ПРОСТОЕ РУКОВОДСТВО ПО RAG

Поисково-дополненная генерация

Абхинав Кимоти

БОМБОРА
ИЗДАТЕЛЬСТВО
Москва

УДК 004.8
ББК 32.813
К40

Abhinav Kimothi
A SIMPLE GUIDE TO RETRIEVAL AUGMENTED GENERATION

Кимоти, Абхинав.

К40 Простое руководство по RAG : поисково-дополненная генерация / Абхинав Кимоти ; [перевод с английского Ю. Н. Смирновой]. — Москва : Эксмо, 2026. — 288 с. — (Manning: профессиональные книги для ИТ-специалистов).

ISBN 978-5-04-232419-2

Это исчерпывающее руководство по технологии Retrieval Augmented Generation (RAG), которая делает современные большие языковые модели (LLM) точными и надежными. Абхинав Кимоти пошагово объясняет архитектуру RAG: от создания баз знаний и конвейеров индексирования до продвинутых стратегий оптимизации и оценки. Практические примеры на Python помогут инженерам и разработчикам легко внедрить RAG в свои проекты. Отличный инструмент для освоения передового AI — от основ до экспертных решений в реальных задачах.

УДК 004.8
ББК 32.813

ISBN 978-5-04-232419-2

Краткое содержание

Содержание

ЧАСТЬ 2. СОЗДАНИЕ RAG-СИСТЕМ

ЧАСТЬ 3. RAG В ПРОДАКШЕНЕ

ЧАСТЬ 4. ДОПОЛНИТЕЛЬНЫЕ СООБРАЖЕНИЯ

Предисловие

Каким образом машины понимают человеческие намерения? Этот вопрос всегда вызывал у меня глубокий интерес. Хотя мое путешествие в мир искусственного интеллекта (ИИ) и машинного обучения началось в 2007 году, по-настоящему я заинтересовался обработкой естественного языка (NLP, Natural Language Processing) в начале 2016-го, когда работал над созданием виртуального аналитика данных. Когда в 2018 году исследователи из Google представили языковую модель BERT, я убедился, что NLP находится на пороге настоящей революции.

В 2022-м, после релиза модели text-davinci-002 из серии GPT-3 от OpenAI, я решил присоединиться к Yarnit — платформе для контент-маркетинга на основе генеративного ИИ, чтобы разработать ее технологическую основу. Нашей целью было создать платформу, где корпоративные маркетинговые команды могли бы быстро, масштабно, с высокой точностью и минимальными затратами генерировать маркетинговые материалы: посты в соцсетях, блоги, электронные письма и многое другое. Вскоре стало очевидно, что без учета знаний о бренде и доступа к закрытым данным ни одна генеративная модель не справится с этой задачей сколь-нибудь эффективно. Это осознание подтолкнуло меня к изучению RAG (Retrieval-Augmented Generation, генерации с дополнением данными).

Большие языковые модели (LLM, Large Language Models) часто не оправдывают ожиданий пользователей. Они невероятно эффективны в хранении и генерации знаний, но при этом склонны к галлюцинациям — уверенно высказанным ошибочным выводам. Именно здесь RAG совершает прорыв, позволяя LLM получать релевантную, фактически точную и свежую информацию перед генерацией ответа. Прелесть RAG заключается в простоте концепции в сочетании с тонкостями реализации. Способность RAG радикально преодолевать ключевые ограничения LLM — вот что удерживает внимание как исследователей, так и практиков.

Когда я начал изучать RAG, эта область была еще мало исследована. Официальных материалов для изучения RAG практически не существовало, сведения были разбросаны по блогам, публикациям в соцсетях, научным статьям и дискуссионным форумам. Я активно делился собственными находками на социальных платформах и в блогах. Со временем у меня созрела идея объединить все эти знания в единую книгу.

Поставив перед собой цель создать простой и практичный ресурс для технических специалистов, разрабатывающих приложения на основе LLM, я приступил к работе над книгой в середине 2024 года. Со временем она превратилась в фундаментальное руководство по RAG, охватывающее тему как вширь, так и вглубь, с акцентом на практической реализации через понятные объяснения и лаконичный код на Python.

Я убежден, что знание RAG обязательно для каждого, кто работает с ИИ-приложениями, а его освоение требует прочной концептуальной базы. Эта книга призвана ее обеспечить. Работа над книгой дала мне невероятно ценный опыт, я многому научился, пока писал ее. Надеюсь, чтение моей книги окажется для вас столь же познавательным и увлекательным.

Благодарности

Чтобы написать книгу, требуются бесчисленные часы исследований и кропотливой работы, особенно когда речь идет о такой быстроразвивающейся области, как RAG, где новые исследования появляются почти каждую неделю. Эта книга не увидела бы свет без безграничной любви и поддержки моей жены Паллави (Pallavi). Ее поддержка и терпение помогли мне пройти этот путь, за что я ей бесконечно благодарен.

