УДК 004.855.5

ИСПОЛЬЗОВАНИЕ RAG-СИСТЕМ ДЛЯ МАСШТАБИРОВАНИЯ КЛИЕНТСКОГО ПОТОКА И ОПТИМИЗАЦИИ МАРКЕТИНГОВЫХ РАСХОДОВ

Файзиев Ф.Р. [1] (ИТМО)
Научный руководитель – к.т.н. Бутылкина К.Д. [1] (ИТМО)
[1]Университет ИТМО
faridunfaiziev@mail.ru

Введение

Что такое RAG. RAG (Retrieval-Augmented Generation) — это подход, при котором языковая модель отвечает не только на основе своих обучающих данных, но и с использованием внешнего контекста, извлечённого из базы знаний. Это позволяет обновлять знания системы без дообучения модели и использовать корпоративные документы, FAQ и регламенты как основу для ответов [1].

Что такое агент «Софа». Агент «Софа» в данной работе рассматривается как ИИ-система на базе RAG, которая не только формирует ответ, но и выбирает действие: уточнить запрос, повторить поиск, обратиться к инструменту или передать вопрос оператору. Такой подход делает систему более устойчивой в реальных сценариях клиентского общения [2].

Основная часть

Цель работы — автоматизировать ответы на типовые клиентские вопросы до и после покупки, сохраняя связь ответов с базой знаний, обеспечивая масштабируемость решения и снижая нагрузку на поддержку. Использование RAG позволяет решать эту задачу без постоянного переобучения модели при изменении данных.

Методология включает подготовку корпуса знаний, разбиение документов на фрагменты, построение эмбеддингов, организацию поиска по базе знаний и оценку качества по метрикам извлечения и генерации. Отдельно учитываются риски безопасности, включая промпт-инъекции и утечку данных [3].

В качестве источников данных используются FAQ, внутренняя база знаний, регламенты, инструкции и актуальные продуктовые материалы. Такой набор позволяет отвечать на вопросы, информация по которым отсутствует в обучающих данных модели, но есть во внутренних документах компании [4].

Для индексирования и поиска рассматривается использование векторных индексов. Это обеспечивает быстрый поиск релевантных фрагментов даже при росте объёма данных и клиентского трафика. При масштабировании приоритет отдаётся приближённому векторному поиску как более эффективному по задержкам и вычислительным затратам.

В клиентском потоке выделяются три типа запросов: вопросы до покупки, вопросы после покупки и пограничные случаи, когда данных недостаточно или источники противоречат друг другу. В первых двух ситуациях система ищет релевантные фрагменты и формирует ответ на их основе. В третьем случае агент принимает решение: запросить уточнение, повторить поиск или передать обращение дальше [5].

Бизнес-ценность решения состоит в снижении стоимости обработки типовых обращений, повышении скорости ответа и уменьшении потерь в клиентской воронке.

Внедрение RAG-агента позволяет одновременно повысить качество клиентского сервиса и эффективнее использовать маркетинговый бюджет за счёт более быстрого сопровождения пользователей на этапах выбора и покупки.

Выводы

ИИ-агент «Софа» на основе RAG может быть эффективным инструментом для масштабирования клиентского потока. Его основное преимущество заключается в сочетании автоматизации, опоры на базу знаний и возможности принимать решения в нестандартных сценариях. Это делает систему полезной как для маркетинговых задач, так и для поддержки клиентов.

Для внедрения такого решения необходимо поддерживать актуальность базы знаний, заранее проектировать систему оценки качества и учитывать требования безопасности при работе с корпоративными данными. Дальнейшие исследования могут быть связаны с улучшением качества поиска, маршрутизации сложных запросов и оценкой экономического эффекта внедрения.

Литература

1. Microsoft Learn (ru-RU). «Генерация с дополненным извлечением (RAG) предоставляет знания больших языковых моделей (LLM)» (процесс: разбиение, эмбеддинги, хранение метаданных для ссылок, подача контекста в модель). – Режим доступа: https://learn.microsoft.com/ru-ru/dotnet/ai/conceptual/rag (Дата обращения 26.02.2026).
2. OpenAI Help Center (ru-RU). «Генерация с дополнением извлечением (RAG) и семантический поиск для GPT» (определение RAG как добавления внешнего контекста во время выполнения; пример для поддержки/FAQ). – Режим доступа: https://help.openai.com/ru-ru/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts (Дата обращения 26.02.2026).
3. Microsoft Learn (ru-RU). «Общие сведения об агентах для запуска» (агент как система, которая рассуждает/планирует/действует; модель-инструкции-средства). – Режим доступа: https://learn.microsoft.com/ru-ru/windows/ai/agent-launchers/ (Дата обращения 26.02.2026).
4. OWASP / Хабр (ru-RU). «Перевод OWASP LLM Top 10» (промпт-инъекции, утечки данных и меры защиты). – Режим доступа: https://habr.com/ru/companies/owasp/articles/893712/ (Дата обращения 26.02.2026).
5. YDB Docs (ru-RU). «Векторный индекс — быстрый старт» (построение векторного индекса и поиск с индексом/без). – Режим доступа: https://ydb.tech/docs/ru/recipes/vector-search/vector-index-quickstart?version=v25.2 (Дата обращения 26.02.2026).