINT
book 2025

Краткое содержание

Об авторе.....	11
О научных редакторах.....	12
Предисловие	13
От издательства	19
Глава 1. Почему именно генерация с дополненной выборкой (RAG)	20
Глава 2. Хранение эмбедингов RAG с помощью Deep Lake и OpenAI	52
Глава 3. Реализация индексированной RAG с помощью LlamaIndex, Deep Lake и OpenAI	80
Глава 4. Мультимодальная RAG.....	110
Глава 5. Повышение эффективности RAG за счет обратной связи от эксперта ...	136
Глава 6. Масштабирование банковских данных клиентов в системе RAG с помощью Pinecone	161
Глава 7. Создание масштабируемой RAG на основе графа знаний с помощью API Википедии и LlamaIndex.....	199
Глава 8. Создание динамических RAG с помощью Chroma и Hugging Face Llama.....	227
Глава 9. Улучшение моделей ИИ: тонкая настройка данных RAG и обратная связь.....	253
Глава 10. RAG для создания стоковых видео с помощью Pinecone и OpenAI.....	271
Приложение.....	307

Оглавление

Об авторе.....	11
О научных редакторах.....	12
Предисловие	13
Целевая аудитория.....	13
Что освещается в книге	14
Необходимые знания	16
Файлы с примерами кода	17
Цветные иллюстрации.....	17
Условные обозначения.....	17
От издательства	19
О научном редакторе русскоязычного издания	19
Глава 1. Почему именно генерация с дополненной выборкой (RAG).....	20
Что такое RAG.....	21
Простая, продвинутая и модульная конфигурации RAG.....	23
RAG или тонкая настройка?.....	23
Экосистема RAG.....	25
Ретривер (D).....	26
Генератор (G).....	28
Модуль оценки (E).....	28
Учитель (T).....	29
Простая, продвинутая и модульная RAG в коде.....	29
Часть 1. Основы и базовая реализация	30
Часть 2. Продвинутые техники и оценка.....	35
Резюме	50
Вопросы	50
Ссылки	51
Дополнительная информация	51
Глава 2. Хранение эмбедингов RAG с помощью Deep Lake и OpenAI.....	52
От сырых данных к эмбедингам в векторной базе данных.....	53
Организация RAG в виде пайплайна	55
Пайплайн генеративного ИИ на базе RAG.....	56

Создание пайплайна RAG.....	57
Настройка среды.....	57
Сбор и подготовка данных.....	62
Создание и сохранение эмбедингов.....	65
Информация о векторной базе данных.....	68
Генерация дополненного ввода.....	70
Оценка вывода с помощью метрики косинусного сходства.....	75
Резюме.....	77
Вопросы.....	78
Ссылки.....	79
Дополнительная информация.....	79
Глава 3. Реализация индексированной RAG с помощью LlamaIndex, Deep Lake и OpenAI	80
В чем польза RAG с индексированием.....	81
Архитектура.....	81
Создание механизма семантического поиска и генеративного агента для сферы БПЛА.....	84
Настройка среды.....	85
Компонент 1: сбор и подготовка документов.....	85
Компонент 2: создание и заполнение векторной базы данных Deep Lake.....	89
Компонент 3: RAG с индексированием.....	94
Механизм индексирования на основе векторов.....	96
Ответ на запрос и его источник.....	97
Оптимизация разбивки текста.....	99
Оценка эффективности.....	99
Механизм индексирования на основе дерева.....	100
Оценка эффективности.....	102
Механизм индексирования на основе списка.....	103
Оценка эффективности.....	105
Механизм индексирования на основе таблицы ключевых слов.....	105
Оценка эффективности.....	107
Резюме.....	108
Вопросы.....	109
Ссылки.....	109
Дополнительная информация.....	109
Глава 4. Мультимодальная RAG	110
Что такое мультимодальная RAG.....	111
Создание мультимодальной RAG для сферы БПЛА.....	114
Загрузка датасета для LLM.....	115
Загрузка и визуализация мультимодального датасета.....	117

