УДК 004.8

УЛУЧШЕНИЕ КАЧЕСТВА ГЕНЕРАЦИИ ОТВЕТОВ В СИСТЕМАХ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ЗА СЧЕТ ИНТЕГРАЦИИ ДОПОЛНИТЕЛЬНО НАЙДЕННОЙ РЕЛЕВАНТНОЙ ИНФОРМАЦИИ

Е. В. Ольховик[1)], К. С. Круглик[2)], Д. В. Вылегжанин[3)], О. Л. Яблонский[4)]

[1)] СООО "ХайКво Солюшенс",
ул. Тимирязева, д.67, пом.180, г. Минск, Беларусь, evgeniy.alkhovik@hiqo-solutions.com
[2)] СООО "ХайКво Солюшенс",
ул. Тимирязева, д.67, пом.180, г. Минск, Беларусь, karina.kruglik@hiqo-solutions.com
[3)] СООО "ХайКво Солюшенс",
ул. Тимирязева, д.67, пом.180, г. Минск, Беларусь, denis.vylegzhanin@hiqo-solutions.com
[4)] СООО "ХайКво Солюшенс",
ул. Тимирязева, д.67, пом.180, г. Минск, Беларусь, aleh.yablonski@hiqo-solutions.com

В данной работе рассмотрены различные подходы, используемые в современных системах искусственного интеллекта для генерации ответов, начиная с традиционных методов, основанных на правилах и шаблонах, и заканчивая более продвинутыми техниками, опирающимися на машинное и глубокое обучение. Для наглядности изложенных идей и их практического использования, в работе представлен пример чат-бота, обрабатывающего запросы в рамках корпоративных данных.

Ключевые слова: автоматическая генерация ответа; векторный поиск; большая языковая модель; гибридная система; обработка естественного языка.

В современном мире быстро развивающихся технологий и искусственного интеллекта (ИИ), возрастает потребность в создании высокоэффективных систем автоматической генерации ответов. Это особенно актуально для областей, где требуется моментальная реакция на запросы пользователей, таких как обслуживание клиентов, личные ассистенты и чат-боты. Одной из ключевых задач является повышение релевантности и информативности ответов, что может быть достигнуто за счет интеграции дополнительной информации в процесс генерации ответа.

Обзор существующих подходов к генерации ответов в системах ИИ

Существует несколько основных подходов к генерации ответов в системах ИИ:

1. Правила и шаблоны. Это один из самых ранних подходов, где система использует предопределенные правила и шаблоны для генерации ответов на основе ключевых слов или фраз, встречающихся в запросе пользователя.

2. Извлечение ответов (Retrieval-based). Системы, использующие этот подход, выбирают подходящий ответ из заранее заготовленной базы данных ответов на основе сходства запроса пользователя с имеющимися вопросами, часто используя алгоритмы машинного обучения.

3. Генерация ответов (Generation-based). Подходы, основанные на генерации, используют модели глубокого обучения, такие как рекуррентные нейронные сети или трансформеры, чтобы создавать ответы с нуля. Модели, обученные на больших наборах данных, относятся к категории больших языковых моделей (Large Language Models, LLM); они могут генерировать более разнообразные и естественные ответы благодаря их большому размеру и сложной структуре.

4. Гибридные системы. Гибридные системы сочетают в себе элементы извлечения и генерации ответов, чтобы воспользоваться преимуществами обоих подходов. Они используют извлечение для быстрого нахождения релевантного содержимого, а затем применяют генеративные модели для переформулировки или дополнения ответа. Примером гибридной системы является Retrieval Augmented Generation (RAG) система.

5. *Интерактивное обучение.* Некоторые системы используют подходы, при которых ИИ учится в процессе взаимодействия с пользователями, подстраиваясь под их предпочтения и получая обратную связь для улучшения качества генерации ответов. Этот процесс, известный как обучение с подкреплением, позволяет ИИ оптимизировать свои стратегии ответов на основе вознаграждений, получаемых за успешные взаимодействия [1].

Методы поиска дополнительной информации

Для повышения точности и релевантности ответов, системы ИИ могут интегрировать контекстную информацию и данные из различных источников: внешних и внутренних. Для эффективного поиска данных система ИИ использует разнообразные методы, включая:

- *Использование поисковых движков в интернете.* Этот подход позволяет системе ИИ оперативно интегрировать в свои ответы самые свежие данные, обеспечивая пользователей актуальной информацией из различных открытых источников, таких как Википедия или результаты Google поиска, используя API поисковых систем.

- *Векторный поиск в корпоративных базах данных.* Векторный поиск трансформирует текстовые элементы в многомерные числовые векторы, что позволяет вычислять степень их семантической близости и эффективно находить содержание, наиболее релевантное заданному пользовательскому запросу.

- *Поиск по интернету с использованием специализированных инструментов.* Данный подход представляет собой интеграцию с веб-скрейперами для извлечения данных с веб-страниц и использование API, предоставляемых платформами, для законного извлечения информации о пользователях, например, загрузка профессионального опыта и навыков из LinkedIn или списка репозиториев и вклада в проекты с GitHub.

- *Интеграция с внутренними системами управления знаниями.* В рамках интеграции с внутренними системами управления знаниями осуществляется подключение к корпоративным базам данных. Например, для SQL баз данных при поступлении вопроса от пользователя, система генерирует соответствующий запрос на языке SQL для извлечения актуальных и релевантных данных.

