Глава 3. Практическое применение RAG

Это продолжение перевода книги Кит Борн. Раскрытие потенциала данных с помощью генеративного ИИ и технологии RAG. В главе 1 мы перечислили несколько применений генерации, дополненной поиском (RAG), а в главе 2 разобрали код конвейера RAG.

В этой главе мы рассмотрим следующие темы:

- Поддержка клиентов и чат-боты
- Автоматизированная отчетность
- Поддержка электронной коммерции
- Использование баз знаний
- Поиск инноваций и анализ тенденций
- Персонализация контента для медиа и контент-платформ
- Персонализированные рекомендации в маркетинговых коммуникациях
- Обучение и образование
- Лаборатория кода 3. Добавление исходных кодов в RAG

Предыдущая глава Содержание Следующая глава

Примеры, представленные в этой главе, не претендуют на исчерпывающий характер, а скорее на то, чтобы предоставить реальные сценарии, демонстрирующие потенциал RAG. Изучив некоторые из этих примеров, вы получите ценную информацию о том, как можно использовать RAG для улучшения существующих процессов, повышения эффективности и стимулирования инноваций. В конце главы мы приведем код, продолжающий пример главы 2 для извлечения данных из документов в ответе запрос.

Мы начнем с обсуждения того, как улучшить чат-боты с помощью RAG и генеративного ИИ (GenAI). Код для этой главы размещен в репозитории GitHub.

Поддержка клиентов и чат-боты с RAG

Чат-боты эволюционировали от простых закодированных ответов до сложных, управляемых RAG разговорных агентов. RAG привнесла новую волну инноваций в чат-боты, включив передовые системы вопросов и ответов в возможности чат-ботов таким образом, чтобы они стали более разговорным и естественным для пользователя. RAG сочетает в себе лучшее из обоих миров: возможность извлекать информацию из обширных наборов данных о вашей компании и ваших клиентах и способность генерировать согласованные, контекстуально релевантные ответы. Это оказалось весьма перспективным в сценариях поддержки клиентов, где возможность быстрого доступа к данным, относящимся к компании, таким как прошлые взаимодействия с клиентами, ответы на часто задаваемые вопросы и документы поддержки, значительно повысила качество обслуживания клиентов.

RAG позволяет чат-ботам предоставлять персонализированные, эффективные и релевантные ответы на запросы пользователей таким образом, что производительность намного превосходит производительность более ранних моделей, которые полагались исключительно на предварительно запрограммированные ответы или базовую обработку естественного языка (NLP).

Это не просто техническое усовершенствование. Уровень, которого достигают чат-боты, представляет собой трансформационный сдвиг в том, как компании могут взаимодействовать с клиентами. GenAI позволил компаниям гораздо глубже изучить все данные, которые у них есть о каждом клиенте, например, прочитать все ваши банковские выписки в формате PDF, что позволило настроить обслуживание один-на-один для каждого клиента. Чат-боты, усовершенствованные RAG, могут просеивать огромные объемы данных, чтобы найти наиболее релевантную информацию, эффективно отвечая на гораздо больше вопросов, возникающих у пользователя, гарантируя, что ответы будут точными и адаптированными к конкретному контексту каждого взаимодействия.

Клиенты ожидают быстрых, актуальных и персонализированных ответов на свои запросы, и использование RAG является наиболее эффективным способом их доставки. Имейте в виду, что системы на основе RAG все еще находятся в зачаточном состоянии. Всего через пару лет можно ожидать, что чат-боты на основе RAG раздвинут границы гораздо шире. Вы увидите разговорных чат-ботов, которые обрабатывают самые уникальные и специфические запросы без необходимости

вмешательства человека. Их естественный язык (natural language, NL), способность работать с большинством языков и доступ к огромным объемам памяти и вычислительной мощности произведут революцию в том, как компании будут взаимодействовать с клиентами.

Преимущества RAG в вопросах и ответах и чат-ботов распространяются на различные сектора, такие как техническая поддержка, финансовые услуги, здравоохранение и электронная коммерция. Мы кратко рассмотрим некоторые из этих популярных примеров, начиная с технической поддержки.

Техническая поддержка

Рассмотрим случай, когда возникает повторяющаяся техническая проблема. Например, большая часть обращений к кабельным провайдерам связана с одними и теми же техническими проблемами. Чат-бот с улучшенной RAG-версией может распознать проблему на основе предыдущих взаимодействий и немедленно предоставить индивидуальный набор шагов по устранению неполадок, учитывая прошлые попытки клиента решить проблему и корректируя текущий ответ. Это не только демонстрирует понимание клиентского опыта, но и укрепляет доверие и уверенность в процессе поддержки.