Навигация по мультимодальному датасету	120
Создание мультимодального механизма обработки запросов.....	124
Обобщение мультимодального вывода.....	129
Оценка эффективности	130
Резюме.....	134
Вопросы	135
Ссылки	135
Дополнительная информация	135
Глава 5. Повышение эффективности RAG за счет обратной связи от эксперта	136
Адаптивная RAG	136
Создание гибридной адаптивной RAG на Python.....	140
1. Ретривер.....	140
2. Генератор	143
3. Оценка	153
Резюме.....	159
Вопросы	160
Ссылки	160
Дополнительная информация	160
Глава 6. Масштабирование банковских данных клиентов в системе RAG с помощью Pinecone	161
Масштабирование с помощью Pinecone.....	162
Архитектура.....	162
Компонент 1: сбор и подготовка данных.....	164
Сбор и обработка датасета.....	165
Изучение и анализ данных (EDA).....	169
Обучение модели ML	172
Компонент 2: масштабирование индекса Pinecone.....	176
Сложности управления векторными базами данных	177
Настройка среды	179
Обработка датасета.....	179
Создание чанков и их преобразование в эмбединги	181
Создание индекса Pinecone	185
Загрузка.....	186
Запрос к индексу Pinecone	189
Компонент 3: генеративный ИИ на базе RAG	191
RAG с GPT-4o	191
Извлечение релевантных текстов	194
Дополнение промта.....	194
Генерация на основе дополненного запроса.....	195

Резюме	197
Вопросы	198
Ссылки	198
Дополнительная информация	198
Глава 7. Создание масштабируемой RAG на основе графа знаний с помощью API Википедии и LlamaIndex	199
Архитектура RAG для семантического поиска на основе графа знаний	200
Создание графов из деревьев	203
Компонент 1: сбор и подготовка документов	206
Извлечение данных и метаданных из Википедии	206
Подготовка данных к загрузке	210
Компонент 2: создание и заполнение векторной базы данных Deep Lake	212
Компонент 3: RAG на основе индексируемого графа знаний	212
Генерация индекса графа знаний	214
Вывод графа	216
Взаимодействие с индексом графа знаний	217
Установка пакетов для измерения сходства и определение функций	218
Переупорядочивание	218
Примеры оценки	220
Резюме	225
Вопросы	225
Ссылки	226
Дополнительная информация	226
Глава 8. Создание динамических RAG с помощью Chroma и Hugging Face Llama	227
Архитектура динамической RAG	228
Настройка среды	231
Hugging Face	231
Chroma	233
Запуск отсчета времени сеанса	234
Скачивание и подготовка датасета	235
Создание эмбедингов и их загрузка в коллекцию Chroma	237
Выбор модели	238
Создание эмбедингов и сохранение завершенного текста	239
Вывод эмбедингов	240
Обращение к коллекции	241
Создание промта и извлечение ответа	245
RAG с Llama	247
Удаление коллекции	250

Общая продолжительность сеанса.....	250
Резюме.....	251
Вопросы.....	252
Ссылки.....	252
Дополнительная информация.....	252
Глава 9. Улучшение моделей ИИ: тонкая настройка данных RAG	
и обратная связь.....	253
Схема тонкой настройки статических данных RAG.....	254
Экосистема RAG.....	255
Настройка среды.....	256
Подготовка датасета к тонкой настройке.....	257
Скачивание и визуализация датасета.....	257
Подготовка датасета к тонкой настройке.....	259
Тонкая настройка модели.....	260
Отслеживание тонкой настройки.....	262
Использование настроенной модели OpenAI.....	264
Метрики.....	266
Резюме.....	269
Вопросы.....	270
Ссылки.....	270
Дополнительная информация.....	270
Глава 10. RAG для создания стоковых видео с помощью Pinecone и OpenAI	271
Архитектура RAG для создания видео.....	272
Среда для производства видео.....	275
Импорт модулей и библиотек.....	275
GitHub.....	276
OpenAI.....	277
Pinecone.....	278
Компонент 1: генератор и комментатор.....	278
Генерация датасета с помощью ИИ.....	279
Генератор и комментатор.....	287
Компонент 2: администратор векторной базы данных.....	295
Запрос к индексу Pinecone.....	296
Компонент 3: видеоэксперт.....	300
Резюме.....	305
Вопросы.....	305
Ссылки.....	306
Дополнительная информация.....	306
Приложение	307