В сфере искусственного интеллекта нет ничего более парадоксального, чем скорость устаревания знаний. За то время, пока готовится к печати очередная техническая книга, появляются новые фреймворки, алгоритмы и подходы, которые кардинально меняют ландшафт технологий. RAG-системы не исключение – эта область развивается настолько стремительно, что любая попытка зафиксировать «окончательное» знание обречена на устаревание ещё до выхода из типографии. Однако именно поэтому необходимо создавать фундаментальные руководства, которые не просто описывают текущие инструменты, а дают глубокое понимание принципов работы технологий. Эта книга – не попытка угнаться за трендами, а стремление зафиксировать базовые концепции и архитектурные решения, которые останутся актуальными независимо от смены версий библиотек и появления новых стартапов. Издание предназначено разработчикам, архитекторам систем, продуктовым менеджерам и всем, кто хочет не просто использовать готовые RAG-решения, но понимать, как они устроены изнутри. Каждая глава сочетает теоретические основы с практическими примерами, позволяя читателю не только изучить концепции, но и немедленно применить их на практике. В эпоху, когда ИИ-системы становятся критически важной инфраструктурой, понимание принципов их работы превращается из преимущества в необходимость.

Содержание

Предисловие

Дорогие читатели!

Более тридцати лет я занимаюсь искусственным интеллектом. За эти годы я воочию видел, как наша отрасль переживала и зимы, и весны, как горделивые надежды сменялись разочарованием, а скепсис – новыми прорывами. Но последние несколько лет стали особенными. Мы стали свидетелями появления технологий, которые впервые приблизили нас к мечте о настоящем искусственном интеллекте.

RAG-системы – не столько очередная аббревиатура в бесконечной очереди терминов, но фундаментальный сдвиг в том, как мы представляем взаимодействие между машинами и человеческими знаниями. За последний год я написал более 40 статей и заметок о различных аспектах RAG-технологий, опубликованных в телеграм-канале «Технооптимисты» (https://t.me/drv_official). Каждая из них была попыткой зафиксировать момент истины в стремительно развивающейся области.

Когда издательство «ДМК Пресс» предложило систематизировать эти знания в книге, я понял, что стою перед классической дилеммой технического писателя: как написать книгу о технологии, которая развивается быстрее, чем пишется текст. Каждую неделю появляются новые фреймворки, подходы, архитектурные решения. К моменту выхода книги многие конкретные инструменты и версии библиотек неизбежно устареют.

Но именно поэтому эта книга нужна. В эпоху информационного шума критически важно не просто следить за новинками, но понимать фундаментальные принципы. RAG-системы – это не мода, это новая парадигма работы с информацией. И как любая парадигма, она имеет базовые законы, архитектурные принципы и философию, которые останутся актуальными независимо от смены конкретных инструментов.

Я видел, как компании тратят месяцы на внедрение RAG-решений, не понимая базовых принципов их работы. Как разработчики пытаются «прикрутить» векторный поиск к существующим системам, получая в результате дорогие и неэффективные решения. Как менеджеры продуктов ожидают от RAG-систем магии, а получают очередной источник галлюцинаций и технического долга.

Цель этой книги – дать читателю именно то понимание, которого так часто не хватает на практике. Не просто «как запустить библиотеку X с параметрами Y», а «почему эта архитектура работает именно так», «какие компромиссы скрываются за каждым решением», «как оценить качество системы до того, как вложить в неё серьёзные ресурсы».

Я пытался сохранить баланс между теоретической глубиной и практической применимостью. В каждой главе есть примеры кода, но они служат ил-

люстрацией концепций, а не самоцелью. Главная задача – научить читателя мыслить категориями RAG-архитектур, понимать их возможности и ограничения, принимать обоснованные технические решения.

Особое внимание в книге уделено отраслевым применениям RAG-систем. За эти годы я консультировал компании от стартапов до крупных корпораций, работал с проектами в юриспруденции, медицине, образовании, финансах. В каждой отрасли RAG решает свои уникальные задачи, имеет специфические требования и ограничения. Понимание этого контекста критически важно для успешного внедрения технологий.

Я благодарен всем своим коллегам и студентам, читателям канала «Технооптимисты» и слушателям одноимённого подкаста, чьи вопросы и комментарии помогали глубже понять нюансы RAG-технологий. Особая благодарность команде издательства «ДМК Пресс» за терпение и профессионализм в работе с материалом, который буквально менялся по ходу написания под влиянием новых технологических решений.

RAG-системы – это мост между традиционными подходами к поиску информации и будущими интеллектуальными помощниками. Построение этого моста требует понимания как классических алгоритмов информационного поиска, так и современных достижений в области больших языковых моделей. Надеюсь, эта книга поможет вам не просто освоить RAG-технологии, но и стать их осознанными архитекторами и пользователями.