Финансовые услуги

Чат-боты, усовершенствованные RAG для финансовых услуг, также могут помочь с запросами по счетам, проблемами транзакций и персонализированными финансовыми консультациями, опираясь на историю транзакций клиента и детали счета. Когда вы в последний раз пытались узнать процентную ставку по кредитной карте?

На этот вопрос легко ответить, но для банков часто не так просто помочь агенту по обслуживанию клиентов, поскольку данные похоронены в базах данных, PDF-документах и за барьерами безопасности. Но после того, как вы безопасно идентифицировали себя в Интернете, банковский чат-бот может использовать RAG, чтобы получить доступ ко всем вашим финансовым документам и быстро ответить 21,9%, а также предоставить сопутствующую информацию, такую как: «Вы платили проценты только дважды».

Если вам затем нужно поговорить о своем ипотечном счете, вам не нужно переключаться на другого агента, так как чат-бот на основе RAG может помочь вам и в этом! Он может справиться и с более сложными вопросом: «Насколько процентная ставка по моей кредитной карте больше, чем по моей ипотеке?» Кроме того, если вы зададите дополнительный вопрос, он поймет контекст всего разговора и сможет говорить о ваших финансовых счетах очень человеческим и естественным образом. Такой уровень поддержки может значительно повысить качество обслуживания клиентов, повышая лояльность и доверие к услугам компании в гораздо большей степени, чем это могут сделать существующие системы.

Здравоохранение

В здравоохранении чат-боты на базе RAG могут оказывать поддержку пациентам, получая доступ к медицинским записям (с соответствующими разрешениями), чтобы предлагать персонализированные советы по здоровью или облегчать запись на прием. Если пациент спрашивает о лечении недавно диагностированного диабета, чат-бот может проанализировать его историю болезни, лекарства и последние результаты лабораторных исследований, чтобы предоставить индивидуальные рекомендации по изменению образа жизни, диете и управлению лекарствами. Чат-бот даже может назначать последующие визиты к врачу и отправлять напоминания, создавая более комплексную поддержку пациентов. Такой уровень персонализированного ухода может значительно улучшить результаты лечения пациентов, повысить доверие к системе здравоохранения и снизить нагрузку на медицинских работников.

Однако RAG предназначена не только для чат-ботов. Давайте обсудим еще одну область, в которой RAG активно работает: автоматизация отчетности для анализа данных.

RAG для автоматизированной отчетности

Компании, которые используют RAG для анализа данных и автоматизированной отчетности, отмечают значительное улучшение своих возможностей и времени, необходимого для выполнения анализа. Это инновационное приложение RAG служит мостом между огромными озерами неструктурированных данных и практическими выводами, которые ежедневно необходимы предприятиям для принятия ключевых решений и внедрения инноваций. Используя RAG для

автоматизированной отчетности, компании могут значительно оптимизировать свои процессы отчетности, повысить точность и выявить ценную информацию, скрытую в данных.

Как используется RAG с автоматизированной отчетностью

Несмотря на то, что существует бесчисленное множество способов настройки этого типа автоматизированной отчетности, обычно она ориентирована на области, где отчеты являются относительно стандартными, что упрощает их кодификацию в среде RAG. Часто вы начинаете с автоматизированного отчета, который уже есть с точки зрения кода, но вы можете применить методы GenAI и/или RAG, чтобы сделать эту автоматизацию еще более эффективной. Например, набор вопросов для первоначального анализа может быть передан в систему RAG без ввода данных пользователем, и этот контент может быть добавлен в первоначальный отчет. Этот первоначальный анализ может включать в себя не только диаграммы и графики, но и комментарии из большой языковой модели (LLM), и все это в качестве начального компонента конвейера RAG.

Во многих сценариях автоматический отчет часто является лишь первым шагом, помогающим лицу, принимающему решения понять данные. Очень часто лица, принимающие решения, просят провести дополнительный анализ на основе того, что они видят в автоматизированном отчете. Если вы сделаете свои автоматизированные отчеты частью более крупной системы RAG, это может существенно заменить и/или ускорить обмен данными, который обычно используется в этих сценариях анализа данных, помогая лицам, принимающим решения, быстрее получить доступ к анализу ключевых данных и принимать более эффективные решения.