- *Использование графов знаний.* Графы знаний представляют собой сетевую структуру, в которой узлы соответствуют сущностям, таким как объекты, концепции или люди, а рёбра отражают отношения между ними, обеспечивая глубокое понимание их взаимосвязей. Графы знаний улучшают семантическую точность и эффективность анализа систем ИИ за счет моделирования правил и обработки больших объемов данных [2].

Пример реализации чат-бота, использующего корпоративную документацию в виде источника данных

Целью работы являлось создание чат-бота, который в качестве дополнительного источника данных использует корпоративную документацию. Для реализации данного подхода выбрана система RAG (Retrieval Augmented Generation) [3]. Архитектура RAG системы состоит из нескольких ключевых компонентов, которые обеспечивают эффективное и точное взаимодействие с пользователем (см. рисунок).

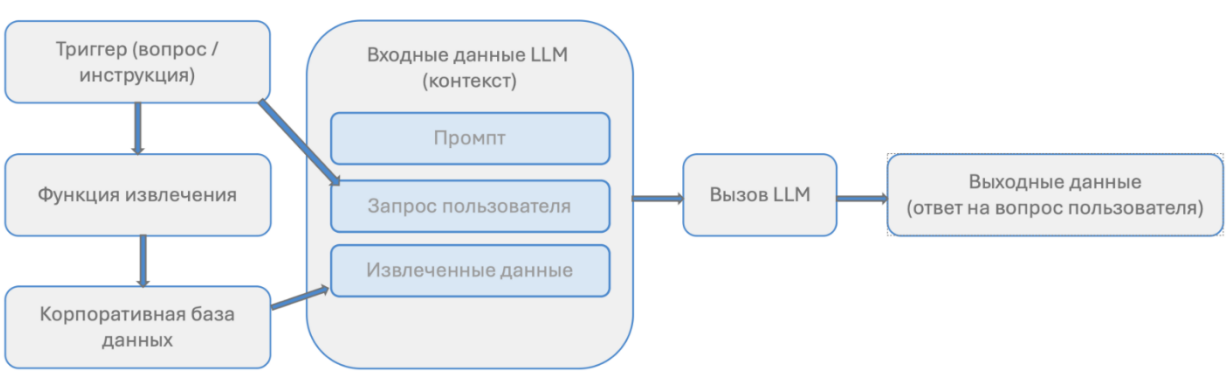

Схема RAG системы

Начальный этап в RAG системе заключается в построении индекса для всей используемой корпоративной документации. Индекс является оптимизированной структурой данных, которая позволяет быстро находить информацию в большом объеме данных. Для обеспечения быстрого и точного поиска используется векторное представление запросов и документов. В результате построения индекса вся корпоративная документация разбивается на фрагменты и каждый фрагмент данных представляется в виде многомерного вектора – эмбеддинга [4]. Функция извлечения занимается поиском и получением релевантной информации из корпоративной базы данных.

После того как векторный поиск находит наиболее релевантные документы, извлеченная информация передается в большую языковую модель (LLM). Входные данные для LLM включают не только тексты релевантных документов, но также и запрос пользователя, что позволяет модели лучше оценить контекст задачи и формировать ответы, наиболее точно соответствующие намерениям пользователя. Основываясь на знаниях, извлеченных из документов, на выходе LLM генерирует ответ, удовлетворяющий запросу пользователя.

Промпт, или инструкция, выполняет функцию связующего звена между процессом поиска информации и её последующей обработкой, ориентируя LLM на создание ответов, которые максимально точно соответствуют запросу пользователя. Разработка эффективного промпта требует грамотной подготовки и точной формулировки, чтобы языковая модель могла корректно интерпретировать запрос и предоставить релевантный ответ.

Важно также обеспечить защиту от так называемых "промпт инъекций", когда злоумышленник попытается манипулировать моделью, вводя специальным образом сформулированные запросы, чтобы получить непредназначенные для публичного доступа данные или решать не предусмотренные функционалом задачи. Это требует строгих мер безопасности и валидации ввода на всех этапах обработки запросов.

Заключение

Интеграция дополнительно найденной релевантной информации в процесс генерации ответов является ключевым фактором для создания эффективных систем ИИ, способных обеспечить высокий уровень пользовательского сервиса. Развитие алгоритмов поиска и обработки информации, а также улучшение механизмов ее интеграции позволяет достичь высокого уровня взаимодействия между человеком и системой ИИ.

Библиографические ссылки

1. Chen X., Jia S., Xiang Y. A review: Knowledge reasoning over knowledge graph // Expert systems with applications. 2020. Т. 141. С. 112948.

2. Саттон Р. С., Барто Э. Г. Обучение с подкреплением, 2-е изд. : пер. с англ. / А. А. Слинкина. М.: ДМК пресс, 2020. 552 с.

3. Retrieval-augmented generation for knowledge-intensive NLP Tasks / P. Lewis [et al.] // Advances in Neural Information Processing Systems. 2020. Т. 33. С. 9459-9474.

4. Improving text embeddings with large language models [Electronic resource] / L. Wang [et al.] // arXiv preprint arXiv:2401.00368. 2023. URL: https://arxiv.org/abs/2401.00368 (date of access: 25.03.2024).