Формат <кода>

Специализированная выставка «Робототехника и искусственный интеллект» - 12 марта 2024 г.

Жуков
Александр Евгеньевич

Директор по развитию бизнеса

Формат <кода>

RAG-архитектура

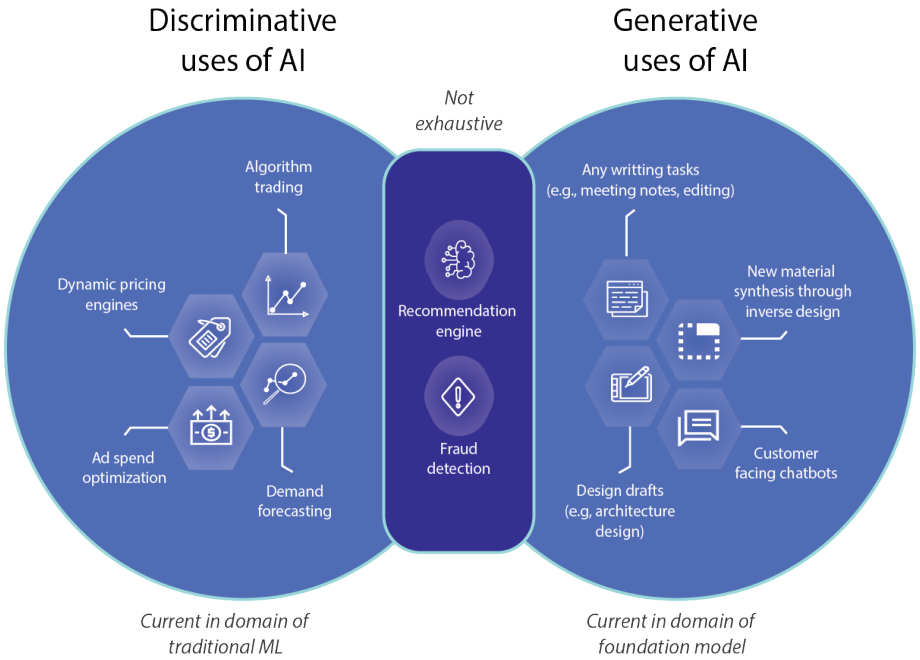
RAG и LLM как средства сделать языковые модели полезными в бизнесе и промышленности

Генеративные модели AI

В отличие от традиционных ML-моделей, генеративный AI способен на создание новой информации

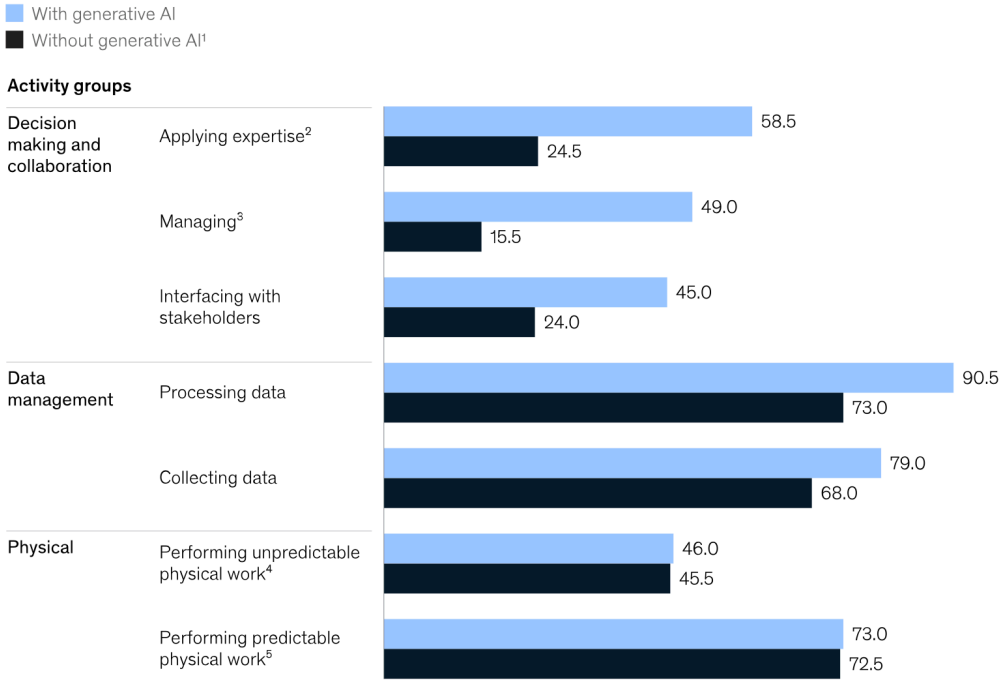
До 2023 года использование генеративных языковых моделей в критических областях бизнеса было затруднено из-за вероятной недостоверности получаемой информации.

