Автор: Алимбекова С. Р., Дмитриев О. О.
18.05.2026 09:44 -

RAG-АССИСТЕНТ В СИСТЕМАХ ДИСТАНЦИОННОГО ОБУЧЕНИЯ: СТРУКТУРНЫЕ ОГРАНИЧЕНИЯ И ОНТОЛОГИЧЕСКИЙ ПОДХОД К ИХ УСТРАНЕНИЮ

Алимбекова С. Р., д.т.н.,

Дмитриев О. О., аспирант,

АНО ВО «Московский университет «Синергия»,

г. Москва, Россия

Аннотация. Описана архитектура RAG-ассистента, разработанного для интеграции в систему дистанционного обучения. Приведены результаты экспериментальной оценки по метрикам RAGAS. Установлено несоответствие между высокими значениями стандартных метрик качества и функциональными возможностями системы применительно к структурным задачам образовательного домена: построению персонализированных траекторий обучения, диагностике пробелов в знаниях, поиску по зависимостям между темами. Идентифицированы архитектурные причины этих ограничений и обозначено направление их устранения на основе онтологически-управляемой организации знаний.

Ключевые слова: RAG, графы знаний, онтология, образовательный ассистент,

системы дистанционного обучения, адаптивное обучение, векторный поиск.

Введение

Большие языковые модели (LLM) демонстрируют низкую точность в специализированных предметных доменах: параметрические знания модели ограничены данными предобучения и не отражают содержание конкретного курса. Одним из распространённых способов решения этой проблемы в образовательных системах стала архитектура Retrieval Augmented Generation (RAG), в рамках которой перед генерацией ответа модель обращается к внешней базе знаний и использует найденные фрагменты в качестве контекста [5]. Это снижает риск «галлюцинаций» (фактически неверных ответов). Помимо этого, такой подход даёт возможность настроить систему под конкретный домен без дообучения весов модели, которое требует значительных вычислительных ресурсов.

В работе описан опыт разработки и эксплуатации подобного ассистента, встроенного в платформу дистанционного обучения. Несмотря на высокие значения стандартных метрик качества, в процессе эксплуатации выявился класс задач, с которыми плоский векторный поиск принципиально не справляется. Далее рассматриваются архитектура реализованной системы, результаты её оценки, выявленные ограничения и возможные пути их устранения.

Архитектура ассистента

В основе поискового механизма лежит гибридная схема, совмещающая два принципиально разных способа оценки релевантности. Семантический поиск

реализован средствами библиотеки Facebook (запрещённая на территории Российской Федерации) Artificial Intelligence Similarity Search (FAISS) и опирается на близость векторных представлений в многомерном пространстве, тогда как лексическое ранжирование по алгоритму Best Matching 25 (BM25) дополняет его, учитывая частоту терминов и их обратную документную частоту. Такое сочетание позволяет компенсировать характерные слабости каждого из подходов в отдельности: чисто семантический поиск нередко упускает точные терминологические совпадения, тогда как лексический плохо справляется с перефразированными запросами. Для получения векторных представлений русскоязычных текстов была выбрана многоязычная модель multilingual-e5-large – в первую очередь из-за её устойчивой работы с кириллическим контентом, что для учебных материалов на русском языке оказывается существенным практическим требованием. Исходные документы предварительно разбиваются на фрагменты алгоритмом RecursiveCharacterTextSplitter и размещаются в тематически изолированных индексах.

Обработка входящего запроса реализована в виде многоуровневого конвейера. Классификацию типа обращения выполняет ансамбль моделей, параллельно с этим механизм перефразирования формирует несколько расширенных вариантов запроса для улучшения поиска. На основе найденного контекста большая языковая модель генерирует итоговый ответ. Прежде чем ответ будет передан пользователю, он проходит через модуль верификации на предмет фактической достоверности, отсутствия галлюцинаций и строгого соответствия контексту. Взаимодействие с системой организовано через веб-интерфейс и REST API.

Экспериментальная оценка качества работы RAG-ассистента

Для оценки релевантности ответов использовался фреймворк Retrieval Augmented Generation Assessment (RAGAS), разработанный для измерения RAG-систем по независимым метрикам [3]. Данный подход позволяет анализировать достоверность генерации и точность извлечения раздельно, не сводя всё к одной агрегированной оценке. Тестовый набор формировался из вопросов по содержанию курса. Каждый

пример включал вопрос, эталонный ответ и набор контекстных фрагментов. По итогам тестирования получены следующие значения:

- Faithfulness – 0,995: галлюцинаций в ответах практически нет.
- Answer Relevancy – 1,000: ответы по смыслу соответствуют запросу.
- Context Precision – 0,900: система отбирает преимущественно нужные фрагменты.
- Context Utilization – 0,800: извлечённые данные задействуются в подавляющем большинстве случаев.

Дополнительно проводилось сравнение с режимом прямого запроса к LLM без привлечения базы знаний. Среднее значение сходства с эталонными ответами составляет 0,884 для RAG против 0,288 при прямом обращении. Разрыв воспроизводится на всём тестовом наборе. В узком предметном домене наличие специализированной базы знаний оказывает на итоговое качество ответов существенно большее влияние, чем параметры самой модели.