Вы можете сделать данные отчета, а также исходные данные, на основе которых он был сгенерирован, доступными для системы RAG, что, в свою очередь, позволяет нетехническому персоналу задавать многочисленные и очень широкие вопросы по данным, не дожидаясь дополнительного анализа, который будет выполнен аналитиком данных. К источникам, доступным для системы RAG, могут быть добавлены дополнительные источники данных, что обеспечивает еще большую глубину обсуждения и анализа, потенциально намного превосходящую то, что может быть предоставлено в рамках простого автоматизированного анализа.

Преобразование неструктурированных данных в полезную аналитику

Подавляющее большинство данных, доступных сегодня организациям, не структурированы, начиная от статей и исследовательских работ и заканчивая лентами социальных сетей и веб-контентом. Эта неструктурированная природа усложняет задачу обработки и анализа с помощью традиционных средств анализа данных. RAG выступает в качестве мощного подхода к анализу этих данных, извлекая информацию из данных, которые традиционно были труднодоступны, создавая первоначальные черновики и резюме, которые можно далее настроить под конкретного человека, выделяя важную для этого человека информацию на основе его роли или интересов.

RAG может служить гораздо более сложным инструментом для пользователей по сравнению с непосредственным использованием LLM, поскольку она может заменить многие задачи, которые им обычно приходится выполнять вручную при взаимодействии с LLM. Эта концепция также может быть применена в мире анализа данных и автоматизированной отчетности, где многие шаги могут быть воспроизведены в системе RAG, что делает их более быстрыми и эффективными. Этот процесс не только экономит время, но и гарантирует, что лица, принимающие решения, смогут быстро понять суть данных.

Важно отметить, что когда речь идет о неструктурированных данных, это часто шаг на этапе индексации RAG, где эти данные преобразуются в новый формат, что делает их более полезными для всей системы RAG. Например, PDF-файл может быть распакован на множество различных элементов с разными уровнями важности. Заголовки могут иметь больший вес, чем абзацы, по уровню их важности. Изображения могут быть определены как таблицы, которые затем могут быть извлечены в виде таблиц, векторизованы и использованы в системе RAG. Именно эти шаги делают неструктурированные данные более доступными для автоматизированной отчетности.

При поиске областей в вашей организации, на которых следует сосредоточиться для автоматизированной отчетности, начните с тех, где своевременная информация является наиболее важной. Например, при анализе рынка RAG может быстро обобщить новостные статьи, финансовые отчеты и информацию о конкурентах, предоставляя компаниям сжатое представление о рыночных

тенденциях и динамике. Такая быстрая обработка позволяет быстро реагировать на изменения рынка, извлекать выгоду из возможностей или оперативно снижать риски для вашей компании.

Совершенствование процесса принятия решений и стратегического планирования

Возможности автоматизированной отчетности и поиска инноваций RAG значительно улучшают процессы принятия решений и стратегического планирования. Предоставляя руководителям и стратегам краткие, обобщенные отчеты и информацию об отраслевых тенденциях, технологических достижениях и конкурентной среде, RAG позволяет принимать обоснованные и стратегические решения. Добавление ориентированной на RAG возможности задавать дополнительные вопросы о данных, которые создали эти отчеты, значительно повышает ценность для пользователей.

Инициативы RAG в этой области зависят от их способности быстро усваивать и анализировать различные наборы данных. Это позволяет компаниям, использующим RAG, применять более проактивный подход к формулированию стратегии. Вместо того, чтобы реагировать на изменения в отрасли, компании могут предвидеть изменения на рынке или в технологиях и соответствующим образом корректировать свои стратегии, гарантируя, что они остаются на шаг впереди конкурентов.

Поддержка электронной коммерции

Электронная коммерция является ключевой областью, которая может значительно выиграть от применения RAG. Давайте рассмотрим пару областей, где можно применить RAG, начиная с описаний товаров.

Динамические онлайн-описания товаров

Способность RAG создавать персонализированные описания продуктов меняет правила игры для бизнеса электронной коммерции. Используя возможности RAG, компании могут создавать целенаправленные и убедительные описания продуктов, которые находят отклик у клиентов, стимулируя продажи и повышая лояльность к бренду. RAG может создавать персонализированные описания продуктов или выделять функции, специально адаптированные к прошлому поведению и предпочтениям пользователя, принимая во внимание способность RAG анализировать огромные объемы данных о клиентах, включая историю просмотров, прошлые покупки и даже взаимодействия в социальных сетях.