Интересно, что корни моего интереса к проблемам создания истинных интеллектуальных систем уходят в далекие девяностые. Ещё будучи студентом МИФИ, на пороге получения диплома, я вместе с моим товарищем Владимиром Юрьевичем Степаньковым написал дерзкую работу «Об истинной интеллектуальной системе». В ней мы, молодые и амбициозные, критиковали современные нам подходы к искусственному интеллекту за фундаментальную проблему GIGO (garbage in – garbage out) – неспособность систем генерировать новые знания сверх заложенных разработчиками.

Мы отправили эту работу на студенческую конференцию, не согласовав с научным руководителем – поступок, который чуть не стоил нам дипломов – нас грозились отчислить из университета «с волчьим билетом». Но именно в той неопубликованной и потерянной на почти 30 лет[1] работе были

[1] Мы написали статью «Об истинной интеллектуальной системе» в 1998 году и отправили её на студенческую конференцию, которая проходила в Судаке, а после отказа оргкомитета и жёсткой выволочки со стороны руководства кафедры мы забыли про неё (но я-то помнил всегда). Временами за чашечкой душистого чая я напоминал Владимиру Юрьевичу о том, что когда-то мы написали такую статью, но он отрицал сам факт этого – вероятно, настолько его поразил получившийся эффект, что мозг решил компенсировать это буквальной выборочной амнезией. И вот в 2023 году во время празднования 60-летия кафедры № 22 «Кибернетика» НИЯУ МИФИ мы с ним нашли оригинальный текст этой статьи на одном из старых компьютеров учебно-научной лаборатории «Системы искусственного интеллекта» кафедры. Полный текст этой статьи с минимальными правками, касающимися исправления пары орфографических ошибок и типографики, приведён в приложении Г.

сформулированы требования к интеллектуальным системам, которые сегодня обретают новую актуальность в контексте RAG-технологий. Мы писали о необходимости самостоятельного целеполагания, автоматической генерации новых знаний, взаимодействии с человеческим социумом – всё то, что сегодня становится реальностью в современных RAG-системах.

Добро пожаловать в увлекательное время развития поистине интеллектуальных систем!

Роман Викторович Душкин
Эксперт в области ИИ
Генеральный директор ООО «А-Я эксперт»
Руководитель образовательной программы
«Искусственный интеллект» НИЯУ МИФИ
Серпухов – Москва, октябрь 2025 года

Введение

Мы живём в эпоху интеллектуального переворота. За последние три года произошло нечто, что изменило наше понимание возможностей искусственного интеллекта – от экспериментальных прототипов мы перешли к системам, способным рассуждать, анализировать и создавать контент на уровне, который ещё недавно казался недостижимым. В центре этой революции находится технология RAG (Retrieval-Augmented Generation) – подход, который превратил большие языковые модели из «умных стохастических попугаев» в действительно интеллектуальные системы.

Цифры говорят сами за себя: доля приложений с RAG выросла с 31 % до 51 % всего за один 2024 год. Более половины компаний теперь используют RAG-технологии, тогда как доля дорогостоящего файн-тюнинга упала до 9 %. Венчурные инвестиции в стартапы, специализирующиеся на генеративном искусственном интеллекте, в третьем квартале 2024 года достигли рекордных 3,9 млрд дол. – на 65 % больше, чем годом ранее. В России объём проектов по работе с большими данными и искусственным интеллектом вырос на 40 %, и среди наиболее востребованных технологий эксперты называют именно RAG.

Почему именно сейчас?

RAG стал катализатором массового внедрения корпоративных систем искусственного интеллекта по простой причине: он решает фундаментальную проблему доступа к знаниям. Традиционные языковые модели, какими бы впечатляющими ни были их возможности, ограничены данными, на которых они обучены. Они не знают о событиях, произошедших после их обучения, не имеют доступа к корпоративной информации, внутренним документам, специализированным базам знаний.

RAG кардинально изменил эту ситуацию, позволив языковым моделям динамически обращаться к актуальным внешним источникам информации. Вместо попыток «втиснуть» все знания мира в параметры модели система получила возможность искать нужную информацию в момент формирования ответа, комбинируя мощь современных моделей с гибкостью поисковых систем.

Масштаб трансформации

Влияние RAG-технологий выходит далеко за рамки технических улучшений. Это новая парадигма работы с корпоративными знаниями. В технической поддержке RAG снижает количество эскалаций на 30–40 %, а время обработки стандартных заявок сокращается на 50–60 %. В юридических и аналити-

ческих подразделениях системы высвобождают до 8 часов рабочего времени еженедельно, автоматизируя поиск и анализ документов.

Более того, RAG становится основой для следующего поколения ИИ-агентов – систем, способных не просто отвечать на вопросы, но выполнять сложные многоэтапные задачи, обращаясь к различным источникам данных и принимая обоснованные решения. По текущим экспертным прогнозам, к 2026 году более 80 % корпоративных ИИ-проектов будут использовать гибридные архитектуры, включающие большие языковые модели и методы доступа к внешним данным.