2023 год сильно поменял правила игры



Generative AI could have the biggest impact on collaboration and the application of expertise, activities that previously had a lower potential for automation.

Overall technical automation potential, comparison in midpoint scenarios, % in 2023



Note: Figures may not sum, because of rounding.
[1]Previous assessment of work automation before the rise of generative AI.
[2]Applying expertise to decision making, planning, and creative tasks.
[3]Managing and developing people.
[4]Performing physical activities and operating machinery in unpredictable environments.
[5]Performing physical activities and operating machinery in predictable environments.
Source: McKinsey Global Institute analysis

Generative AI's impact on productivity could add trillions of dollars in value to the global economy. Our latest research estimates that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion annually across the 63 use cases we analyzed—by comparison, the United Kingdom's entire GDP in 2021 was $3.1 trillion. This would increase the impact of all artificial intelligence by 15 to 40 percent. This estimate would roughly double if we include the impact of embedding generative AI into software that is currently used for other tasks beyond those use cases.

Проблемы

Галлюцинации языковых моделей

Если современную LLM-модель до-обучить на массиве новых документов, то она сможет отвечать некоторые на вопросы по ним, однако в случаях, когда у модели не хватает информации, она переходит в режим "галлюцинирования", когда она генерирует правдоподобный текст не основываясь на фактах.

Смысловые нюансы связанных документов влияют на выводы

При обобщении информации, зачастую нельзя сделать выводы на базе единственного поступившего документа, всегда нужно анализировать связанную с ним информацию (например ГК РФ)

Безопасность

Большинство LLM-моделей работают в облаке, при этом неизвестно количество человек, имеющих доступ к данным модели. Это крайне нежелательно для работы с документацией, содержащей персональные данные и коммерческую тайну.

Кроме того...

Проекты по использованию LLM в традиционном подходе требуют

- До-обучения модели на своих данных
- Использования платных облачных сервисов
- Сложно с аналитикой релевантности и качества ответов

Рынок испытывает серьезный дефицит специалистов и подвергается множеству рисков, целевые сроки срываются, процессы внедрения происходят медленно и дорого.

Формат <кода>

Задачу можно сильно облегчить

В 2023 году появились две технологии, которые позволяют снизить уровень затрат человеческого труда:

- локальные LLM на базе LLama, преимуществом которых является полная безопасность и дешевизна.
- RAG (Retrieval Augmented Generation) которая обеспечивает минимизацию галлюцинаций и релевантность ответов

RAG

Retrieval Augmented Generation представляет собой способ избавить большие языковые модели (LLM) от галлюцинаций и недостоверных фактов

- Жестко задает контекст в виде фрагментов текста, на базе которых LLM долдна скомпоновать ответ
- Использует LLM для извлечения информации из цепочек связанных документов путем интеллектуального анализа, а не разметки страниц
- Позволяет использовать локальные LLM модели

LLama.cpp

Позволяет запускать локально (не в облаке) генеративные языковые модели уровня GPT 3.5

- Позволяет использовать различные топологии и AI-модели
- Не требует облачных интеграций и передачи в публичное облако критических данных
- Существуют совместимые модели, обученные на русском языке и способные хорошо понимать и формулировать результаты