Предположим, что Райли — пользователь, который часто покупает экологически чистые продукты. RAG можно использовать для того, чтобы подчеркнуть аспекты экологичности продукта в его описании, когда Райли просматривает веб-сайт электронной коммерции. Это приводит к улучшению пользовательского опыта для Райли, а также к повышению вероятности ее покупки этого продукта. Теперь предположим, что Райли ранее купил такой товар, как кроссовки, и часто работает с контентом, связанным с марафонскими тренировками. Когда Райли просматривает кроссовки для бега на сайте электронной коммерции на базе RAG, описания продуктов могут быть сгенерированы динамически, чтобы подчеркнуть такие характеристики, как амортизация, устойчивость и долговечность — все это ключевые факторы для бегунов на длинные дистанции. Описание также может включать дополнительную информацию, относящуюся к интересам Райли, например, о том, как обувь была протестирована марафонцами или была разработана для снижения риска распространенных травм при беге.

Кроме того, RAG можно использовать для анализа отзывов клиентов, чтобы выявить наиболее часто упоминаемые плюсы и минусы продукта. Эта информация может быть включена в описания товаров, предоставляя потенциальным покупателям более сбалансированный и информативный обзор товара. Решая распространенные проблемы или выделяя высоко оцененные функции, описания, созданные RAG, могут помочь клиентам принимать обоснованные решения о покупке, снижая вероятность возвратов или негативных отзывов.

RAG также можно использовать для создания описаний продуктов на нескольких языках, что облегчает компаниям выход на международные рынки. Обучая RAG на разнообразном наборе языковых данных, компании могут гарантировать, что описания их продуктов соответствуют культурным особенностям и находят отклик у клиентов в разных регионах.

Товарные рекомендации для сайтов электронной коммерции

Рекомендательные системы являются важным применением ИИ в деловом мире. GenAI и RAG могут сделать эти рекомендательные механизмы еще более эффективными. Одним из ключевых

преимуществ использования RAG для рекомендаций продуктов является его способность анализировать огромные объемы данных о клиентах, включая историю просмотров, прошлые покупки, поисковые запросы и даже взаимодействия в социальных сетях. Понимая уникальные интересы, предпочтения в стиле и покупательские привычки клиента, RAG может генерировать рекомендации по продуктам, которые не только актуальны, но и очень привлекательны для каждого человека. RAG может выявлять закономерности и предпочтения на гораздо более глубоком уровне, чем предыдущие методы, чтобы рекомендовать продукты, которые пользователь может найти наиболее привлекательными. RAG позволяет нам выйти далеко за рамки традиционных рекомендательных систем, интегрируя более глубокое понимание индивидуальных предпочтений пользователей, что приводит к точным и персонализированным рекомендациям.

Предположим, Обри – один из VIP-клиентов, часто покупает снаряжение для активного отдыха и в последнее время просматривает походные ботинки на сайте электронной коммерции. RAG может анализировать данные Обри и рекомендовать не только наиболее подходящие походные ботинки на основе ее предпочтений и прошлых покупок, но и дополнительные товары, такие как походные носки, рюкзаки и туристические палки. Представляя тщательно подобранные товары, которые соответствуют интересам Обри, RAG может повысить вероятность того, что Обри совершит покупку нескольких товаров, и улучшить общее впечатление от покупок.

Более того, RAG может вывести рекомендации по продуктам на новый уровень, учитывая факторы, выходящие за рамки просто истории покупок клиента. Например, RAG может анализировать отзывы и рейтинги клиентов о продуктах, чтобы получить представление об удовлетворенности предыдущими покупками. Эта информация может быть использована для уточнения будущих рекомендаций, гарантируя, что клиентам будут представлены продукты, которые не только соответствуют их интересам, но и соответствуют их ожиданиям в области качества.

Применение RAG для персонализации рекомендаций по продуктам выходит за рамки веб-сайтов электронной коммерции. Используя возможности RAG, компании могут предоставлять целенаправленный и привлекательный опыт в различных точках соприкосновения, включая медиа, контент-платформы и цифровые маркетинговые кампании.

Поскольку электронная коммерция продолжает расти и развиваться, важность персонализированных описаний продуктов невозможно переоценить. Используя возможности RAG, компании могут создавать описания, которые не только информируют, но и убеждают, стимулируя продажи и выстраивая долгосрочные отношения с клиентами. Во многом это связано со способностью RAG выполнять поиск в огромных объемах данных. Далее давайте обсудим, как те же самые возможности поиска и доступа к данным могут помочь компании лучше использовать знания внутри компании.

Использование баз знаний с помощью RAG

RAG может получать доступ к базам знаний, как внутренним, так и внешним, и использовать их.