Вызовы и реальность внедрения

Однако путь к успешному внедрению RAG-систем не так прост, как может показаться из маркетинговых материалов. Качество RAG-системы напрямую зависит от качества данных, с которыми она работает. Дублирующиеся документы, устаревшая информация, противоречивые данные – все эти проблемы усиливаются в RAG-системах и могут привести к генерации некорректных ответов.

Исследования показывают, что RAG может значительно увеличить вероятность получения опасных или некорректных ответов, особенно при работе с неструктурированными или плохо организованными данными. Это подчёркивает критическую важность правильной архитектуры системы и глубокого понимания принципов её работы.

От хайпа к реальной пользе

За несколько лет работы с RAG-технологиями я видел, как компании переходят от первоначального энтузиазма к более взвешенному пониманию возможностей и ограничений этих систем. Успешные внедрения объединяет несколько общих черт: чёткое понимание бизнес-задач, тщательная подготовка данных, правильный выбор архитектурных решений и непрерывный мониторинг качества системы.

RAG – не универсальное решение для всех задач корпоративной ИИ-системы. Для генерации креативного контента лучше использовать классические языковые модели без поискового компонента. Для задач, требующих глубокого понимания узкой предметной области, может потребоваться специализированное дообучение модели. RAG наиболее эффективен там, где нужно сочетать общие языковые способности с доступом к специфическим, часто обновляющимся знаниям.

Структура и цели книги

Эта книга построена как практическое путешествие от базовых концепций к реальным корпоративным внедрениям. Первая часть знакомит читателя с фундаментальными принципами RAG-систем, их архитектурой и ключевыми компонентами, а вторая часть погружает его в технические детали: от выбора правильных фреймворков до настройки компонентов поиска и генерации.

Третья часть посвящена критически важным вопросам качества и надёжности – как оценить эффективность системы, как бороться с галлюцинациями, как обеспечить стабильную работу в производственной среде. Четвёртая часть исследует отраслевые применения RAG: от корпоративных систем управления знаниями до специализированных решений в медицине, юриспруденции, образовании.

Заключительная часть рассматривает перспективные направления развития: мультимодальные RAG-системы, интеграцию с агентными архитектурами, новые подходы к масштабированию и оптимизации. Каждая глава сочетает теоретические основы с практическими примерами, позволяя читателю не только понять концепции, но и немедленно применить их на практике.

Для кого написана эта книга

Эта книга написана для практиков, а не теоретиков. Она адресована разработчикам, которые хотят создавать эффективные RAG-системы, архитекторам, проектирующим корпоративные ИИ-платформы, продуктовым менеджерам, принимающим решения о внедрении ИИ-технологий, и руководителям, которым нужно понимать возможности и ограничения современных интеллектуальных систем.

Предполагается, что читатель имеет базовые знания в области программирования и машинного обучения, но при этом не требуется глубокая экспертиза в области обработки естественного языка или теории информационного поиска. Главная цель – дать практическое понимание того, как RAG-системы работают, почему они работают именно так и как их можно эффективно использовать для решения реальных бизнес-задач.

Методология подачи материала

В отличие от чисто академических трудов или поверхностных обзоров, эта книга следует принципу «от проблемы к решению». Каждая глава начинается с реальной практической задачи, затем объясняет теоретические основы её решения и завершается конкретными примерами реализации. Такой подход позволяет читателю не просто изучить инструменты, но понять логику их применения.

Особое внимание уделяется архитектурным принципам и паттернам проектирования. RAG-системы – это не просто набор библиотек и API, а сложные информационные системы, требующие продуманного подхода к проектированию, тестированию и эксплуатации. Понимание этих принципов критически важно для создания надёжных и масштабируемых решений.

Актуальность и перспективы

RAG-технологии развиваются с головокружительной скоростью. Многообразие инструментов для повышения точности RAG-систем стремительно растёт: от продвинутых техник переранжирования до гибридных архитектур,

объединяющих несколько типов поиска. В январе 2025 года появились исследования, расширяющие концепцию RAG на видеоконтент, что открывает новые возможности для мультимодальных моделей.

Однако в погоне за технологическим прогрессом важно не потерять понимание фундаментальных принципов. Независимо от того, какие новые фреймворки и подходы появятся в ближайшие годы, базовые концепции поиска релевантной информации, оценки её качества и включения в процесс генерации останутся актуальными. Именно эти принципы и составляют основу данной книги.

RAG – это не просто техническое решение, это новый способ мышления об искусственном интеллекте как о системе, способной учиться, адаптироваться и расти вместе с накопленными знаниями. В эпоху, когда объём информации растёт экспоненциально, а скорость принятия решений становится критически важным конкурентным преимуществом, понимание и эффективное применение RAG-технологий превращается из полезного навыка в стратегическую необходимость.