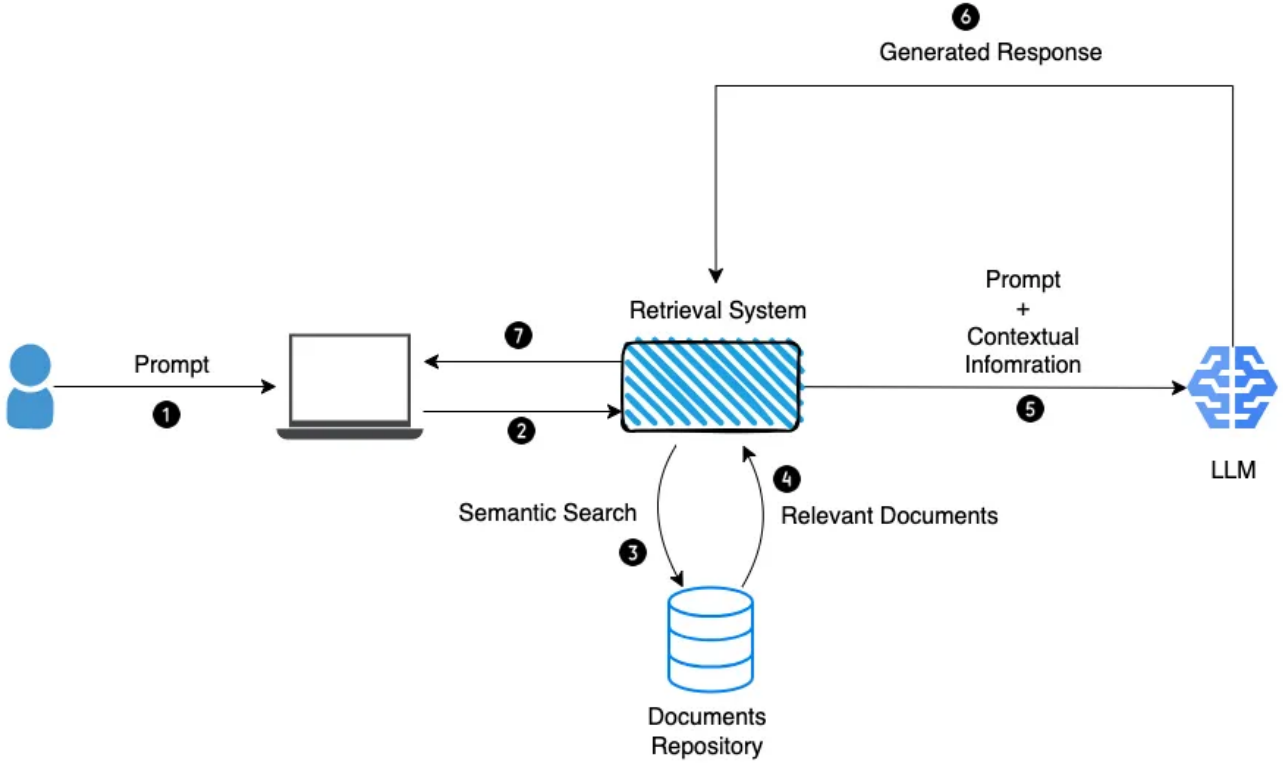
Формат <кода>

Эмбеддинги и векторные базы

Отдельное ML-решение, переводящее фрагменты текста в вектора для поиска. Позволяют сохранить и искать по архиву документов, используя "смысловую близость" вопросов и ответов

- Позволяет легко обнаружить релевантную к запросу информацию
- Задает исходные контексты для генеративной языковой модели
- Работает на локальном оборудовании без публичных облачных решений