Возможность поиска и полезность внутренних баз знаний

Сочетая концепцию расширенного поиска информации с самыми передовыми LLM, RAG показывает значительные успехи в области внутренних поисковых систем. Сочетание расширенных возможностей поиска и аналитики трансформирует корпоративные операции, позволяя нам получать доступ к данным гораздо более сложными способами, лучше используя огромные объемы данных, которые накапливают компании. Эта революционная технология не только упрощает поиск информации, но и повышает качество представляемых данных, делая ее краеугольным камнем для принятия решений и повышения операционной эффективности во многих секторах.

RAG значительно расширяет возможности поиска и полезность внутренних баз знаний, выступая в качестве катализатора для улучшения доступа к информации и управления ею. Внутренние базы знаний, которые включают в себя широкий спектр документов в различных неструктурированных форматах (PDF, Word, Google Docs, электронные таблицы, слайды и т. д.), часто используются недостаточно из-за их обширности и сложности извлечения. RAG может решить эту проблему несколькими способами, извлекая данные и обрабатывая их. Например, создание кратких резюме документов облегчает сотрудникам понимание сути содержимого без необходимости просматривать весь документ. Более того, RAG может предоставлять прямые ответы на запросы, анализируя содержимое этих ресурсов и используя генеративную силу LLM для предоставления согласованных и точных ответов на контент, который может быть похоронен в миллионах страниц

неструктурированных данных. Эта возможность прямого ответа особенно полезна для быстрого решения конкретных запросов, что может привести к значительному сокращению времени, затрачиваемого сотрудниками на поиск информации.

Внедрение RAG во внутренние поисковые системы также способствует более организованному и эффективному способу управления знаниями. Благодаря лучшей категоризации и поиску информации сотрудники могут быстрее получать доступ к соответствующим данным, что оптимизирует работу. Это особенно полезно в быстро меняющихся условиях, где время имеет решающее значение, а быстрый доступ к точной информации может значительно повлиять на принятие решений и результаты проекта.

В то время как внутренние знания являются ключевым стратегическим преимуществом для большинства компаний, внешние знания также могут играть важную роль в поддержании конкурентного преимущества в вашей отрасли. Давайте рассмотрим, как RAG позволяет сотрудникам компании лучше использовать внешние данные.

Использование внешних баз знаний

Помимо внутренних ресурсов, RAG распространяет свои преимущества на внешние базы знаний, что имеет решающее значение в областях, требующих актуальных знаний законов, нормативных актов, отраслевых стандартов, соблюдения требований к исследованиям в медицине, знаний патентов. В этих областях объем информации не только огромен, но и постоянно обновляется, что затрудняет поддержание актуальности. RAG упрощает эту задачу, извлекая и обобщая соответствующую информацию из обширных баз данных. Например, в юридическом секторе и комплаенсе RAG может быстро просмотреть тысячи документов, чтобы найти соответствующие прецедентные законы, нормативные акты и руководства по соблюдению нормативных требований, что значительно сокращает время, затрачиваемое юристами и специалистами по соблюдению нормативных требований.

Аналогичным образом, в области исследований и разработок RAG может ускорить процесс, предоставляя исследователям быстрый доступ к актуальным исследованиям, патентам и техническим документам, имеющим отношение к их работе. Эта возможность неоценима для того, чтобы избежать дублирования усилий и зажечь новые идеи на основе существующих знаний. В области медицины способность RAG извлекать соответствующие тематические исследования, научные статьи и рекомендации по лечению может помочь медицинским работникам в принятии обоснованных решений, улучшая уход за пациентами.

Одна из основных проблем с использованием LLM, таких как ChatGPT, для проведения собственных исследований заключается в том, что они часто могут дать вам вымышленную, но убедительную информацию (называемую галлюцинациями). RAG — отличный способ уменьшить эти галлюцинации, поскольку она сохраняет компонент LLM в конвейере RAG на основе реальных данных, что сокращает время, затрачиваемое на проверку ответов. Вы можете добавить несколько вызовов к LLM в конвейере RAG, которые не только помогут ответить на ваши исследовательские вопросы, но и проверят, что ответ имеет отношение к исходному вопросу, прежде чем предоставить ответ. Вы также можете настроить вывод таким образом, чтобы предоставить всю подтверждающую документацию и цитаты.

Использование внешних данных с помощью RAG может быть не только доступом к общим базам знаний. Давайте обсудим концепцию инновационного поиска и то, как организации используют RAG для дальнейшего внедрения инноваций и более быстрого выявления тенденций в своей отрасли.