Я глубоко признателен своим сооснователям Джьотирмою (Jyotirmoy) и Акашу (Akash), а также всей команде Yarnit, которые внесли значительный вклад в мое понимание RAG. Практический опыт создания реальных ИИ-приложений, несомненно, обогатил эту книгу, сделав ее более ценной для читателей.

Также я хотел бы выразить искреннюю благодарность коллегам и наставникам: Ашишу Риши (Ashish Rishi), Сатьякаму Моханти (Satyakam Mohanty), Прадипте Мишре (Pradeepta Mishra), Мегхе Джон (Megha John), Сандипу Ачарье (Sandeep Acharya), Акшиту Шарме (Akshit Sharma), Вишалу Синхе (Vishal Sinha) и многим другим — за содержательные беседы и наставничество на протяжении многих лет. Их взгляды сформировали мою философию и подход к data science и ИИ.

Отдельная благодарность выдающейся команде Manning Publications во главе с Энди Валдроном (Andy Waldron) за предоставленную возможность.

Я глубоко признателен Яну Хафу (Ian Hough) за его бесценные советы в процессе написания книги. Особую признательность выражаю техническому редактору Артуро Гейгелю (Arturo Geigel) за тщательную проверку написанного и содержательные замечания, которые значительно улучшили книгу. Большое спасибо Азре Дедич (Azra Dedic) за существенное улучшение графики в книге, а также Робину Кэмпбеллу (Robin Campbell) и Аире Дуцич (Aira Ducic) за их выдающуюся работу по продвижению и маркетингу книги. Также приношу благодарность сотрудникам производственного отдела за усердный труд по подготовке книги к публикации.

Я глубоко признателен сообществу исследователей ИИ, чье неустанное стремление к знаниям и инновациям продолжает расширять границы возможного. Во многом эта книга отражает коллективные знания исследователей, участников open-source проектов и практиков, щедро делившихся своими идеями в статьях, блогах и на форумах.

Я искренне признателен всем рецензентам: Абхишеку Гупте (Abhishek Gupta), Алехандро Куэвасу Риверо (Alejandro Cuevas Rivero), Алексу Маклинтоку (Alex McLintock), Алиреза Агамохаммади (Alireza Aghamohammadi), Амиту Дикситу (Amit Dixit), Анидите Нат (Anindita Nath), Аниндьядипу Санниграхи (Anindyadeep Sannigrahi), Арьяну Джадону (Aryan Jadon), Ашишу Саркару (Ashish Sarkar), Аушиму Нагаркатти (Aushim Nagarkatti), Авинашу Тивари (Avinash Tiwari), Баблу Кумару (Babloo Kumar), Баладжи Дхамодарану (Balaji Dhamodharan), Балакришнану Баласубраманиану (Balakrishnan Balasubramanian), Берту Голлнику (Bert Gollnick), Бхаргобу Деке (Bhargob Deka), Брайану Дейли (Brian Daley), Чарану Акири (Charan Akiri), Кристоферу Дж. Фраю (Christopher G. Fry), Харчарану С. Каббаю (Harcharan S. Kabbay), Харшвардану С. Фартале (Harshwardhan S. Fartale), Игорю Свиленкову Божичу (Igor Svilenkov Božić), Иван Морено (Iván Moreno), Лалиту Чоурей (Lalit Chourey), Луису Луангкесорну (Louis Luangkesorn), Луи-Франсуа Бушарду (Louis-François Bouchard), Манасу Талукдару (Manas Talukdar), Марсио Ф. Ногейре (Márcio F. Nogueira), Марин Серре (Marine Serré), Наге Сантошу Редди Вутукури (Naga Santhosh Reddy Vootukuri), Нилеш Патерия (Neelesh Pateriya), Питеру Котронео (Peter Cotroneo), Питеру Моргану (Peter Morgan), Риче Талдар (Richa Taldar), Ридхибен Суниткумар Шах (Riddhiben Sunitkumar Shah), Роберту Винсу (Robert Vince), Самиту Сонаване (Sameet Sonawane), Сашанку Даре (Sashank Dara), Стивену Вольфу (Stephen Wolff), Субхашу Кумару Периасами (Subhash Kumar Periasamy), Винешу Гудле (Vinesh Gudla) и Янки Луо (Yanqi Luo) — за их ценные идеи и предложения, которые помогли улучшить качество этой книги.

В заключение я хочу сердечно поблагодарить всех, кто был частью этого пути. Ваша поддержка, мудрость и щедрость помогли воплотить в жизнь не только эту книгу, но и мою мечту стать автором.

О книге

RAG трансформирует сферу прикладного генеративного искусственного интеллекта. Термин RAG впервые был предложен Льюисом (Lewis) и его коллегами в основополагающей работе Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (*https://arxiv.org/abs/2005.11401*). Технология RAG быстро стала краеугольным камнем современного ИИ, повышая надежность и достоверность результатов больших языковых моделей (LLM).