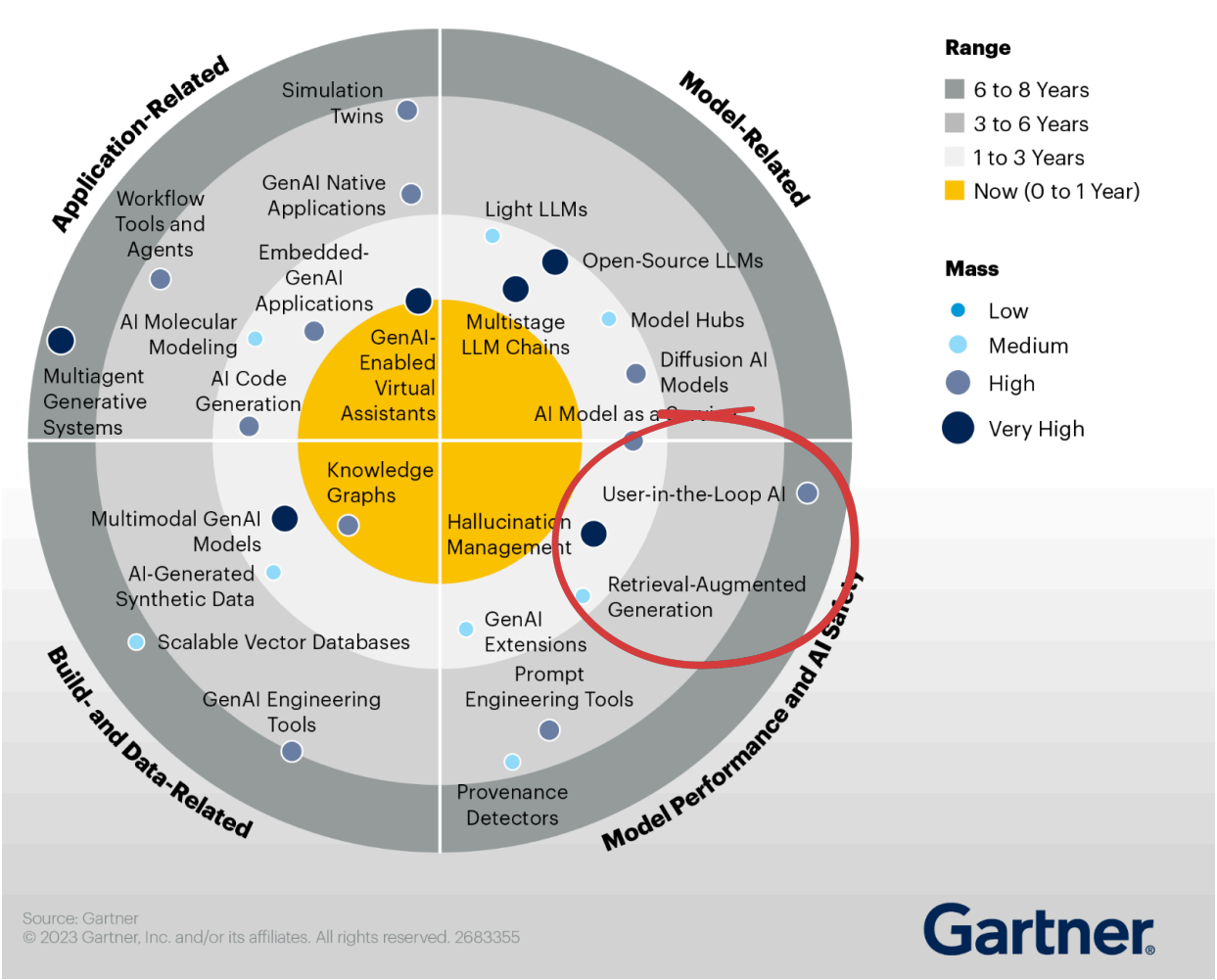
Архитектура RAG: три компонента в одной архитектуре



Формат <кода>

Где экономический эффект?

Вокруг AI много шума, но непонятно, как именно его применить



Gartner

Прогнозы аналитиков

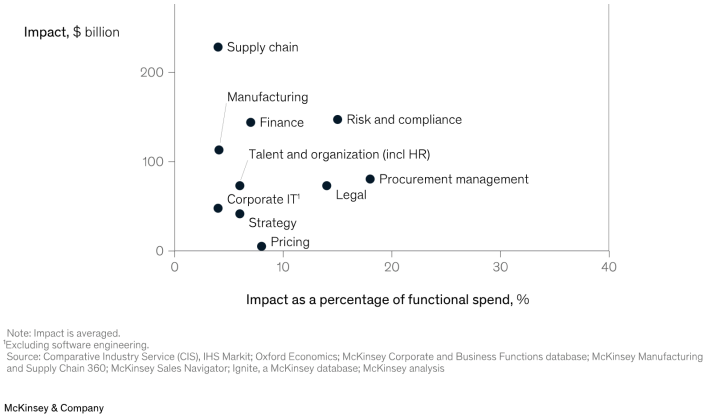
Считается что RAG войдет в обиход практических применений в течение 1-3 лет.