Поиск инноваций и анализ тенденций

Сканирование и обобщение информации из различных качественных источников также может сыграть важную роль в поиске инноваций и анализе тенденций. RAG помогает компаниям определить новые тенденции и потенциальные области для инноваций, которые соответствуют их специализации. Это особенно актуально в быстро развивающихся отраслях, где для сохранения конкурентных преимуществ важно оставаться на шаг впереди.

В технологическом секторе RAG может анализировать патенты, технические новости и научные публикации для выявления новых технологий и инновационных моделей. Это позволяет компаниям

более эффективно направлять свои усилия в области исследований и разработок, фокусируясь на областях с высоким потенциалом роста и прорыва рынка.

В отрасли, в которой я сейчас работаю, в фармацевтической промышленности, RAG используется для ускорения процесса выявления новых результатов исследований и потенциальных возможностей разработки лекарств путем анализа свежих медицинских журналов, отчетов о клинических испытаниях и патентных баз данных. Это ускоряет темпы инноваций и помогает фармацевтическим компаниям эффективно распределять бюджеты и ресурсы на исследования.

Применение RAG выходит за рамки поиска инноваций и анализа тенденций. Еще одна область, в которой RAG добивается значительных успехов, — это персонализация контента для медиа и контент-платформ. В сегодняшнем цифровом ландшафте, где пользователи завалены огромным количеством контента, RAG предлагает мощное решение для устранения шума и предоставления узконаправленного, персонализированного опыта.

Далее давайте обсудим этот целостный подход к персонализации, который может привести к повышению лояльности клиентов, повышению коэффициента конверсии и более успешному бизнесу.

Использование RAG для персонализированных рекомендаций в маркетинговых коммуникациях

RAG также может стать значительным шагом вперед для компаний, которые внедряют его для повышения вовлеченности и удовлетворенности пользователей в средствах массовой информации, на контент-платформах и в цифровых маркетинговых кампаниях. RAG можно использовать для создания персонализированных наборов продуктов или коллекций на основе предпочтений клиента, которые можно рекламировать и включать в цифровые кампании, нацеленные на каждого клиента. Анализируя данные о клиентах и определяя дополнительные продукты, RAG может предложить предварительно собранные комплекты, которые обеспечивают удобство и ценность.

RAG также может повысить эффективность кампаний e-mail-маркетинга, предоставляя персонализированные рекомендации по продуктам прямо в маркетинговых сообщениях. Анализируя данные клиента и адаптируя предложения продуктов к его интересам, интернет-магазины могут создавать высоко таргетированный и привлекательный контент для писем, который увеличивает показатели кликабельности (CTR) и конверсии.

По мере того, как маркетинг продолжает развиваться, а компании ищут новые способы привлечения клиентов, потенциал RAG выходит за рамки персонализированных рекомендаций по продуктам на веб-сайтах. RAG можно использовать для создания персонализированного опыта в различных цифровых точках взаимодействия, повышая вовлеченность пользователей и способствуя долгосрочной лояльности клиентов.

Далее давайте рассмотрим, как возможности RAG могут быть применены в другой важной для бизнеса области: тренингах и обучении сотрудников.

Обучение и образование

RAG может использоваться образовательными организациями, университетами и средней школой. Она также может быть использована во внутренних корпоративных программах обучения, чтобы держать сотрудников в курсе огромного количества постоянно меняющихся концепций, о которых они должны знать. RAG помогает создавать или настраивать учебные материалы на основе конкретных потребностей, уровня знаний и функций учащихся.

RAG продемонстрировала невероятные перспективы для улучшения обучения и развития в корпоративной среде. Для предприятий часто бывает непростой задачей держать своих сотрудников в курсе последних отраслевых знаний и навыков, особенно учитывая сегодняшние быстрые темпы изменений во многих отраслях. RAG помогает решить эту проблему, предлагая персонализированное обучения, которое адаптируется к потребностям каждого сотрудника, стилю и темпу их обучения.

RAG может анализировать роль сотрудника, его опыт и историю обучения, чтобы составить индивидуальную схему обучения, которая фокусируется на наиболее актуальных и необходимых навыках. Такой подход гарантирует, что сотрудники проходят обучение, которое напрямую способствует их профессиональному росту и соответствует целям компании.

Кроме того, RAG также можно использовать для создания интерактивных учебных материалов, таких как викторины, тематические исследования и симуляции, адаптированных к стилю обучения и прогрессу каждого сотрудника. Такой адаптивный подход к обучению поддерживает вовлеченность и мотивацию сотрудников, поскольку они получают контент, который ставит перед ними задачи на нужном уровне и обеспечивает немедленную обратную связь об их работе.