Структурные ограничения стандартной архитектуры RAG

Приведённые метрики [3] характеризуют качество ответа на отдельный вопрос. Они не измеряют способность системы решать задачи, естественно возникающие в учебном

процессе: диагностику готовности студента к новому разделу, выявление пробелов в знаниях, построение персонализированной траектории, поиск по зависимостям между темами курса. Причина такого ограничения носит архитектурный характер. Векторное хранилище – это набор фрагментов в метрическом пространстве: семантически близкие тексты располагаются рядом, далёкие – на расстоянии. Такая организация решает задачу локального поиска, но не содержит средств для представления структурных отношений между единицами знания. Бинарное отношение предшествования между учебными темами невозможно вывести из косинусного расстояния их векторных представлений, так как описания зависимых тем могут быть представлены с использованием принципиально разной лексики, и значение метрики близости окажется случайной величиной. Структурные связи требуют явного представления.

Следовательно, описанные ограничения не могут быть устранены путём инкрементальных улучшений. Такие меры, как замена модели эмбеддингов, более тонкое чанкование или переранжирование результатов, влияют на точность локального поиска, но не добавляют в систему того, чего в ней нет, а именно формального представления зависимостей между сущностями. Для решения структурных задач требуется структурный слой. Систематизация основных образовательных задач и применимость стандартной архитектуры RAG для их решения представлены в таблице 1.

Таблица 1. Соответствие образовательных задач возможностям RAG

Задача	**Доступна в RAG**

Причина
Ответить на вопрос по теме
Да
Семантически близкий фрагмент найдётся
Определить зависимости между темами
Нет
Структурные связи не хранятся
Построить траекторию обучения
Нет
Отсутствует модель студента и структуры курса
Диагностировать пробелы в знаниях

Нет

Нет диагностического компонента

Ответить на вопрос, связывающий две темы

Частично

Зависит от чанкования — оба аспекта могут не попасть в один фрагмент

Направление решения

Структурные связи между темами, явно зафиксированные на уровне учебной программы, подлежат явному представлению в виде графа знаний с заданной заранее схемой – онтологией. Семантический поиск при этом отвечает за работу с содержанием учебных текстов, граф за работу со структурой курса. Эти два слоя не конкурируют, а закрывают принципиально разные подзадачи.

Важно, что образовательный домен обладает здесь структурным преимуществом. Структура знаний задана явно через учебную программу. Рабочая программа дисциплины, тематический план, перечень формируемых компетенций – это, по существу, и есть описание образовательной онтологии в нестандартизированной форме [1]. Онтология не строится с нуля, а формализуется из уже существующих документов. Это снижает стоимость её создания и поддержки – один из главных инженерных рисков при работе с графами знаний [2, 4].

Ключевыми сущностями образовательной онтологии являются Курс, Модуль, Тема, Понятие, Компетенция, Задание, Профиль студента. Центральное отношение – это «предшествует» между темами: оно формализует зависимости и задаёт логически обоснованный порядок освоения. Профиль студента обновляется динамически через отношение «освоил» и связывается с конкретными узлами графа по результатам выполненных заданий.

Поверх онтологии формируется гибридный поисковый механизм. На первом шаге семантический ретривер находит стартовые узлы графа по смысловой близости запроса к привязанным текстовым фрагментам. На втором система проходит по структурным связям и строит подграф, охватывающий релевантные узлы. Этот подграф передаётся языковой модели как контекст, что позволяет обрабатывать запросы, требующие связи нескольких тем, недостижимой при плоском векторном поиске. Параллельно диагностический контур, опираясь на тот же граф и профиль студента, определяет, какие темы доступны для изучения на текущем этапе, и формирует обоснованные рекомендации о следующем шаге. Сравнительные характеристики векторного, графового и предлагаемого онтологического подходов суммированы в таблице 2.

Таблица 2. Сравнение подходов к организации знаний в образовательных системах

Аспект
Векторный RAG

Граф по корпусу

Онтологический подход

Структурные связи между темами

Отсутствуют

Частично, из текстов

Явные, заданы схемой

Модель состояния студента

Нет

Нет

Поддерживается

Персонализация траектории

Нет

Нет

Возможна

Контроль качества знаний

Слабый

Слабый

Высокий

Заключение

В ходе разработки и тестирования RAG-ассистента для платформы дистанционного обучения зафиксировано следующее. Доступ к специализированной базе знаний определяет качество ответов в предметном домене – среднее значение сходства с эталонными ответами составило 0,884 для RAG против 0,288 при прямом запросе к LLM, причём разрыв стабилен по всему тестовому набору. Метрики RAGAS это подтверждают: Faithfulness – 0,995, Answer Relevancy – 1,000, Context Precision – 0,900, Context Utilization – 0,800.

Вместе с тем перечисленные метрики характеризуют исключительно качество ответа на отдельный вопрос. Построение персонализированных траекторий, диагностика пробелов в знаниях, поиск по зависимостям между темами – всё это за их пределами.

Проблема архитектурна: плоское векторное хранилище не располагает средствами для хранения структурных отношений между единицами знания. Переход к онтологическому слою – формальной схеме курса, объединяющей содержательный и структурный уровни – позволяет устранить это ограничение. В образовательном домене онтология формализуется из уже существующих программных документов, что снижает стоимость перехода по сравнению с большинством других областей. Детальная проработка механизма адаптивной диагностики составляет предмет дальнейшего исследования.

Литература

1. Ломов П. А. Использование онтологий для контекстуализации запросов к большим языковым моделям // Онтология проектирования. – 2025. – № 2 (56).

2. da Cruz T., Tavares B., Belo F. Ontology Learning and Knowledge Graph Construction: A Comparison of Approaches and Their Impact on RAG Performance // arXiv preprint arXiv: 2511.05991. – 2025.

3. Es S. et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. – 2024. – P. 150-158.

4. Keet C. M. An Introduction to Ontology Engineering. – Version 1. – 2018.

5. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // CoRR. – 2020. – Vol. abs/2005.11401.