При этом, в рамках технологий направленных на адаптацию AI в реальных приложениях RAG находится в компании сложно отделяемых от него фич (управление галлюцинациями и совместная работа с человеком)

Практическое использование

RAG как технология уже здесь

Практически все крупные игроки облачных решений для AI уже запустили свои продукты, которые поддерживают технологию. Дело за продуктами.



Note: Impact is averaged.
[1]Excluding software engineering.
Source: Comparative Industry Service (CIS), IHS Markit; Oxford Economics; McKinsey Corporate and Business Functions database; McKinsey Manufacturing and Supply Chain 360; McKinsey Sales Navigator; Ignite, a McKinsey database; McKinsey analysis

McKinsey & Company

Oracle AI

Announcing the OCI Generative AI Agents RAG service

January 23, 2024 | 4 minute read

Barry Mostert
Senior Director, Artificial Intelligence and Analytics

We're excited to introduce the beta availability of Oracle Cloud Infrastructure (OCI) Generative AI Agents Retrieval-Augmented Generation (RAG) service, your organization's own customizable solution for conversing with and acting on diverse knowledge bases. The RAG service is the first of a series of AI Agents, with an focus on OpenSearch. Upcoming releases are expected to support a wider range of large language models (LLMs) and provide access to Oracle Database 23c with AI Vector Search and MySQL HeatWave with Vector Store.

AUQUAN UNDER THE HOOD

Actionable intelligence you can trust — powered by Retrieval Augmented Generation (RAG)

Auquan's Intelligence Engine uses Retrieval Augmented Generation (RAG), a cutting-edge enhancement to Natural Language Processing (NLP) and generative AI that addresses common AI pitfalls in the enterprise. RAG combines the strengths of retrieval-based models and generative models: the ability to pull real-time, accurate information from vast amounts of unstructured data — and the ability to craft natural, accurate summaries.

Формат <кода>

Примеры приложений

1. Анализ документации и автоматизация бизнес-процессов
2. Обобщение информации по массивам документов, протоколов и поддержка принятия решений

Поддержка принятия решений

Информирование руководства о различных показателях процесса обычно состоит в следующих этапах:
-Занесение данных в систему (человеческий фактор)
-Подготовка данных для отчета (человеческий фактор)
-Подготовка самого отчета (время и человеческий фактор)
Время и степень стандартизации этого процесса могут быть различными, однако задержки, искажения и неполнота – типичная проблема при принятии решения.

Вместо этого пути можно подключить источники информации непосредственно к RAG-модели, а обобщение информации предоставить ИИ. В этом случае для принятие решения нужно будет только буквально задавать правильные вопросы

Постоянный анализ входящей информации

Если в организации много входящей документации (неважно как поступающей - в бумаге или в электронном виде), то зачастую старт бизнес-процесса по этим документам задерживается до тех пор, пока ответственный сотрудник не занесет метаданные в СЭД или в ERP. При этом такая система плохо устойчива к нестандартным ситуациям. Традиционная альтернатива тут - писание скриптов или настройка OCR, это обычно дорого и сложно.

Для автоматизации можно легко использовать RAG-схему, что позволит задать решению простой человеческий вопрос: "Сформируй пожалуйста JSON с содержимым этого счета-фактуры с полями и значениями" и получить на выходе готовый набор метаданных.

При этом система может принимать во внимание не только содержание документа, но и дополнительные инструкции и правила их обработки.

Спасибо!

Жуков
Александр Евгеньевич

Директор по развитию бизнеса

Готов ответить на все ваши вопросы

+7 (921) 952-1521

alexander.zhukov@formatkoda.ru

formatkoda.ru