Способность RAG быстро обобщать и представлять актуальную информацию из обширных баз знаний также делает ее бесценным инструментом для обучения и поддержки производительности на рабочем месте. Сотрудники могут использовать систему RAG для быстрого доступа к информации, необходимой им для решения проблем, принятия решений и более эффективного выполнения задач. Такой подход к обучению «точно в срок» (JIT) снижает потребность в длительных формальных учебных занятиях и дает сотрудникам возможность контролировать свое обучение и развитие.

RAG также обеспечивает эффективное сотрудничество и обмен знаниями между сотрудниками. Анализируя знания и навыки каждого человека, RAG может определить экспертов в предметной области в организации и облегчить связи между сотрудниками, которые могут учиться друг у друга. Этот внутренний поток знаний способствует формированию культуры непрерывного обучения и помогает организациям сохранять и использовать свой коллективный опыт.

RAG обладает огромным потенциалом для трансформации различных аспектов деятельности компании, от персонализированного обучения и развития сотрудников до взаимодействия с клиентами и оптимизации процессов. Тем не менее, успешное внедрение RAG требует стратегического подхода, который согласуется с целями и приоритетами компании.

Одним из важнейших аспектов многих приложений RAG, упомянутых ранее, является включение соответствующих данных и источников в генерируемые ответы. Например, при использовании RAG для сканирования юридических документов или научно-исследовательских работ крайне важно ссылаться на источники, чтобы обеспечить достоверность и поддержку представленной информации. Давайте рассмотрим, как мы можем расширить пример кода из главы 2 , чтобы включить эту функциональность.

Лаборатория кода 3. Добавление исходных кодов в RAG

Многие из упомянутых выше приложений включают в себя добавления данных в ответ. Например, вы захотите процитировать источники ответа, если у вас есть конвейер RAG, который сканирует юридические документы или научно-исследовательские работы. Мы продолжим работу над кодом из главы 2 и добавим шаг возврата полученных документов в ответ RAG.

```
from langchain_core.runnables import RunnableParallel
```

Это новый импорт: объект RunnableParallel из LangChain runnables. Это вводит концепцию параллельного запуска ретривера и вопроса. Это может повысить производительность, позволяя ретриверу получать контекст во время одновременной обработки вопроса:

```
rag_chain_from_docs = (
   RunnablePassthrough.assign(context=(
      lambda x: format_docs(x["context"])))
   | prompt
   | llm
   | StrOutputParser()
)

rag_chain_with_source = RunnableParallel(
   {"context": retriever,
    "question": RunnablePassthrough()}
).assign(answer=rag_chain_from_docs)
```

Сравните это с нашим исходным объектом rag_chain:

```
rag_chain = (
   {"context": retriever | format_docs,
    "question": RunnablePassthrough()}
```

```
    | prompt
    | llm
    | StrOutputParser()
)
```

В оригинальном коде rag_chain создается с использованием словаря, который сочетает в себе retriever и функцию format_docs для "context" и RunnablePassthrough() для "question". Затем этот словарь передается по конвейеру (|) через prompt, в llm и StrOutputParser().

В новой версии с названием rag_chain_from_docs, конструкция rag_chain разделена на две части:

- rag_chain_from_docs создается с помощью RunnablePassthrough.assign() для форматирования документов, извлеченных из контекста. Затем он передает отформатированный контекст через prompt в llm и StrOutputParser().
- rag_chain_with_source создается с помощью RunnableParallel() для параллельного запуска retriever и RunnablePassthrough() для "context" и "question" соответственно. Затем результат присваивается «ответу» с помощью rag_chain_from_docs.

Основное различие в функциональности между этими двумя подходами заключается в том, что новый подход отделяет извлечение контекста от форматирования и обработки полученных документов. Это обеспечивает большую гибкость в обработке полученного контекста перед его передачей через командную строку, LLM и выходной синтаксический анализатор.

Наконец, нам нужно изменить имя цепочки, которую мы передаем пользовательскому запросу, чтобы оно соответствовало новому имени цепочки, rag_chain_with_source. Как и в прошлом, мы вызываем метод invoke, rag_ chain_with_source.invoke(), передавая ему вопрос, который запускает параллельное выполнение retriever и вопроса, за которым следует форматирование и обработка полученного контекста с использованием rag_chain_from_docs для генерации окончательного ответа:

rag_chain_with_source.invoke("What are the Advantages of using RAG?")