«Простое руководство по дополненной генерации» — это фундаментальное пособие для всех, кто хочет изучить RAG. Книга представляет доступное, но всестороннее введение в концепцию, а также практические рекомендации, которые помогут эффективно использовать RAG в своих целях.

Для кого эта книга?

Книга предназначена для технических специалистов, которые хотят познакомиться с концепцией RAG и создавать приложения на основе LLM. Она будет полезна как новичкам, так и опытным профессионалам. Если вы специалист по данным, инженер данных, инженер машинного обучения, разработчик программного обеспечения, техлид или студент, интересующийся разработкой приложений на основе генеративного ИИ, вы найдете эту книгу полезной. После прочтения вы:

- разберетесь в основах, компонентах и практическом применении RAG;
- узнаете, как создавать непараметрические базы знаний и как с ними работать;

- сможете построить RAG-систему с глубоким погружением в пайплайн индексирования и генерации;
- получите полное представление об оценке RAG-систем и о модульных стратегиях оценки;
- познакомитесь с современными стратегиями и постоянно развивающимся ландшафтом RAG;
- приобретете знания об инструментах, технологиях и фреймворках для создания и развертывания систем RAG корпоративного уровня;
- узнаете о передовых вариантах RAG, таких как мультимодальный и агентный RAG;
- получите представление о текущих ограничения RAG и познакомитесь с популярными новейшими техниками для их дальнейшего изучения.

Эта книга является фундаментальным руководством и не требует от читателя глубокого понимания концепций, однако будет полезно предварительно познакомиться с миром машинного обучения, генеративного ИИ и LLM. Вы глубже поймете суть LLM, прочитав первую главу.

Книга содержит примеры кода на Python с использованием фреймворка LangChain. Важно отметить, что эти фрагменты кода служат лишь дополнительной иллюстрацией концепций и предназначены для читателей, которые хотят получить практический опыт. Для работы с этим кодом достаточно базового понимания Python и API.

Генеративный ИИ остается областью прорывных, непрерывно развивающихся технологий. Эта книга поможет вам повысить квалификацию и открыть новые возможности в текущей и будущей профессиональной деятельности.

Как организована эта книга: краткий обзор

Книга состоит из девяти глав, распределенных по четырем частям.

Часть 1 посвящена основополагающим понятиям RAG.

- Глава 1 начинается с определения RAG и объяснения его необходимости и значимости в той области ИИ, где применяются LLM. Здесь вы также найдете несколько практических примеров применения систем, использующих технологию RAG.
- Глава 2 посвящена разбору основных компонентов RAG-системы. В ней представлены два ключевых пайплайна: пайплайн индексации и пайплайн генерации. Кроме того, в главе затрагиваются концепции оценки RAG.

Часть 2 посвящена созданию базовой RAG-системы с основными пайплайнами и их оценке.

- В главе 3 мы рассмотрим сквозной пайплайн индексации для создания базы знаний RAG-системы. Вы на примерах познакомитесь

с загрузкой данных, разбиением на чанки, эмбеддингами (преобразованием текста в векторное представление) и векторным хранилищем.

- Глава 4 освещает пайплайн генерации, обеспечивающий доступ к базе знаний и взаимодействие с LLM для генерации контекстных и точных ответов в режиме реального времени. Мы поговорим о ретриверах, стратегиях извлечения релевантных данных и промпт-инжиниринге для RAG, а также получим общее представление о доступных LLM.
- В главе 5 подробно рассматриваются различные методы оценки RAG и проводится их анализ с точки зрения вопроса, ответа и контекста. Также мы обсудим создание эталонного набора данных (ground truth dataset) и его значимость. Из этой главы вы узнаете подробности о популярных фреймворках и бенчмарках для оценки RAG-систем.

Часть 3 предоставит руководство по улучшению RAG-пайплайна и расскажет о слоях, необходимых для построения RAG-системы, готовой для применения в продакшене.

- В главе 6 рассматриваются современные концепции RAG: наивный, продвинутый и модульный RAG. Мы поговорим о важных компонентах, а также о предпоисковых и постпоисковых стратегиях. В этой главе также представлены методы оптимизации для повышения производительности RAG-системы.
- Глава 7 знакомит с различными инструментами и технологиями, которые обеспечивают работу стека RAGOps (RAG Operations). Вы узнаете о критически важных слоях, без которых любая RAG-система не сможет функционировать, а также о слоях, необходимых для повышения производительности системы, и о дополнительных слоях, ориентированных на удобство использования, масштабируемость и эффективность.

В части 4 вы познакомитесь с популярными современными вариантами RAG и общей схемой разработки RAG.

- Глава 8 посвящена новейшим вариантам RAG, включая мультимодальный, графовый и агентный RAG.
- В завершающей, 9-й главе представлена общая схема разработки RAG, которая поможет вам спланировать создание RAG-системы.

Книга рассчитана на последовательное чтение, причем в последней главе дается обзор всех рассмотренных в книге концепций.

О коде

Все примеры кода в этой книге написаны на Python. Исполняемые фрагменты кода доступны в онлайн-версии книги на платформе liveBook по