Вывод будет выглядеть следующим образом (я сократил часть текста, чтобы вместить его в эту книгу, но вы должны увидеть полный ответ при запуске этого коде):

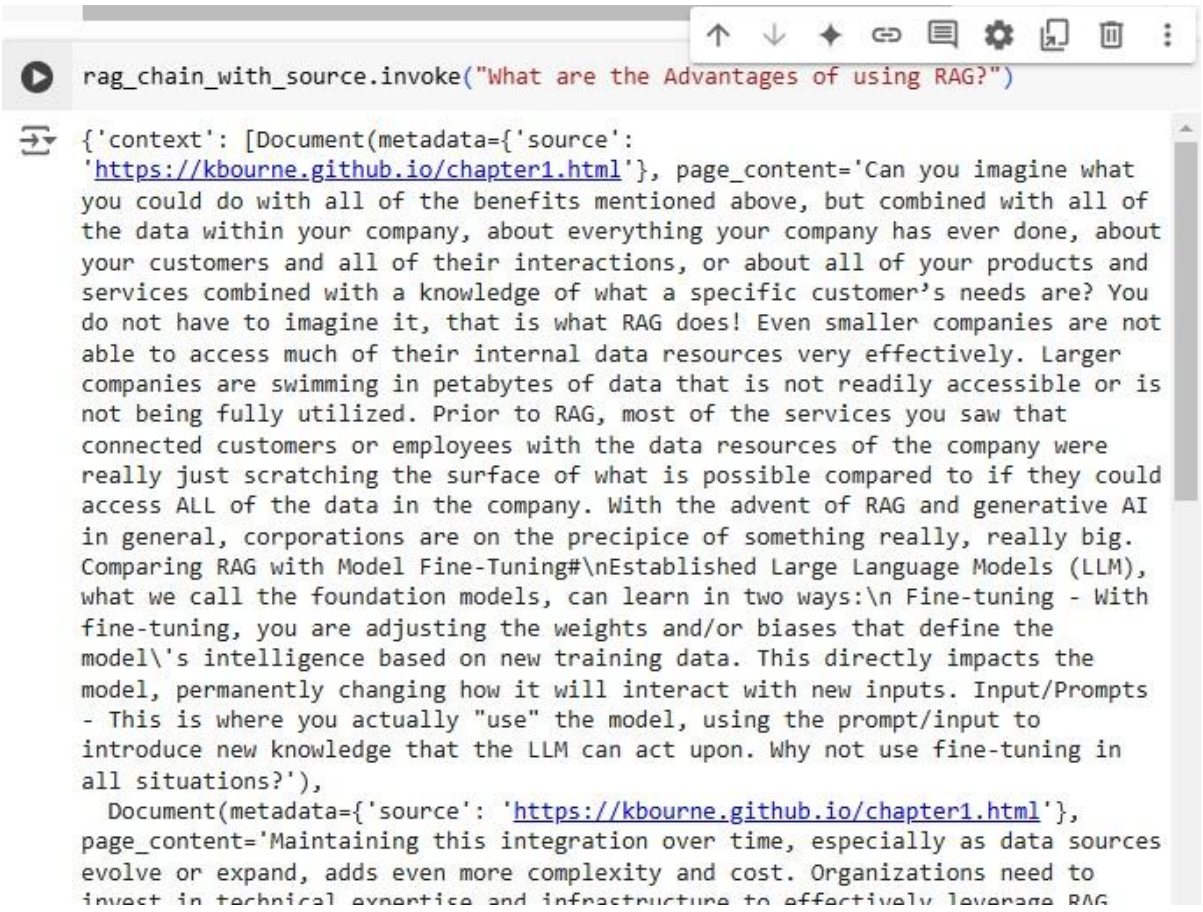

Рис. 3.1. Вывод запроса

Это больше похоже на код, чем наш предыдущий окончательный вывод, но он содержит всю информацию, которую вы могли бы предоставить пользователю для указания источника ответа, который вы ему предоставили. Во многих случаях этот источник материала очень важен для того, чтобы помочь пользователю понять, почему реакция была такой, какой она была, и проверить источник.. Обратите внимание на источник метаданных, указанный после каждого экземпляра page_content , который вы должны предоставить в качестве ссылки на источник. В ситуациях, когда в результатах поиска есть несколько документов, это может выглядеть по-разному для каждого отдельного документа, возвращенного на этапе извлечения, но здесь мы используем только один документ.

В следующей главе мы углубимся в технические компоненты, составляющие систему RAG, исследуем тонкости индексации, извлечения и генерации, а также то, как эти этапы интегрируются для предоставления возможностей, которые мы обсуждали в этой главе. Разбивая каждый компонент, вы получите подробное представление о внутренней работе и о том, как они взаимодействуют друг с другом, создавая улучшенные генеративные результаты.